*Biometrika* (2010), **97**, 2, *pp*. 465–480
© 2010 Biometrika Trust
*Printed in Great Britain*

doi: 10.1093/biomet/asq017
Advance Access publication 28 April 2010

# The horseshoe estimator for sparse signals

BY CARLOS M. CARVALHO, NICHOLAS G. POLSON

*Booth School of Business, University of Chicago, Chicago, Illinois 60637, U.S.A.*
carlos.carvalho@chicagobooth.edu nicholas.polson@chicagobooth.edu

AND JAMES G. SCOTT

*McCombs School of Business, The University of Texas, Austin, Texas 78712, U.S.A.*
james.scott@mccombs.utexas.edu

SUMMARY

This paper proposes a new approach to sparsity, called the horseshoe estimator, which arises from a prior based on multivariate-normal scale mixtures. We describe the estimator's advantages over existing approaches, including its robustness, adaptivity to different sparsity patterns and analytical tractability. We prove two theorems: one that characterizes the horseshoe estimator's tail robustness and the other that demonstrates a super-efficient rate of convergence to the correct estimate of the sampling density in sparse situations. Finally, using both real and simulated data, we show that the horseshoe estimator corresponds quite closely to the answers obtained by Bayesian model averaging under a point-mass mixture prior.

*Some key words*: Normal scale mixture; Ridge regression; Robustness; Shrinkage; Sparsity; Thresholding.

## 1. INTRODUCTION

### 1·1. *The proposed estimator*

Suppose we observe a $p$-dimensional vector $y \mid \theta \sim N(\theta, \sigma^2 I)$. If $\theta$ is believed to be sparse, we propose using the following model for estimation and prediction:

$$\theta_i \mid \lambda_i \sim N(0, \lambda_i^2), \quad \lambda_i \mid \tau \sim \mathrm{C}^+(0, \tau), \quad \tau \mid \sigma \sim \mathrm{C}^+(0, \sigma),$$

where $\mathrm{C}^+(0, a)$ is a standard half-Cauchy distribution on the positive reals with scale parameter $a$. Crucially, each $\theta_i$ is mixed over its own $\lambda_i$, and each $\lambda_i$ has a half-Cauchy prior with common scale $\tau$. Additionally, we assume Jeffreys' prior for the variance, $p(\sigma^2) \propto 1/\sigma^2$. The prior for $\tau$ also follows the treatment of Jeffreys, in that it is scaled by $\sigma$, the standard deviation of the error model (Jeffreys, 1961, Ch. 5).

We estimate $\theta$ using the posterior mean under this model, which we call the horseshoe prior. This name arises from the observation that, for fixed values $\sigma^2 = \tau^2 = 1$,

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

where $\kappa_i = 1/(1 + \lambda_i^2)$, and where $E(\kappa_i \mid y)$ can be interpreted as the amount of shrinkage towards zero, a posteriori. The half-Cauchy prior on $\lambda_i$ implies a horseshoe-shaped Be$(1/2, 1/2)$ prior for the shrinkage coefficient $\kappa_i$. The left side of the horseshoe, $\kappa_i \approx 0$, yields virtually no shrinkage,

Is this a Beta distribution?

and describes signals. The right side of the horseshoe, $\kappa_i \approx 1$, yields near-total shrinkage and describes noise.

Is the sigma determined by the data (i.e. the variance of y)?

Unlike other similar procedures, the horseshoe prior is free of user-chosen hyperparameters, since the priors for $\lambda_i$, $\tau$ and $\sigma$ are all fully specified without additional inputs. Nonetheless, the prior is robust and highly adaptive, yielding strong performance across a variety of situations.

This paper's goal, aside from introducing the horseshoe prior as a modelling tool, is to propose a theoretical framework under which the model can be compared with other similar shrinkage priors. The sparse normal-means problem, while simple, can be thought of as a proving ground for methodology aimed at solving many of the common challenges in modern statistics, such as regression, classification, function estimation and regularization of covariance matrices.

We consider two major issues: robustness to large signals and shrinkage of noise. To address the first issue, we 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 estimators, and highlights the advantages of the horseshoe compared to other potential default priors. To address the second issue, we 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 behaviour 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. In short, it will leave obvious signals unshrunk, even in the face of significant noise. This unique combination of super-efficiency when the real answer is sparse, coupled with robustness when the real answer is not sparse, proves to be quite powerful in forming a low-risk estimator that can accurately separate signals from noise.

### 1·2. *The horseshoe density function*

The univariate horseshoe density function lacks an analytic form, but very tight bounds are available. For ease of notation we assume fixed values of $\sigma^2 = \tau^2 = 1$ and suppress conditioning on these terms in writing $p(\theta)$, though in general we use the priors specified in the previous section.

THEOREM 1. *The univariate horseshoe density $p(\theta)$ satisfies the following:* (*a*) $\lim_{\theta \to 0} p(\theta) = \infty$. (*b*) *For* $\theta \neq 0$,

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

*where* $K = 1/(2\pi^3)^{1/2}$.

*Proof*. See the Appendix. □

It is also possible to integrate explicitly over $\tau$, yielding a marginal density for $\lambda_i$ given by $p(\lambda_i) = 2 \log \lambda_i^2 / \{\pi^2(\lambda_i^2 - 1)\}$, though of course the terms are not independent once $\tau$ 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 conditionally independent priors $p(\lambda_i \mid \tau)$ rather than a complex joint prior $p(\lambda_1, \ldots, \lambda_p)$ over $\mathbb{R}^p$.

