The Horseshoe Estimator for Sparse Signals

Carlos M. Carvalho
Nicholas G. Polson
James G. Scott

October 2008

Abstract

This paper proposes a new approach to sparse-signal detection called the horseshoe estimator. We show that the horseshoe is a close cousin of the lasso in that it arises from the same class of multivariate scale mixtures of normals, but that it is almost universally superior to the double-exponential prior at handling sparsity. A theoretical framework is proposed for understanding why the horseshoe is a better default "sparsity" estimator than those that arise from powered-exponential priors. Comprehensive numerical evidence is presented to show that the difference in performance can often be large. Most importantly, we show that the horseshoe estimator corresponds quite closely to the answers one would get if one pursued a full Bayesian model-averaging approach using a "two-groups" model: a point mass at zero for noise, and a continuous density for signals. Surprisingly, this correspondence holds both for the estimator itself and for the classification rule induced by a simple threshold applied to the estimator. We show how the resulting thresholded horseshoe can also be viewed as a novel Bayes multiple-testing procedure.

Keywords: shrinkage; lasso; Bayesian lasso; multiple testing; empirical Bayes.

1 Introduction

1.1 Estimating signals of unknown sparsity

The primary goal of this paper is to introduce a novel procedure for estimating a sparse vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. Our estimator arises as the posterior mean under an exchangeable model $\pi_H(\theta_i)$ that we call the horseshoe prior. The name "horseshoe" relates not to the shape of the density itself, but rather to the shape of the implied prior for the shrinkage coefficients applied to each observation. This line of thought will take some development, but for a "sneak peak", flip to Figure 4.

Sparse-signal detection is fundamentally a two-groups question: given observations arising from a sparse vector, which components are signals, and which are noise? Modern Bayesian and empirical-Bayes approaches are capable of giving quite sophisticated two-groups answers to this question through the use of discrete mixtures. These methods can effectively characterize both the groups themselves, and group membership of individual components, via simultaneous shrinkage and selection in a way that can appeal both to Bayesians and frequentists alike.

A very different approach to sparsity involves a one-group model—for example, the lasso—that describes both signals and noise with a single continuous prior distribution. One-group models enjoy enormous popularity due to their computational simplicity and various asymptotic guarantees. Yet to Bayesians, one-group models appear to dodge the fundamental two-group question of "signal versus noise" altogether, arriving at sparse solutions only through artifice.

Our approach differs from past work in that it does not rely upon the posterior mode to induce zeroes in $\hat{\boldsymbol{\theta}}$. But neither does it employ a discrete mixture comprising both a continuous density and a point mass at zero. Despite this, the horseshoe prior turns out to be quite adept at handling cases in which many components of $\boldsymbol{\theta}$ are exactly or approximately 0, a situation that is ubiquitous in modern scientific problems involving high-throughput experiments. Like the lasso, the horseshoe is a one-group model; unlike the lasso, it gives virtually the same answers as a well-developed two-group model.

We choose to study sparsity in the simplified context where $\boldsymbol{\theta}$ is a vector of normal means: $(\mathbf{y}|\boldsymbol{\theta}) \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I)$, where σ^2 may be unknown. It is here that the lessons drawn from a comparison of different approaches for modeling sparsity are most readily understood, but these lessons generalize straightforwardly to more difficult problems—regression, covariance regularization, function estimation—where many of the challenges of modern statistics lie.

Of course, if both the degree and nature of underlying sparsity in $\boldsymbol{\theta}$ are well understood beforehand, then these assumptions should be used to construct a well-tuned estimator (which will always be optimal with respect to its assumed prior distribution). Often, however, neither of these two things are known even approximately. Perhaps most of the entries in $\boldsymbol{\theta}$ are identically zero; this is often called strong sparsity. But perhaps instead most entries are nonzero yet small compared to a handful of large signals; this is often called weak sparsity. For example, $\boldsymbol{\theta}$ may be of bounded l^α norm for some suitably small α, or its entries may decay in absolute value according to some power law. Indeed, both conditions can be true at the same time, with some θ_i's being zero and the rest being themselves weakly sparse.

We recommend the horseshoe as "jack-of-all-trades" estimator for just these situations. This paper will focus on three main strengths of the horseshoe: adaptivity, both to unknown sparsity and to unknown signal-to-noise ratio; robustness to large, outlying signals; and multiplicity control, or the level of control over the number of false-positive flags in vectors with many zeros. In developing these ideas, we hope

that different versions of our methodology will hold appeal for three groups:

- Non-Bayesians interested in shrinkage and classification, and who use lasso-type estimators to induce zeros in $\boldsymbol{\theta}$ (for example, Tibshirani, 1996).
- Bayesians who do not trust the zeros induced by lasso-type estimators, and who want to see a full posterior distribution for each θ_i (for example, Park and Casella, 2008; Hans, 2008).
- Bayesians and empirical-Bayesians who seek to classify observations using a two-group model for signals and noise (for example, Johnstone and Silverman, 2004; Scott and Berger, 2006; Efron, 2008).

This last point is of particular interest. Despite being an estimation procedure, our approach turns out to share one of the most appealing features of Bayesian and empirical-Bayes model-selection techniques: it exhibits an automatic penalty for multiple hypothesis testing (Berry, 1988). The nature of this multiple-testing penalty is well understood in discrete mixture models, and we will clarify how a similar effect occurs when the horseshoe prior is used instead.

1.2 The proposed estimator

Assume that each θ_i arises independently from a location-scale density $\pi_H(\theta_i|0,\tau)$, where π_H has the following interpretation as a scale mixture of normals:

$$(\theta_i \mid \lambda_i, \tau) \sim \mathrm{N}(0, \lambda_i^2\tau^2) \tag{1}$$

$$\lambda_i \sim \mathrm{C}^+(0,1)\,, \tag{2}$$

where $\mathrm{C}^+(0,1)$ is a standard half-Cauchy distribution on the positive reals. We call the λ_i's the *local* shrinkage parameters, and τ the *global* shrinkage parameter. Estimation of the global shrinkage parameter is very important and will be considered in Section 6, but for ease of understanding we assume τ to be fixed at 1.