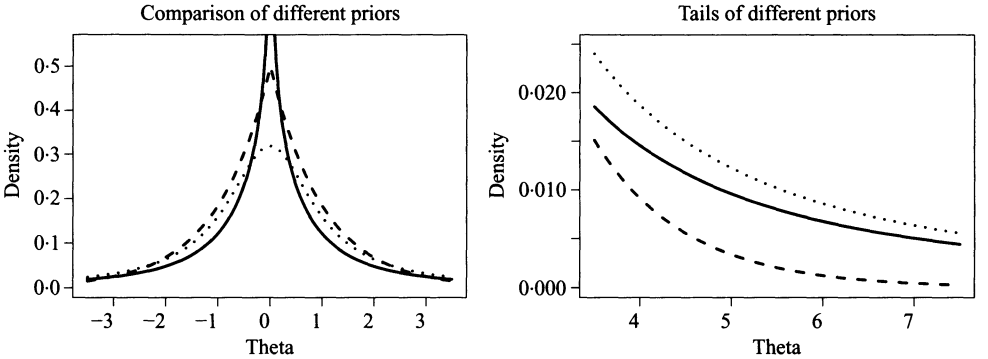


Fig. 1. Comparison of the horseshoe (solid), Cauchy (dotted) and double-exponential (dashed) densities.

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

### 1·3. *Relationship with similar methods*

The horseshoe prior assumes independent mixing densities upon $p$ idiosyncratic scale terms $\lambda_i$, 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 $\tau$. This section, while far from exhaustive, summarizes some other popular local shrinkage rules that have been considered in the literature.

The discrete mixture prior, $\theta_i \sim wg(\theta_i) + (1-w)\delta_0$, can also be represented as a variance mixture, with $\lambda_i \sim wh(\lambda_i) + (1-w)\delta_0$. The choice of $h$ will induce the form of the nonnull density $g$. If, for example, $h$ is a point mass at $\tau$, then $g$ is a $N(0, \tau^2)$ distribution. Scott & Berger (2006) study this prior extensively.

The Student-$t$ prior, $\theta_i \sim t_\xi(0, \tau)$, is defined by an inverse-gamma mixing density, $\lambda_i^2 \sim \mathrm{IG}(\xi/2, \xi\tau^2/2)$. Tipping (2001) uses this model for sparsity by finding posterior modes under the assumption that $\xi \to 0$.

The double-exponential prior has mixing density $p(\lambda_i^2 \mid \tau^2) = (2\tau^2)^{-1} \exp\{-\lambda_i^2/2\tau^2\}$, $\tau^2 \sim \mathrm{IG}(\xi/2, \xi d^2/2)$. The standard Markov chain Monte Carlo algorithm for working with the double-exponential model is due to Carlin & Polson (1991), and the use of this model in robust Bayesian inference dates at least to Pericchi & Smith (1992). A theory for the wider class of powered-exponential priors appears in West (1987). More recently, Park & Casella (2008) and Hans (2009) have revitalized interest in this prior as a Bayesian alternative to the lasso (Tibshirani, 1996).

The normal-Jeffreys prior has been studied by Figueiredo (2003) and Bae & Mallick (2004). This improper prior is induced by placing the 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, posing issues that are considered more carefully in § 3·1.

The Strawderman–Berger prior (Strawderman, 1971; Berger, 1980) lacks an analytic form, but arises from assuming $\theta_i \mid \kappa_i \sim N(0, \kappa_i^{-1} - 1)$, with $\kappa_i \sim \mathrm{Be}(1/2, 1)$. Johnstone & 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 also consider variations on this prior in an Imperial College technical report from 2000.

The normal-exponential-gamma family of priors, proposed by Griffin and Brown in a 2005 technical report from the University of Kent and further analyzed by Scheipl & Kneib (2009), is

What does "up to constants" mean here?

Table 1. *Priors for $\lambda_i$ and $\kappa_i$ associated with some common local shrinkage rules. For the normal-exponential-gamma prior, it is assumed that $d = 1$. Densities are given up to constants.*

| Prior for $\theta_i$ | Density for $\lambda_i$ | Density for $\kappa_i$ |
|---|---|---|
| Double-exponential | $\lambda_i \exp(-\lambda_i^2/2)$ | $\kappa_i^{-2} \exp\{-1/(2\kappa_i)\}$ |
| Cauchy | $\lambda_i^{-2} \exp\{1/(2\lambda_i^2)\}$ | $\kappa_i^{-1/2}(1-\kappa_i)^{-3/2} \exp[-\kappa_i/\{2/(1-\kappa_i)\}]$ |
| Strawderman–Berger | $\lambda_i\,(1+\lambda_i^2)^{-3/2}$ | $\kappa_i^{-1/2}$ |
| Normal-exponential-gamma | $\lambda_i\,(1+\lambda_i^2)^{-(c+1)}$ | $\kappa_i^{c-1}$ |
| Normal-Jeffreys | $\lambda_i^{-1}$ | $\kappa_i^{-1}(1-\kappa_i)^{-1}$ |
| Horseshoe | $(1+\lambda_i^2)^{-1}$ | $\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$ |

Jeffreys prior is most similar to horseshoe (w.r.t. K) but horseshoe changes more slowly at the extremes (perhaps implies the Jeffreys has even heavier tails at the K = 0 end?)

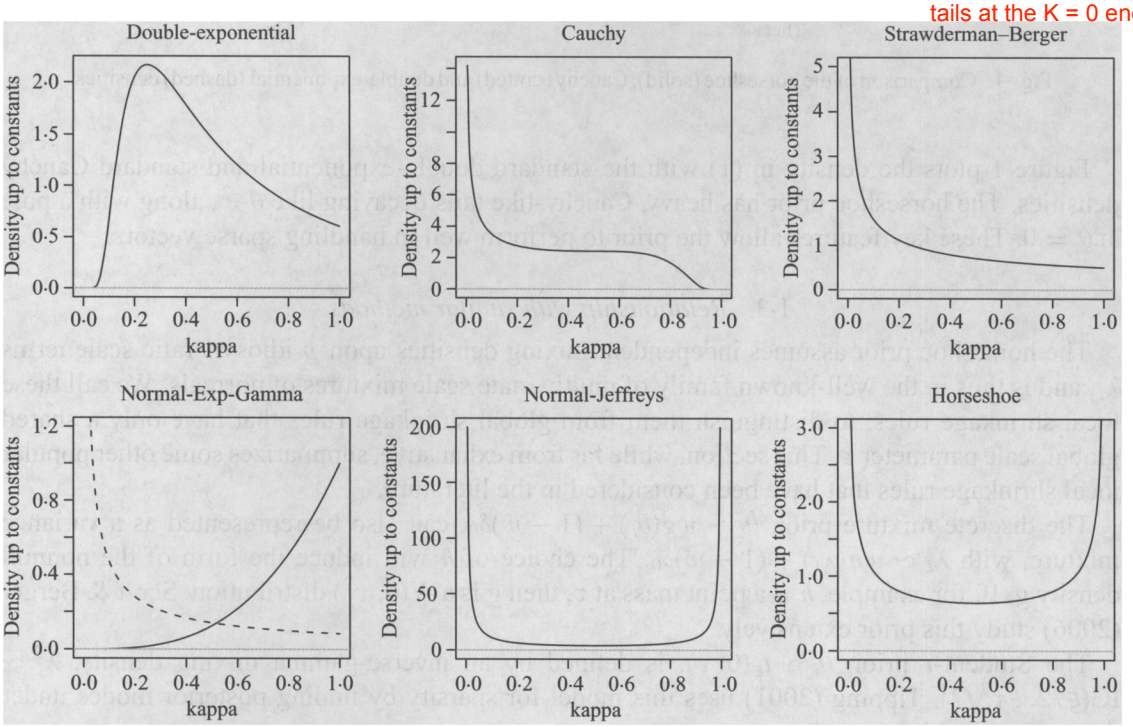


Fig. 2. The implied densities $p(\kappa_i)$ up to proportionality for six priors: the double exponential, Cauchy, Strawderman–Berger, normal-exponential-gamma, normal-Jeffreys and horseshoe. In the bottom-left panel for the normal-exponential prior, the solid line is for $c = 4$ and $d = 1$, while the dashed line is for $c = 1/4$ and $d = 1$.

also based upon the exponential mixing density, but generalizes the lasso specification using a $\mathrm{Ga}(c, d^2)$ density to mix over the exponential rate parameter. This leads to

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

The two hyperparameters $c$ and $d$ allow control over tail weight and scale, respectively.

### 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 local shrinkage rules, since $E(\theta_i \mid y_i) = \{1 - E(\kappa_i \mid y)\}\, y_i$ under a multivariate normal scale-mixture prior. The behaviour of $p(\kappa_i)$ near $\kappa_i = 0$ will control the tail robustness of the prior, and the behaviour near $\kappa_i = 1$ will control the shrinkage of noise. Table 1 lists the priors for $\lambda_i$ and $\kappa_i$ implied by the six different local shrinkage rules described above. Figure 2 also plots these

six priors on the $\kappa$ 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, reflecting their poles 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 behaviour of the posterior mean when the true vector is sparse.

The heavier-tailed priors, which are the 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, which are the double-exponential and normal-exponential-gamma with $c \geqslant 1$, all cause $p(\kappa_i)$ to vanish at $\kappa_i = 0$. These differences affect the treatment of large, obvious signals.

To provide additional insight as to how $p(\kappa_i)$ affects the behaviour of the resulting estimator, observe that if $\kappa_i \sim \text{Be}(a, b)$, then the implied prior for $\lambda_i$ is $p(\lambda_i) \propto \lambda_i^{2b-1}\,(1+\lambda_i^2)^{-(a+b)}$. This behaves like $\lambda_i^{2b-1}$ near the origin, and like $\lambda_i^{-(2a+1)}$ in the upper tail. The horseshoe prior thus marks a sharp phase transition between two extremes. If $b < 1/2$, then $p(\lambda_i)$ will be unbounded at zero, unlike under the horseshoe prior. Yet if $b > 1/2$, then $p(\lambda_i)$ vanishes at zero, and consequently $p(\theta_i)$ will be bounded. Choosing $b = 1/2$ is the only choice for which $p(\lambda_i)$ tends to a nonzero constant at the origin. The horseshoe prior does just this, yet it remains fairly noninformative on the $\kappa$ scale, since it places $1/3$ of its mass on $1/4 \leqslant \kappa_i \leqslant 3/4$.

The normal-Jeffreys prior, of course, is the improper limiting case of $\kappa_i \sim \text{Be}(\epsilon, \epsilon)$ as $\epsilon \to 0$. This will lead to tails that are even heavier, and a pole at $\theta = 0$ that is even more pronounced, compared to the horseshoe prior. Plotting this prior on the $\kappa$ scale shows just how informative it truly is, since most of the probability is highly concentrated near the extremes of 0 and 1.