It must be emphasized that the horseshoe prior involves something fundamentally different than simply placing a half-Cauchy prior on a common variance component. Rather, each mean θ_i is mixed over its own λ_i, each of which has an independent half-Cauchy prior.

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

Theorem 1.1. *Let $K = 1/\sqrt{2\pi^3}$. The horseshoe prior has the following properties:*

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

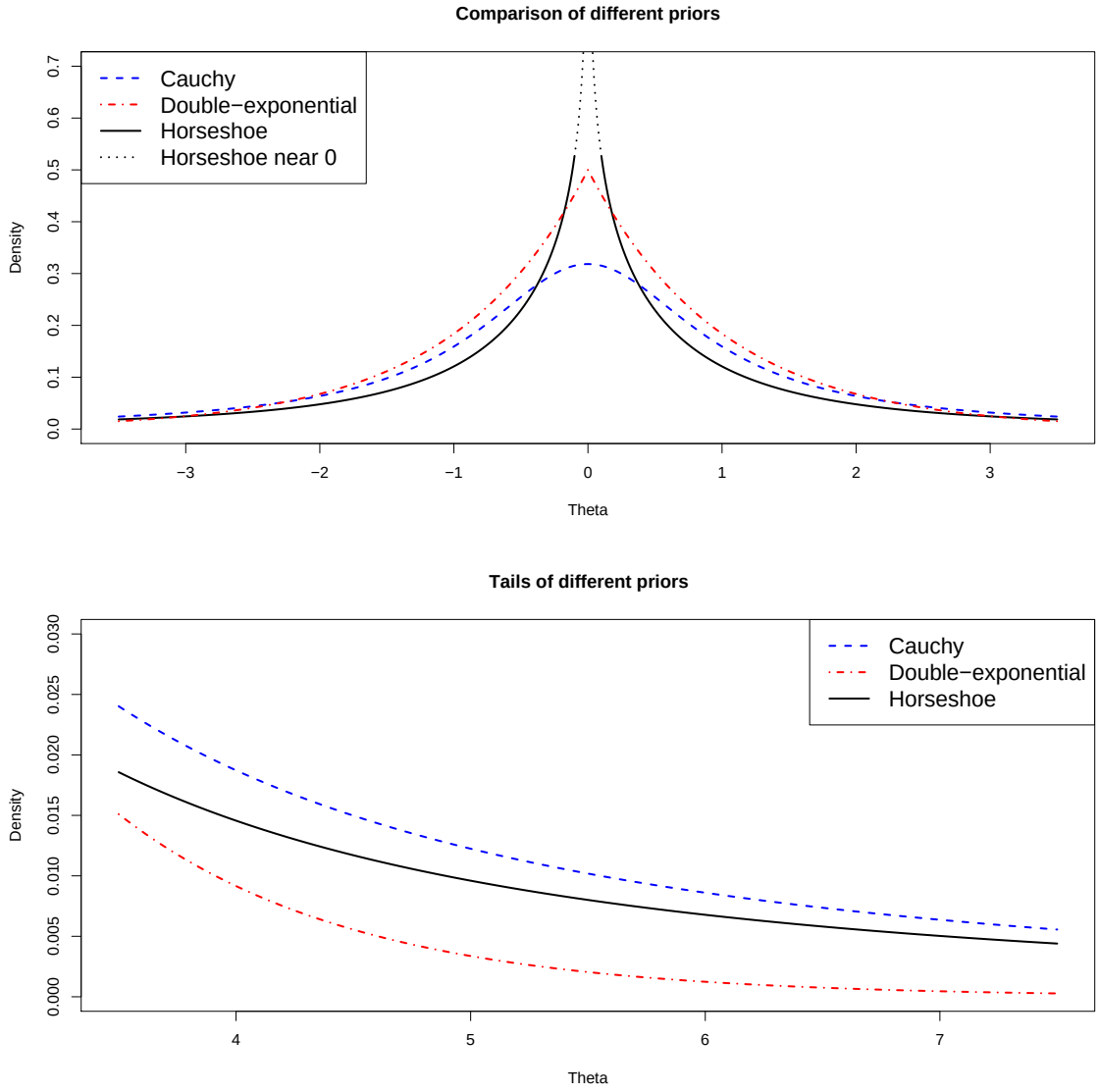

Figure 1: A comparison of π_H versus standard Cauchy and double-exponential densities; the dotted lines indicate that π_H approaches ∞ near 0.

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

$$\frac{K}{2}\log\left(1+\frac{4}{\theta^2}\right) < \pi_H(\theta) < K\log\left(1+\frac{2}{\theta^2}\right). \tag{3}$$

Proof. See Appendix A. □

Plots of π_H compared to a standard normal and a standard Cauchy density can be seen in Figure 1. The next several sections will study this prior in detail, but its