## 2. Robustness to large signals

### 2·1. *A representation of the posterior mean*

The following theorem characterizes an estimator's tail robustness, or its behaviour in situations where $y$ is very different from the prior mean. Tail robustness is a useful property in sparse settings, where one would like to shrink observations near zero much more forcefully than those far from zero.

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

$$m^\star(y) = \int_{\mathbb{R}} p(|y-\theta|)\, p^\star(\theta)\, d\theta, \quad p^\star(\theta) = \int_{\mathbb{R}^+} p(\theta \mid \lambda)\, p^\star(\lambda)\, d\lambda, \quad p^\star(\lambda) = \lambda^2 p(\lambda).$$

*Then*

$$E(\theta \mid y) = \frac{m^\star(y)}{m(y)} \frac{d}{dy} \log m^\star(y) = \frac{1}{m(y)} \frac{d}{dy} m^\star(y). \tag{2}$$

*Proof*. See the Appendix. □

If $p(|y-\theta|)$ is a normal likelihood, then (2) reduces to

$$E(\theta \mid y) = y + \frac{\mathrm{d}}{\mathrm{d}y} \log m(y). \tag{3}$$

Versions of (3) appear in Masreliez (1975), Polson (1991) and Pericchi & 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 into an estimator's behaviour in the presence of large signals: Bayesian robustness may be achieved by choosing a prior for $\theta$ 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 (3) will lead to $E(\theta \mid y) \approx y$, 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* $(\theta \mid \lambda) \sim N(0, \tau^2\lambda^2)$ *and* $\lambda \sim C^+(0, 1)$. *Let* $E(\theta \mid y)$ *denote the posterior mean. Then* $|y - E(\theta \mid y)| \leqslant b_\tau$ *for some* $b_\tau < \infty$ *that depends upon* $\tau$, *and* $\lim_{|y|\to\infty} d \log m(y)/dy = 0$.

*Proof*. See the 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 1. *The value of* $E_{y\mid\theta}(\|\theta - \hat{\theta}^H\|^2)$ *is bounded for all* $\theta$.

*Proof*. Regardless of $\theta$, the risk satisfies

$$E\left\{\sum_{i=1}^{p}(\theta_i - \hat{\theta}_i^H)^2\right\} \leqslant E\left\{\sum_{i=1}^{p}(|\theta_i - y| + b_\tau)^2\right\} = p + pb_\tau^2. \qquad \square$$

Finally, although the horseshoe prior itself has no analytic form, it does yield an expression for the posterior mean:

$$E(\theta_i \mid y_i) = y_i\left\{1 - \frac{2\Phi_1(1/2, 1, 3/2, y_i^2/2, 1 - 1/\tau^2)}{3\Phi_1(1/2, 1, 5/2, y_i^2/2, 1 - 1/\tau^2)}\right\}, \tag{4}$$

where $\Phi_1(\alpha, \beta, \gamma, x, y)$ is the degenerate hypergeometric function of two variables (Gradshteyn & Ryzhik, 1965, 9.261). Combining (4) with the marginal density in (A1) allows an empirical-Bayes estimate $E(\theta \mid y, \hat{\tau})$ to be computed very rapidly.

## 3. EFFICIENCY IN HANDLING SPARSITY

### 3·1. *Joint distribution for* $\tau$ *and the* $\lambda_i$*s*

With the exception of Corollary 1, the above results describe the behaviour of the horseshoe estimator for each $\theta_i$ when $\tau$ is known. Usually, however, $\tau$ is unknown, leading to a joint distribution $p(y, \tau, \lambda_1, \ldots, \lambda_p)$ under the assumed half-Cauchy prior for $\tau$. Inspecting this joint distribution yields an understanding of how sparsity is handled under the global-local framework of the horseshoe model.

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

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

Some straightforward algebra leads to

$$p(y, \kappa, \tau^2) \propto p(\tau^2)\,\tau^p \prod_{i=1}^{p} \frac{\exp\left(-\kappa_i y_i^2/2\right)}{(1-\kappa_i)^{1/2}} \prod_{i=1}^{p} \frac{1}{\tau^2\kappa_i + 1 - \kappa_i}, \tag{5}$$

which yields several insights. As in other common multivariate scale mixtures, the global shrinkage parameter $\tau$ is conditionally independent of $y$, given $\kappa$. Similarly, the $\kappa_i$s are conditionally independent of each other, given $\tau$.

More interestingly, (5) clarifies that the global shrinkage parameter $\tau$ is estimated by the average signal density. To see this, observe that if $p$ is large, the conditional posterior distribution for $\tau^2$, given $\kappa$, is well approximated by substituting $\bar{\kappa} = p^{-1}\sum_{i=1}^{p}\kappa_i$ for each $\kappa_i$. Ignoring the contribution of the prior for $\tau^2$, this gives

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

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

Shared global parameters are of fundamental importance in high-dimensional inference. This is the insight of Stein (1956), and it applies regardless of whether sparsity is present. This fact is also central to the work of Johnstone & 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 & Berger, 2006), where global parameters play a central role in controlling the rate of Type I errors. In fact, one way of viewing our procedure is that we are asking $\tau$ 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 $\kappa_i$ 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 $\tau$ 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 $\tau$ can also lead to strong shrinkage near the origin. This shrinkage, however, can severely compromise performance in the tails. Results from Pericchi & Smith (1992) and Mitchell (1994) show that the posterior mean $E(\theta_i \mid y_i) = w_i(y_i + b) + (1-w_i)(y_i - b)$, where