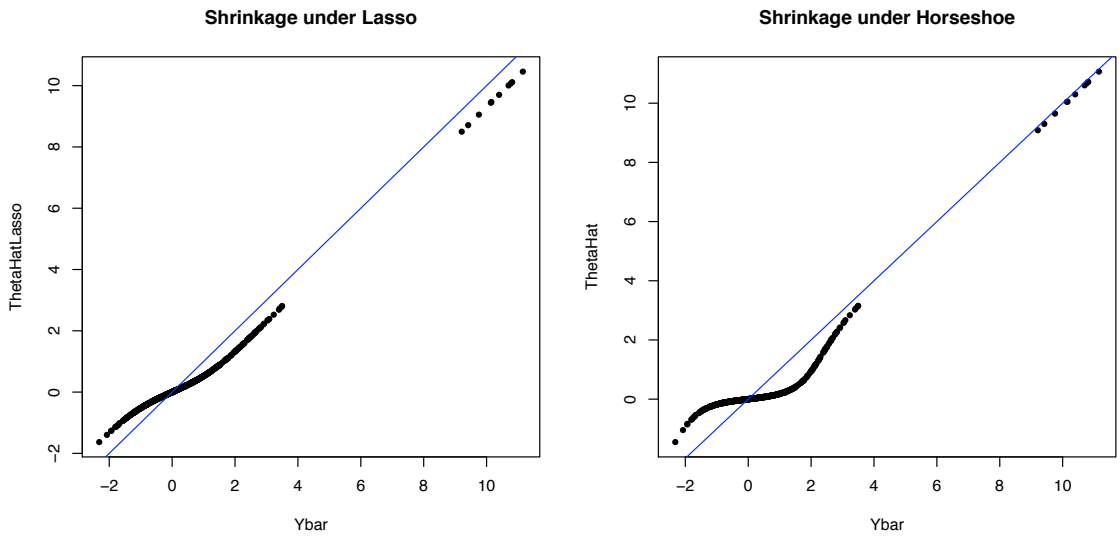

Figure 2: Plots of $\bar{y}_i$ versus $\hat{\theta}_i$ for Bayesian lasso (left) and horseshoe (right) priors on data where most of the means are zero. The diagonal lines are where $\hat{\theta}_i = \bar{y}_i$.

most interesting features are the following:

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

These features make $\pi_H(\theta)$ a useful shrinkage prior for sparse signals: its flat tails allow each θ_i to be large if the data warrant such a conclusion, and yet its infinitely tall spike at the origin means that the estimate can also be quite severely shrunk back to zero. Figure 2 gives an indication of the practical differences between using horseshoe priors and lasso/double-exponential priors in a situation where most of the means are zero save a handful of large signals. (A full discussion of this example can be found in Section 3.3.)

The focus of our paper is the posterior mean, which is widely known to be optimal under quadratic loss:

$$\hat{\theta}_i^H = \mathrm{E}_{\pi_H}(\theta_i \mid \mathbf{y}) = \int_{\mathbb{R}} \theta_i \; \pi_H(\theta_i \mid \mathbf{y}) \; \mathrm{d}\theta_i \,. \tag{4}$$

This can be thought of as a multivariate half-Cauchy mixture of ridge estimators.

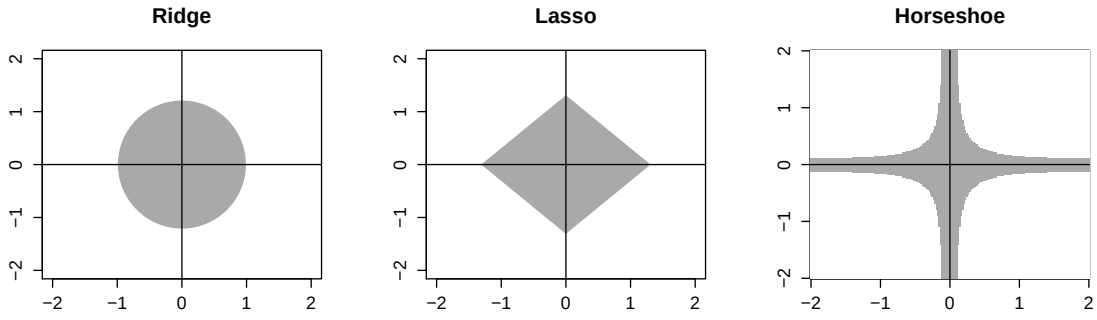

Figure 3: A comparison of the constraint regions implied by the ridge, lasso, and horseshoe penalty terms, where the posterior mode can be thought of as maximizing the likelihood over the points inside the shaded area.

An alternative horseshoe estimator arises from treating $\ln \pi_H$ as a regularization penalty and estimating $\boldsymbol{\theta}$ by the posterior mode. From Theorem 1.1, it follows that the horseshoe penalty is approximately

$$\ln \pi_H(\boldsymbol{\theta}) \approx \ln \sum_{i=1}^{n} \ln \left(1 + 2\theta_i^{-2}\right) , \tag{5}$$

which differs noticeably from other common penalty terms, including ridge (l^2), lasso (l^1), and bridge penalties (l^α). These penalized-likelihood estimators can be thought of as maximum-likelihood estimates subject to a constraint on the θ_i's; Figure 3 shows example constraint regions in two dimensions for the horseshoe, ridge, and lasso penalties, with the size of the constraint region depending upon the variability in the prior. See also Rissanen (1983), who describes the similar "universal prior" over the integers.

In all of our examples, we will estimate $\boldsymbol{\theta}$ using the posterior mean, though in principle the mode could also be used. We will focus on issues for which this distinction is not very important. While it is true that the mode can yields zeros under the right circumstances, the trustworthiness of these zeros can never be better than the trustworthiness of the underlying model for sparsity. Our focus is on showing the horseshoe to be a reasonable default model.

1.3 Outline of paper

In Section 2, we show that the horseshoe prior is just one member of a general class of distributions, the class of multivariate scale mixtures of normals. This class unites many seemingly different procedures, such as proper Bayes minimax estimation and

the Bayesian lasso, under a single umbrella. These local shrinkage rules are capable of generating a wide variety of possible estimators, and in Section 3, specific versions will be compared with two alternative classes of model-based procedures: global shrinkage rules and discrete mixture rules. This discussion will motivate our decision to single out the horseshoe prior for special consideration as a default "sparsity prior."

We then describe a set of detailed simulation studies that compare the risk properties of the horseshoe prior as an estimation procedure (Section 4) and as a classification procedure (Section 5). Section 6 then discusses Bayesian and empirical-Bayes approaches for handling two important features of the problem: the degree of sparsity in $\boldsymbol{\theta}$, and the scale of the non-sparse θ_i's compared to the error variance σ^2. We also note a surprising difference between the two approaches. Section 7 then concludes with a discussion of our results.

2 Model-based Rules for Shrinkage and Sparsity

2.1 Global shrinkage rules

There exists a vast, well-established body of literature on shrinkage estimation under normal error. The famous James-Stein estimator (James and Stein, 1961), for example, can be seen as a special case of a general family of global shrinkage rules of the form

$$E(\theta_i \mid \tau, \sigma, y_i) = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) y_i \,.$$

This is the result of assuming an exchangeable normal prior $\theta_i \sim \mathrm{N}(0, \tau^2)$, which we call a "global shrinkage rule." James and Stein choose τ by estimating $\sigma^2/(\sigma^2 + \tau^2)$ with an unbiased estimator; a more modern approach is to assume a prior distribution $\pi(\tau)$. The marginal prior for $\boldsymbol{\theta}$ is then

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \tau^{-n} \exp\left\{-\frac{\|\boldsymbol{\theta}\|^2}{2\tau^2}\right\} \, \pi(\tau) \, \mathrm{d}\tau \,, \tag{6}$$

from which a wide variety of Bayes and empirical-Bayes estimators can be constructed using different choices for $\pi(\tau)$.

Many different proposals have been made for this prior; some notable choices can be found in Tiao and Tan (1965), Stein (1981), Gelman (2006), and Scott and Berger (2006). We do not attempt a comprehensive review of global shrinkage rules, and instead refer the reader to Efron and Morris (1971) and Copas (1983), and to Section 6, where the specification of hyperparameters in the horseshoe model is considered more fully.

2.2 Discrete mixture rules

While global shrinkage rules are appealing due to their reduction in risk over the maximum-likelihood estimate, they are usually not appropriate for sparse signals, given that the goal is to shrink the noise components quite severely and the signals hardly at all.

Discrete mixture models provide a useful alternative. These models involve augmenting a global shrinkage rule with a point mass at $\theta_i = 0$, explicitly accounting for the presence of sparsity:

$$\theta_i \sim (1 - p)\delta_0 + p \cdot g(\theta_i)\,, \tag{7}$$

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

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

For these and other conjugate choices choices of g, it is straightforward to compute the posterior inclusion probabilities:

$$w_i = \Pr(\theta_i \neq 0 \mid \mathbf{y})\,.$$

These quantities will adapt to the level of sparsity in the data through shared dependence upon the unknown mixing probability p, yielding strong control over the number of false positive declarations. This effect can most easily be seen if one imagines testing a small number of signals in the presence of an increasingly large number of noise observations. As the noise comes to predominate, the posterior for p concentrates near 0, making it increasingly more difficult for any w_i to be large.

2.3 Local shrinkage rules

Under a discrete mixture model, the posterior mean of θ_i is

$$\mathrm{E}(\theta_i \mid \mathbf{y}) = w_i \cdot \mathrm{E}_g(\theta_i \mid \mathbf{y}, \theta_i \neq 0)\,, \tag{8}$$

an estimator that shrinks both globally through the nonzero mixture component g in (7), and locally through the inclusion probabilities w_i. An alternative line of thought suggests that, since this estimator will never be identically zero unless y_i is itself 0, the w_i terms should be modeled directly rather than derived through the sometimes-unwieldy task of model averaging under a discrete mixture (see, for example, Denison

and George, 2000).

One way of doing this is using the class of multivariate scale mixtures of normals with conditionally independent components:

$$\pi(\boldsymbol{\theta}) = (2\pi)^{-\frac{n}{2}} \int_{\Lambda,\tau} \tau^{-n} |\Lambda|^{-1/2} \exp\left\{ -\frac{\boldsymbol{\theta}^t \Lambda^{-1} \boldsymbol{\theta}}{2\tau^2} \right\} \, \pi(\tau) \, \pi(\Lambda) \, \mathrm{d}\tau \, \mathrm{d}\Lambda \,, \tag{9}$$

where $\Lambda = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_n^2)$. We call these "local shrinkage rules," since the parameter τ can be interpreted as a global shrinkage parameter and the λ_i's as a set of local shrinkage parameters.

Denote by $\mathcal{M}$ the class of priors expressible as in (9) for proper $\pi(\tau)$ and $\pi(\Lambda) = \pi(\lambda_1) \cdots \pi(\lambda_n)$. The horseshoe prior of Section 1.2 is clearly a member of $\mathcal{M}$, since its local shrinkage parameters follow independent half-Cauchy distributions.

This class unites a wide variety of proposed shrinkage rules under a single family. Some of these rules aim at capturing sparsity with a one-group model, the most famous of which is undoubtedly the lasso; see Tibshirani (1996) for a description of the classical lasso, along with Park and Casella (2008) and Hans (2008) for Bayesian versions that use the posterior mean rather than the mode. Under the lasso, the prior for each λ_i expressible in terms of an independent positive-stable random variable (West, 1987). Other similar mixtures can yield Bayesian versions of the entire class of l^α penalized-likelihood estimators for $0 < \alpha \leq 2$. In linear regression, these are called bridge estimators (Fu, 1998) when τ is fixed and the posterior mode, rather than the mean, is used.

Other members of class $\mathcal{M}$ include multivariate Cauchy priors useful in robust shrinkage estimation (Angers and Berger, 1991); the class of Strawderman–Berger priors, which have heavy tails and yield minimax estimators under quadratic loss (Strawderman, 1971; Berger, 1980); the generalized ridge estimators of Denison and George (2000); and the class of normal–exponential–gamma models described in Griffin and Brown (2005).