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

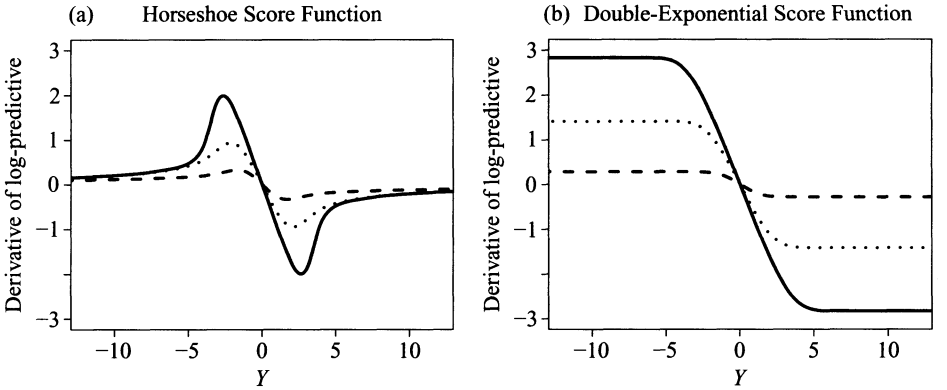


Fig. 3. A comparison of the score function for horseshoe and double-exponential priors for different values of the global scale parameter $\tau$. (a) $\tau = 0{\cdot}1$, solid line; $\tau = 1{\cdot}0$, dotted line; $\tau = 10{\cdot}0$, dashed line. (b) $\tau = 0{\cdot}5$, solid line; $\tau = 1{\cdot}0$, dotted line; $\tau = 4{\cdot}0$, dashed line.

and where $\Phi$ 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 Fig. 3. Small values of $\tau$ may help to reduce risk at the origin, but do so at the expense of increased risk in the tails, since $|E(\theta_i \mid y_i) - y_i| \approx \sqrt{2/\tau}$ for large $|y_i|$.

Therefore, when $\theta$ is sparse, estimation of $\tau$ under the double-exponential model must balance two competing forces: risk due to undershrinking noise, and risk due to overshrinking 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 a redescending score function. As Fig. 3 shows, the horseshoe prior requires no compromise of this sort.

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 $c \leqslant 1/2$ 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.

"Suppressing noise" mostly equivalent to identifying signals?

### 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 behaviour of the prior near the origin to its efficiency in handling sparsity.

True mean = 0 —> truth is sparse?

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 componentwise application of the results for a fixed value of $\tau$.

A preliminary lemma is required. To avoid notational confusion between priors and sampling models, we use $\theta_0$ to denote the true value of the parameter, $p_\theta = p(y \mid \theta)$ to denote a sampling model with parameter $\theta$ and $\mu(A)$ to denote the prior or posterior measure of some set $A$. We also let $L(p_1, p_2) = E_{p_1}\{\log(p_1/p_2)\}$ denote the Kullback–Leibler divergence of $p_2$ from $p_1$.

LEMMA 1. *Let $A_\epsilon = \{\theta : L(p_{\theta_0}, p_\theta) \leqslant \epsilon\} \subset \mathbb{R}$ denote the Kullback–Leibler information neighbourhood of size $\epsilon$, centred at $\theta_0$. Let $\mu_n(d\theta)$ be the posterior distribution under some prior*

*measure $\mu(d\theta)$ after observing data $y_{(n)} = (y_1, \ldots, y_n)$, 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 $p_{\theta_0}$, 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 $\epsilon > 0$:*

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

The proof of this lemma can be found in Clarke & Barron (1990). Intuitively, it follows from the fact that, for any $\theta$, $\{n^{-1} \log(p_{\theta_0}/p_\theta)\}$ converges to $L(p_{\theta_0}, p_\theta)$ almost surely under $p_{\theta_0}$, which allows the following approximation:

$$n^{-1} E_{p_{\theta_0}}\{\log(p_{\theta_0}/\hat{p}_n)\} \approx n^{-1} \log \int \exp\{n L(p_{\theta_0}, p_\theta)\} \mu(d\theta).$$

This lemma can be used to characterize the Kullback–Leibler risk in terms of $\mu(A_\epsilon)$, the amount of prior mass in a neighbourhood of $\theta_0$. The horseshoe prior's pole at zero produces a super-efficient rate of convergence when $\theta_0 = 0$.

THEOREM 4. *Suppose the true sampling model $p_{\theta_0}$ 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 $R_n = O\{n^{-1}(\log n - b \log\log n)\}$, where $b$ is a constant. When $\theta_0 \neq 0$, the optimal rate is $R_n = O(n^{-1} \log n)$.*

(2) *Suppose $p(\theta)$ is any other density that is continuous, bounded above, and strictly positive on a neighbourhood of the true value $\theta_0$. For $\hat{p}_n$ under $p(\theta)$, the optimal rate of convergence of $R_n$, regardless of $\theta_0$, is $R_n = O(n^{-1} \log n)$.*

So horseshoe is great for sparse situations, and average comp. to other priors in non-sparse contexts

*Proof*. See the Appendix. □

Two further remarks help to set this theorem in context. 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 $\theta$ are zero has been explicitly flagged as an interesting possibility. Yet if $\theta_0 \neq 0$, the horseshoe yields no worse a 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 $O(n^{-1} \log n)$ 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 $\theta_i$ is signal or noise. The decision rule, though informal, appears nearly indistinguishable from the formal Bayes rule under the discrete-mixture model and a symmetric loss function where false negatives and false positives are penalized equally. This suggests an interesting correspondence between the two procedures.

Under the discrete mixture-model described in § 1, the Bayes estimator for each component is the posterior mean $\hat{\theta}_i = w_i E_g(\theta_i \mid y_i)$, where $w_i$ is the posterior inclusion probability for $\theta_i$, and $g$ is the distribution of the nonzero means. Accordingly, $w_i$ serves a dual role. First, it is a posterior probability giving rise to a formal Bayes decision rule about whether $\theta_i$ should be

Table 2. *Posterior probabilities $w_i$ under the discrete-mixture model, expressed as percentages, for 10 fixed signals of varying strength. These same ten signals are tested in nine datasets where there are increasingly many standard-normal noise observations, with numbers given in the left-most column. The bracketed numbers are the corresponding shrinkage weights $1 - \hat{\kappa}_i$ under the horseshoe model, also expressed as percentages.*

| | Signal strength | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Noise | 0·5 | 1·0 | 1·5 | 2·0 | 2·5 | 3·0 | 3·5 | 4·0 | 4·5 | 5·0 | FP |
| 25 | 18 (15) | 20 (17) | 23 (18) | 27 (21) | 31 (24) | 35 (28) | 39 (32) | 44 (37) | 49 (42) | 54 (47) | 0 (0) |
| 50 | 8 (11) | 10 (12) | 12 (14) | 15 (17) | 19 (20) | 25 (26) | 32 (33) | 41 (40) | 50 (49) | 60 (57) | 0 (0) |
| 100 | 8 (14) | 10 (17) | 15 (22) | 27 (31) | 46 (46) | 69 (62) | 86 (75) | 95 (85) | 99 (89) | 100 (92) | 0 (0) |
| 200 | 5 (11) | 6 (12) | 11 (17) | 21 (26) | 42 (43) | 70 (63) | 90 (79) | 98 (87) | 100 (91) | 100 (93) | 2 (1) |
| 500 | 1 (4) | 2 (4) | 3 (6) | 6 (10) | 14 (18) | 35 (36) | 67 (61) | 91 (80) | 98 (89) | 100 (92) | 1 (1) |
| 1000 | 0 (1) | 1 (1) | 1 (2) | 2 (3) | 3 (5) | 9 (10) | 24 (25) | 55 (52) | 85 (76) | 97 (88) | 0 (0) |
| 2000 | 0 (1) | 0 (1) | 1 (1) | 1 (2) | 3(4) | 8 (9) | 24 (24) | 59 (54) | 89 (80) | 98 (90) | 0 (0) |
| 5000 | 0 (0) | 0 (0) | 0 (0) | 0 (1) | 1 (1) | 3 (3) | 9 (10) | 32 (33) | 72 (67) | 95 (86) | 0 (0) |
| 10 000 | 0 (0) | 0 (0) | 0 (0) | 0 (1) | 1 (1) | 3 (3) | 9 (10) | 32 (30) | 74 (68) | 96 (88) | 3 (2) |

FP, the number of false positive declarations, reflecting cases where the posterior probability $w_i$ or shrinkage weight $1 - \hat{\kappa}_i$ is larger than 0·5 for a noise observation.

classified as signal or noise. Second, it measures how aggressively $y_i$ should be shrunk to zero when estimating $\theta_i$ under squared-error loss.

For appropriately heavy-tailed $g$, the posterior mean under the discrete-mixture rule is approximately $w_i y_i$. Compare this form to the horseshoe estimator: $\hat{\theta}_i = (1 - \hat{\kappa}_i) y_i$, where $\hat{\kappa}_i$ is the posterior mean of $\kappa_i$. Clearly, the shrinkage weight $1 - \hat{\kappa}_i$ plays the same role as $w_i$ in the discrete-mixture model. It is therefore natural to ask whether these weights, even though they cannot be interpreted as posterior probabilities, can nonetheless be used to construct an informal decision rule for classifying each $\theta_i$ as signal or noise.

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 based on the horseshoe prior is to call $\theta_i$ a signal if $(1 - \hat{\kappa}_i) \geqslant 0{\cdot}5$, and to call it noise otherwise. 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 to nine simulated datasets having an increasingly large number of standard-normal noise observations. We compared the horseshoe shrinkage weights $1 - \hat{\kappa}_i$ to the posterior inclusion probabilities from the discrete-mixture rule using Strawderman–Berger priors. Results are shown in Table 2.

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, there is a striking correspondence between the shrinkage weights from the horseshoe model and the true posterior probabilities from the discrete-mixture model. This can be seen from Table 2, 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 $(1 - \hat{\kappa}_i)$ under the double-exponential prior are not shown, they do not behave at all like the $w_i$ from the discrete mixture model. These results, and many more simulations along these lines, can be found in the third author's unpublished doctoral thesis, available from Duke University.

Table 3. *Realized squared-error loss under different estimators*

| | $w = 0{\cdot}05$ | | $w = 0{\cdot}2$ | | $w = 0{\cdot}5$ | |
|---|---|---|---|---|---|---|
| | $\xi = 2$ | $\xi = 10$ | $\xi = 2$ | $\xi = 10$ | $\xi = 2$ | $\xi = 10$ |
| MLE | 250 | 248 | 249 | 251 | 252 | 251 |
| Double-exponential | 171 | 127 | 237 | 217 | 247 | 235 |
| NEG ($c = 4{\cdot}0, d = 3$) | 121 | 121 | 134 | 134 | 186 | 183 |
| NEG ($c = 2{\cdot}0, d = 3$) | 165 | 164 | 170 | 171 | 187 | 187 |
| NEG ($c = 1{\cdot}0, d = 3$) | 199 | 197 | 201 | 202 | 208 | 208 |
| NEG ($c = 0{\cdot}5, d = 3$) | 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 |

$w$, the degree of sparsity; $\xi$, the tail weight of the true $t_\xi$ signal density; MLE, maximum likelihood estimator; NEG, normal-exponential-gamma model.