3 Detecting Sparse Signals with the Horseshoe

3.1 Overview: one-group and two-group models

We believe that the two-group mixture of Equation (7) is, in some sense, the "right model" for the two-group question intrinsic to sparse situations. To borrow the language of Johnstone and Silverman (2004), the question is one of identifying needles and straws in haystacks. The two-group model captures the notion of strong sparsity quite directly, but is also a good approximation in weakly sparse situations as long as a simple criterion is met regarding the size, in units of σ, of the "nearly zero" elements (Berger and Delampady, 1987). Discrete mixtures have well-understood theoretical properties, and they tend to work well in realistic situations with hybrid loss

functions—that is, where there is interest both in detecting signals and in estimating the size of nonzero signals under, for example, squared-error loss.

But discrete mixtures present computational difficulties. The choice of g is severely restricted due to the need for conjugacy, and the challenge of exploring a discrete model space that grows exponentially in nonorthogonal situations (like regression) can be overwhelming. Non-Bayesians are also reluctant to make the parametric assumptions necessary to fit such a model.

One-group, lasso-type estimators avoid these difficulties while also aiming at the same goal of simultaneous selection and estimation. Yet there is no agreement among Bayesians and non-Bayesians about the right way to proceed here, either. Bayesians observe that the lasso solution produces zeros in $\boldsymbol{\theta}$ as a mere side effect of using the posterior mode—clearly a poor estimator under squared-error loss—and not as a probabilistic statement about inclusion (Park and Casella, 2008). There is also research to suggest that the zeros so induced may not be the same zeros one would get from a full variable-selection approach (Hans, 2008).

Yet the Bayesian-lasso approach of using the posterior mean rather than the mode produces no zeros at all, and so ignores a crucial element of inference in sparse settings: that is, sparsity itself. And even if estimation is the only goal, there is no reason to expect that a lasso-type prior can replicate the local shrinkage properties of a discrete mixture, which estimates the w_i's of Equation (8) in a highly structured way.

Our goal is to show that the horseshoe prior introduced in Section 1.2 can offer a true one-group answer to the two-group question of sparsity:

- Like the classical lasso and the mixture model (but unlike the Bayesian lasso), the horseshoe can accurately classify zeros in $\boldsymbol{\theta}$ once an appropriate threshold is applied. Remarkably, as Section 5 will show, these zeros are nearly identical to those produced by the Bayesian discrete mixture model, in spite of the very different approaches used to find these zeros.

- Like the Bayesian lasso (but unlike the classical lasso), the horseshoe estimator uses the posterior mean rather than the mode to estimate $\boldsymbol{\theta}$, and will consequently do better under squared-error loss. It also allows for the full posterior distribution of $\boldsymbol{\theta}$ to be assessed.

- Like the Bayesian and classical versions of the lasso (but unlike the discrete mixture), the horseshoe does not require computing marginal likelihoods or searching a large discrete model space.

3.2 The motivation for the horseshoe prior

We now attempt to provide some intuition as to why the horseshoe is an appropriate defaulty sparsity prior.

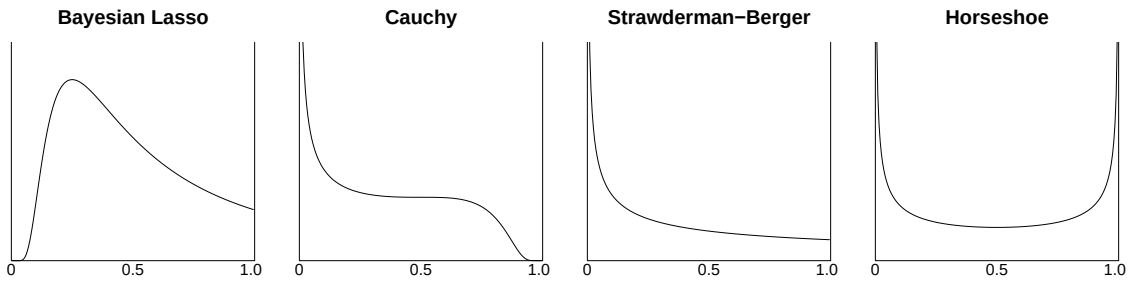

Figure 4: A comparison of the implied density for the shrinkage weights $\kappa_i \in [0, 1]$ for four different priors, where $\kappa_i = 0$ means no shrinkage and $\kappa_i = 1$ means total shrinkage to zero.

Recall that in a multivariate scale mixture, λ_i is the random scale parameter in the normal prior for each θ_i. As a function of λ_i, the Bayes estimator for θ_i under quadratic loss is

$$\hat{\theta}_i(\lambda_i) = \mathrm{E}(\theta_i \mid \mathbf{y}, \lambda_i) = \left(1 - \frac{1}{1+\lambda_i^2}\right) y_i \,,$$

where τ and σ^2 are fixed at 1. Hence by Fubini's theorem,

$$\hat{\theta}_i = \mathrm{E}(\theta_i \mid \mathbf{y}) = \int_0^\infty \left(1 - \frac{1}{1+\lambda_i^2}\right) y_i \; \pi(\lambda_i \mid \mathbf{y}) \; \mathrm{d}\lambda_i = 1 - \mathrm{E}\left(\frac{1}{1+\lambda_i^2} \mid \mathbf{y}\right) \cdot y_i \,.$$

Thus $\kappa_i = 1/(1+\lambda_i^2)$, which lies on $[0, 1]$, has an interpretation as a random shrinkage parameter, with $1 - \kappa_i$ playing a role similar to that of the inclusion probability w_i in the posterior mean under a discrete mixture (8).

We call κ_i the "shrinkage weight", and its opposite $w_i = 1 - \kappa_i$ the "significance weight." Small *a priori* values of κ_i yield $\hat{\theta}_i \approx y_i$ and hence mean high significance/weak shrinkage; this is good for identifying needles in haystacks. Large values allow $\hat{\theta}_i \approx 0$ and hence mean low significance/strong shrinkage; this is good for identifying straw. To identify both needles and straw, the implied prior for κ_i must allow values both very near 0 and very near 1.