## 4. Examples

### 4·1. *Simulated data*

Table 3 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 model due to Johnstone & 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 $\theta_i$ using the posterior median. This last comparison is an especially important benchmark, as it is widely recognized as the gold standard in handling sparsity.

Our study involved simulating from the following sparse model:

$$(y_i \mid \theta_i) \sim N(\theta_i, 1), \quad \theta_i \sim w\, t_\xi(0, \tau) + (1 - w)\delta_0,$$

Can we interpret the w as the probability of a signal/nonzero effect?

where $\delta_0$ is a point mass at zero, and where $t_\xi(0, \tau)$ is a Student-$t$ density centred at zero, with $\xi$ degrees of freedom and scale parameter $\tau$.

In all our simulations, we set $\tau = 3$, and investigated six configurations of tail weight and sparsity by choosing $\xi \in \{2, 10\}$ and $w \in \{0{\cdot}05, 0{\cdot}2, 0{\cdot}5\}$. These combinations span a wide range of behaviours, from very sparse signals with very heavy tails, to mildly sparse signals with much lighter tails. For each combination we simulated 500 datasets.

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, $\sigma$, $\tau$ and $w$ were estimated by marginal maximum likelihood.

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 $0{\cdot}1 \leqslant d \leqslant 10$ and $1/2 \leqslant c \leqslant 8$. We report results for five of these choices in Table 3. 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 $\xi$ and $w$, which could only be judged after the fact.

Our results show the double-exponential prior systematically losing out to the horseshoe. We attribute this to the two features mentioned previously: that exponential tails are insufficiently heavy to estimate large signals when noise is present, and that a pole at zero aids in reducing the substantial amount of noise in these problems.

The horseshoe prior also systematically beats the default normal-exponential-gamma priors, and has a slight edge over the best fixed choice of $c$ and $d$. 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 $w = 0{\cdot}5$. The horseshoe prior was beaten only in the situation when the signal was neither sparse nor heavy-tailed, with $w = 0{\cdot}5$ 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 use the data on Vanguard mutual funds from Carvalho & Scott (2009), which contains $n = 86$ weekly returns for $p = 59$ funds.

The connection with regression is as follows. Suppose we observe a matrix of samples $Y^{\mathrm{T}} = (y^1 \cdots y^n)$, with each $p$-dimensional vector $y^i$ drawn from a zero-mean normal distribution with unknown covariance matrix $\Sigma$. When $p$ is large relative to $n$, traditional estimators of $\Sigma$ are known to perform poorly, and some form of regularization is necessary to reduce their variance. We choose to model the Cholesky decomposition of $\Sigma^{-1}$ and estimate the ensemble of regression models in the implied triangular system $\{Y_j \mid Y_1, \ldots, Y_{j-1}\}_{j=2,\ldots,p}$, where $Y_j$ is the $j$th column of the matrix of samples. Horseshoe priors were assumed for the vector of coefficients in each of these regressions, and posterior means were computed using the Markov chain Monte Carlo method.

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 $\Sigma$: the maximum-likelihood estimate $\hat{\Sigma} = Y^{\mathrm{T}}Y$; the AND and the OR versions of the lasso, described by Meinshausen & Buhlmann (2006); and Bayesian model-averaging over different Gaussian graphical models, using fractional Bayes factors for computing marginal likelihoods and feature-inclusion stochastic search for model determination (Scott & Carvalho, 2008).

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

The results are in Table 4, 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, rather merely to show 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, both on simulated and real data.

Table 4. *Covariance-estimation example. The table entries are risk ratios versus Bayesian model averaging in the out-of-sample prediction exercise*

| | 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 |

SE, squared-error loss; AE, absolute-error loss; MLE, maximum likelihood estimator; BMA, Bayesian model averaging.

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 $\tau$, along with robust local adaptation to signals through the $\lambda_i$s, is unmatched by other common scale-mixture priors.

"Averaging over models instead of picking one" seems much more in tune with the Bayesian ethos

Finally, a word on sparsity. Many similar procedures, most notably the lasso, estimate $\theta$ 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 $\theta$ 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 $\theta$ like ours, in 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, or in this context, averaging over the two peaks at 0 and 1 in the posterior distribution for each local shrinkage parameter $\kappa_i$. 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.

## Acknowledgement

The first author acknowledges the support of the IBM Corporation Scholar Fund at the University of Chicago Booth School of Business. The third author acknowledges the support of a graduate research fellowship from the National Science Foundation, U.S.A. All authors would like to thank the referees for their careful suggestions in improving this work.

## Appendix

*Proof of Theorem* 1. Clearly,

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

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

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

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

$$p(\theta) = K e^{\theta^2/2} \int_1^\infty \frac{1}{z} \exp\left(-\frac{z\theta^2}{2}\right) \mathrm{d}z$$

$$= K \exp\left(\frac{\theta^2}{2}\right) E_1\left(\frac{\theta^2}{2}\right),$$

where $E_1(\cdot)$ is the exponential integral function. This function satisfies tight upper and lower bounds:

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

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

*Proof of Theorem* 2. First, $m^\star(y)$ exists for any proper prior, since it exists for $p(\lambda^2) \equiv 1$, which leads to the harmonic estimator in the case of a normal likelihood. This is sufficient to allow the interchange of integration and differentiation.