For this reason, it is often more intuitive to interpret different priors for λ_i, along with the estimators they give rise to, in terms of the priors they imply for the shrinkage weight κ_i (or equivalently to the significance weight w_i). This will provide crucial insight as to whether a given prior for λ_i will yield an estimator that is robust both to outliers and to false positives.

Table 1 gives four examples: the lasso/double-exponential prior, the Strawderman–Berger prior, the Cauchy prior, and the horseshoe prior. Figure 4 plots these four densities for κ_i. Notice that at $\kappa_i = 0$, both $\pi_C(\kappa_i)$ and $\pi_{SB}(\kappa_i)$ are both unbounded, while $\pi_{BL}(\kappa_i)$ vanishes. This suggests that Cauchy and Strawderman–Berger estima-

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

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

tors will be good, and lasso estimators (both Bayesian and classical versions) will be bad, in terms of outlier robustness—that is, at leaving large signals unshrunk.

Meanwhile, at $\kappa_i = 1$, π_C tends to zero, while both π_{BL} and π_{SB} tend to fixed constants. This suggests that Cauchy estimators will be bad, while Bayesian lasso and Strawderman–Berger estimators will be mediocre, in terms of Type-I-error robustness—that is, at correctly shrinking noise all the way to 0.

What kind of prior for κ_i does the horseshoe prior $\pi_H(\theta)$ imply? Since $\pi(\lambda) \propto 1/(1+\lambda^2)$ is the mixing density giving rise to π_H, it is easily shown that

$$\pi_H(\kappa_i) \propto \kappa_i^{-1/2}(1-\kappa_i)^{-1/2}\,, \tag{10}$$

a Beta$(1/2, 1/2)$ distribution as indicated in Table 1.

The shape of this implied density for κ_i is why we call $\pi_H(\theta_i)$ the "horseshoe prior." It is unbounded both at $\kappa_i = 0$ and $\kappa_i = 1$, suggesting that π_H will be robust in both senses: large outlying θ_i's $\kappa_i \approx 0$ and will not be shrunk, but the remaining θ_i's will have $\kappa_i \approx 1$ *a posteriori* and can thus be shrunk almost all the way to 0. At the same time, $\pi_H(\kappa_i)$ does not bottom out too rapidly for intermediate values, placing one-third of its mass from $\kappa_i = 0.25$ to $\kappa_i = 0.75$ and thus allowing moderate shrinkage of each θ_i if the data warrant it.

Meanwhile, the global shrinkage parameter τ controls the overall degree of sparsity in $\boldsymbol{\theta}$. In all of the examples that we have investigated, this combination of strong global shrinkage with robust local shrinkage has proven very effective.

3.3 An illustrative example

An example will help illustrate these ideas. Two standard normal observations were simulated for each of 1000 means: 10 signals of mean 10, 90 signals of mean 2, and 900 noise of mean 0. To this data set we then fit two models, one that used independent horseshoe priors for each θ_i, and one that used independent double-exponential priors (giving the Bayesian lasso solution). For both models, a half-Cauchy prior was used for the global scale parameter τ following the advice of Gelman (2006) and Scott and

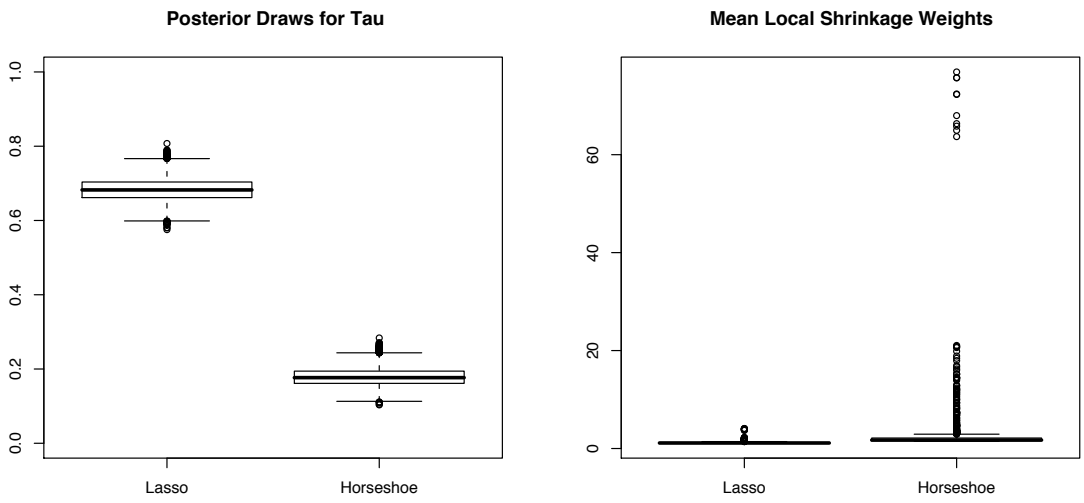

Figure 5: Left: posterior draws for the global shrinkage parameter τ under both lasso and horseshoe for the toy example. Right: boxplot of $\hat{\lambda}_i$'s, the posterior means of the local shrinkage parameters λ_i.

Berger (2006), and Jeffreys' prior $\pi(\sigma) \propto 1/\sigma$ was used for the error variance.

The shrinkage characteristics of these two fits are summarized in Figure 2 from the introduction. These plots show the posterior mean $\hat{\theta}_i^{BL}$ and $\hat{\theta}_i^{H}$ as a function of the observed data $\bar{y}_i$, with the diagonal lines showing where $\hat{\theta}_i = \bar{y}_i$. Key differences occur near $\bar{y}_i \approx 0$ (the noise observations) and $\bar{y}_i \approx 10$ (the ten large signals). Compared with the horseshoe prior, the double-exponential prior tends to shrink small observations not enough, and the large observations too much.