We make use of the following identities:

$$\frac{\mathrm{d}}{\mathrm{d}y} p(y - \theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta} p(y - \theta), \quad \lambda^2 \frac{\mathrm{d}}{\mathrm{d}\theta}\{N(\theta \mid 0, \lambda^2)\} = -\theta N(\theta \mid 0, \lambda^2).$$

Clearly,

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

Using integration by parts and the above identities, we obtain

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

from which the result follows directly. □

*Proof of Theorem* 3. Clearly,

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

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

$$m(y) = \frac{1}{(2\pi^3)^{1/2}} \int_0^1 \exp(-zy^2/2)(1 - z)^{-1/2} \left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) z\right\}^{-1} \mathrm{d}z$$

$$= \frac{2}{\tau(2\pi^3)^{1/2}} \exp\left(-\frac{y^2}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right). \tag{A1}$$

By a similar transformation, it is easy to show that

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

Hence

$$\frac{\mathrm{d}}{\mathrm{d}y} \log m(y) = -\frac{2y \; \Phi_1\left(1/2, 1, 3/2, y_i^2/2, 1 - 1/\tau^2\right)}{3\Phi_1\left(1/2, 1, 3/2, y_i^2/2, 1 - 1/\tau^2\right)}. \tag{A2}$$

Next, we use the following identity from Gordy (1998):

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(x) \sum_{n=0}^\infty \frac{(\alpha)_n (\beta)_n}{(\gamma)_n} \frac{y^n}{n!} {}_1F_1(\gamma - \alpha, \gamma + n, -x), \tag{A3}$$

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

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

The final identities necessary are from Chapter 4 of Slater (1960). For a real number $x$,

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

Hence regardless of the sign of $1 - 1/\tau^2$, expanding (A2) using these identities yields a polynomial of order $y^2$ or greater left in the denominator, from which the redescending score function follows. The bound $|y - E(\theta \mid y)| \leqslant b_\tau$ then follows from the continuity of (A1), which evaluates to 0 at $y = 0$. □

*Proof of Theorem* 4. The optimal rate of convergence, following Clarke & Barron (1990), comes from choosing $\epsilon_n = 1/n$, which reflects the ideal case of independent samples $y_1, \ldots, y_n$.

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

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

since the density is bounded above. Applying Lemma 1, the optimal rate for Part 2 is

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

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 neighbourhood near $\theta_0$. When $\theta_0 = 0$, we can use the bound on the density given previously, $2(2\pi^3)^{1/2}p(\theta) \geqslant \log(1 + 4\theta^{-2})$. Ignoring constant factors not depending upon $n$, this leads to

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

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

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

Upon integrating by parts, we then have

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

This last integral is easily computed and of order $\epsilon^{1/2}$. Setting $\epsilon = 1/n$ and applying Lemma 1 then gives the optimal rate bound as $R_n = O\{n^{-1}(\log n - b \log\log n)\}$, proving Part 1. □

## References


Bae, K. & Mallick, B. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–30.

Berger, J. O. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8**, 716–61.

Berry, D. (1988). Multiple comparisons, multiple tests, and data dredging: a Bayesian perspective. In *Bayesian Statistics* 3, Ed. J. Bernardo, M. DeGroot, D. Lindley and A. Smith, pp. 79–94. New York: Oxford University Press.

Carlin, B. P. & Polson, N. G. (1991). Inference for nonconjugate bayesian models using the Gibbs sampler. *Can. J. Statist.* **19**, 399–405.

CARVALHO, C. M. & SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96**, 497–512.
CLARKE, B. & BARRON, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Info. Theory* **36**, 453–71.
FIGUEIREDO, M. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. Pat. Anal. Mach. Intel.* **25**, 1150–9.
GORDY, M. B. (1998). A generalization of generalized beta distributions. In *Finance and Economics Discussion Series*. Board of Governors of the Federal Reserve System.
GRADSHTEYN, I. & RYZHIK, I. (1965). *Table of Integrals, Series, and Products*. New York: Academic Press.
HANS, C. M. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–45.
JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. New York: Oxford University Press.
JOHNSTONE, I. & SILVERMAN, B. W. (2004). Needles and straw in haystacks: empirical-Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594–649.
MASRELIEZ, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE. Trans. Auto. Contr.* **20**, 107–10.
MEINSHAUSEN, N. & BUHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–62.
MITCHELL, A. F. (1994). A note on posterior moments for a normal mean with double-exponential prior. *J. R. Statist. Soc.* B **56**, 605–10.
PARK, T. & CASELLA, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.* **103**, 681–6.
PERICCHI, L. R. & SMITH, A. (1992). Exact and approximate posterior moments for a normal location parameter. *J. R. Statist. Soc.* B **54**, 793–804.
POLSON, N. G. (1991). A representation of the posterior mean for a location model. *Biometrika* **78**, 426–30.
SCHEIPL, F. & KNEIB, T. (2009). Locally adaptive Bayesian p-splines with a normal-exponential-gamma prior. *Comp. Statist. Data Anal.* **53**, 3533–52.
SCOTT, J. G. & BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plan. Infer.* **136**, 2144–62.
SCOTT, J. G. & CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comp. Graph. Statist.* **17**, 790–808.
SLATER, L. J. (1960). *Confluent Hypergeometric Functions*. New York: Cambridge University Press.
STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. In *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, vol. 1.
STRAWDERMAN, W. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Statist.* **42**, 385–8.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
TIPPING, M. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **1**, 211–44.
WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–8.

[*Received December* 2008. *Revised December* 2009]