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

The horseshoe prior clearly does better at handling both aspects of the problem: leaving large signals unshrunk while squelching most of the noise. This is reflected in their relative mean squared-error—almost 25% lower under the horseshoe model.

3.4 Comparison with discrete mixture models

The discrete mixture can be thought of as adding a point mass at $\kappa_i = 1$, allowing total shrinkage. In broad terms, this is much like the unbounded density under the

horseshoe prior as $\kappa_i \to 1$, suggesting that these models may be similar in the degree to which they shrink noise variables to 0.

It is therefore interesting to consider the differences between shrinkage profiles of the horseshoe prior (π_H) and the heavy-tailed discrete mixture using Strawderman–Berger priors (π_{DM}). These differences are easiest to understand when considering scaled versions of the priors :

$$\pi_H : \ (\theta_i \mid \kappa_i) \sim \mathrm{N}(0, \tau^2(\kappa_i^{-1} - 1)) \text{ with } \kappa_i \sim \mathrm{Be}(1/2, 1/2) \tag{11}$$

$$\pi_{DM} : \ (\theta_i \mid \kappa_i) \sim (1-p) \cdot \delta_0 + p \cdot \mathrm{N}\left(0, \frac{\tau^2 + \sigma^2}{2\kappa_i} - \sigma^2\right) \text{ with } \kappa_i \sim \mathrm{Be}(1/2, 1)\,. \tag{12}$$

The Strawderman–Berger prior has a number of desirable properties for describing the nonzero θ_i's, since it is both heavy-tailed and yet still allows closed-form convolution with the normal likelihood. Both the horseshoe and the discrete mixture have global scale parameters τ and local shrinkage weights κ_i. Yet τ plays a very different role in each model (aside from the fact that the Strawderman–Berger prior is only defined for $\tau > \sigma$). Under the horseshoe prior,

$$\mathrm{E}_H(\theta_i | \kappa_i, \tau, y_i) = \left(1 - \frac{\kappa_i}{\kappa_i + \tau^2(1-\kappa_i)}\right) y_i\,, \tag{13}$$

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

$$\mathrm{E}_{DM}(\theta_i | \kappa_i, w_i, y_i) = w_i \left(1 - \frac{2\kappa_i}{1+\tau^2}\right) y_i\,, \tag{14}$$

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

$$\frac{w_i}{1-w_i} = \frac{p \cdot G^*(y_i \mid \tau)}{(1-p) \cdot \mathrm{N}(y_i \mid 0, 1)}\,. \tag{15}$$

Several differences between the approaches are apparent:

- In the discrete model, local shrinkage is controlled by the Bayes factor in (15), which is a function of the signal-to-noise ratio τ/σ. In the scale-mixture model, local shrinkage is determined entirely by the κ_i's (or equivalently the λ_i's).

- In the discrete model, global shrinkage is primarily controlled through the prior inclusion probability p. In the scale-mixture model, global shrinkage is controlled by τ, since if τ is small in (13), then κ_i must be very close to 0 for extreme shrinkage to be avoided. Hence in the former model, p adapts to the overall sparsity of $\boldsymbol{\theta}$, while in the latter model, τ performs this role.

- There will be a strong interaction between p and τ in the discrete model which

is not present in the scale-mixture model. Intuitively, as p changes, τ must adapt to the scale of the observations that are reclassified as signals or noise.

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

Indeed, if the discrete mixture model is a way of arriving at a good shrinkage estimator by way of a multiple-testing procedure, then our horseshoe estimator goes in the opposite direction—arriving at a good multiple-testing procedure by way of a shrinkage estimator.

4 Estimation Risk

4.1 Overview

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

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

For both the Bayesian lasso and the horseshoe, the following default hyperpriors were used in all studies:

$$\pi(\sigma) \propto 1/\sigma \tag{16}$$

$$(\tau \mid \sigma) \sim \mathrm{C}^{+}(0, \sigma)\,. \tag{17}$$

A slight modification is necessary in the fully Bayesian discrete-mixture model, since

under the Strawderman–Berger prior of (12), τ must be larger than σ:

$$p \sim \text{Unif}(0,1) \tag{18}$$
$$\pi(\sigma) \propto 1/\sigma \tag{19}$$
$$(\tau \mid \sigma) \sim \text{C}(\sigma,\sigma) \cdot \mathbf{1}_{\tau \geq \sigma}\,. \tag{20}$$

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

Experiment 1: Strongly sparse signals

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

$$y_i \sim \text{N}(\theta_i, \sigma^2) \tag{21}$$
$$\theta_i \sim p \cdot t_3(0,\tau) + (1-p) \cdot \delta_0 \tag{22}$$
$$p \sim \text{Be}(1,4)\,, \tag{23}$$

where the nonzero θ_i's follow a t distribution with 3 degrees of freedom, and where $\boldsymbol{\theta}$ has 20% nonzero entries on average.

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

Experiment 2: Weakly sparse signals

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

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

$$y_i \sim \text{N}(\theta_i, \sigma^2) \tag{24}$$
$$(\theta_i \mid \eta, \alpha) \sim \text{Unif}(-\eta c_i, \eta c_i) \tag{25}$$
$$\eta \sim \text{Ex}(2) \tag{26}$$
$$\alpha \sim \text{Unif}(a,b)\,, \tag{27}$$

for $c_i = n^{1/\alpha} \cdot i^{-1/\alpha}$ for $i = 1, \ldots, n$. These $\boldsymbol{\theta}$'s correspond to a weak-l^α bound on the coefficients, as described by Johnstone and Silverman (2004): the ordered θ_i's

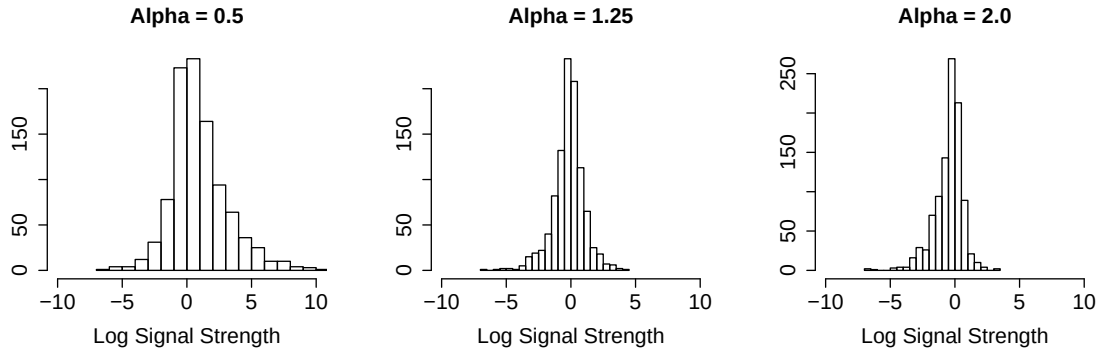

Figure 6: Log signal strength for three weak-l^α vectors of 1000 means, with α at 0.5, 1.25, and 2.0.

follow a power-law decay, with the size of the largest coefficients controlled by the exponentially distributed random bound η, and the speed of the decay controlled by the random norm α.

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

4.2 Results

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

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

Close inspection of some of these simulated data sets led us to conclude that the double-exponential prior loses on both ends here, much as it did in the toy example summarized in Figures 2 and 5. It lacks tails that are heavy enough to robustly estimate the large t_3 signals, and it also lacks sufficient mass near 0 to adequately squelch the substantial noise in $\boldsymbol{\theta}$.

These problems, particularly the tail robustness, also plague the classical lasso. In our experience, the issue is not whether the mean or mode is used, but with the

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

Table 2: Risk under squared-error (SE) loss and absolute-error (AE) loss in experiment 1. Bold diagonal entries are median sum of squared-errors (top half) and absolute errors (bottom half) in 1000 simulated data sets. Off-diagonal entries are average risk ratios in units of σ (risk of row divided by risk of column). BL: Bayesian lasso/double exponential. HS: horseshoe. DMF: discrete mixture, fully Bayes. DME: discrete mixture, empirical Bayes.

underlying model for sparsity as described by the κ densities in Figure 4. Like any shrinkage estimator, the lasso model is fine when its prior describes reality, but suffers in problems that correspond to very reasonable, commonly held notions of sparsity. The horseshoe, on the other hand, allows each shrinkage coefficient κ_i to be arbitrarily close to 0 or 1, and hence can accurately estimate both signal and noise.

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

By and large, the horseshoe estimator, despite providing a one-group answer, acts just like a model-averaged Bayes estimator arising from a two-group mixture. The lasso does not.

5 Classification Risk

5.1 Overview

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

Recall that under the under the discrete mixture model of Equation (7), the

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

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

Bayes estimator for each θ_i is $\hat{\theta}_i^{DM} = w_i E_g(\theta_i \mid y_i)$, where w_i is the posterior inclusion probability for θ_i. For appropriately heavy-tailed g such as the Strawderman–Berger prior of Section 3.4, this expression is approximately $w_i y_i$, meaning that w_i can be construed in two different ways:

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

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

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

We will study two different asympotic scenarios: fixed-k asymptotics, and what we call "ideal signal-recovery" asymptotics. As before, we will benchmark the horseshoe against the Bayesian lasso.

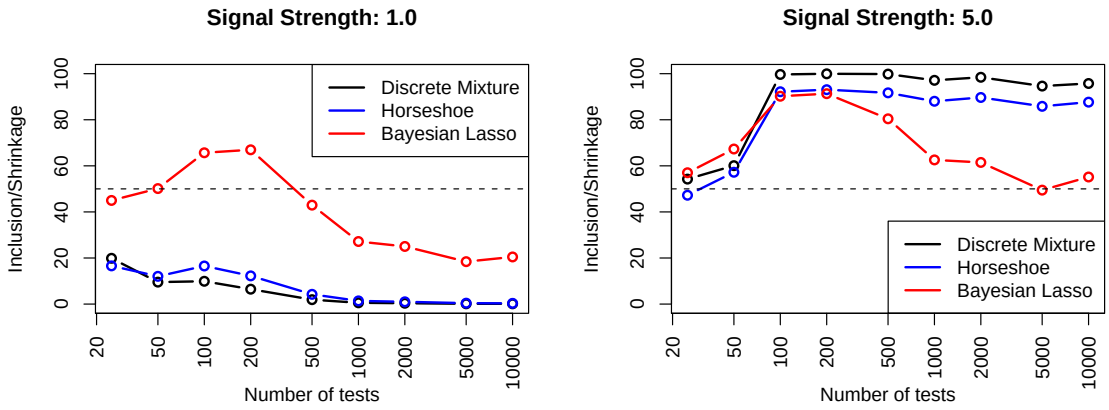

Figure 7: Inclusion probabilities/significance weights w_i versus number of tests n in Experiment 3 (classification under fixed-k asymptotics). Shrinkage weights near 100 indicate Black = discrete mixture; blue = horseshoe; red = lasso.

Experiment 3: Fixed-k asymptotics

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

Experiment 4: Ideal signal-recovery asymptotics

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

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

$$s \to \infty \tag{28}$$

$$p \to 0 \tag{29}$$

$$s^{-1} \log\left(\frac{1-p}{p}\right) \to C \quad \text{for some} \quad 0 < C < \infty\,, \tag{30}$$

then as the number of tests n grows without bound, the probability of a false positive

$(w_i \geq 0.5, \theta_i = 0)$ under converges to 0, while the probability of a false negative $(w_i < 0.5, \theta_i \neq 0)$ converges to a fixed constant less than one, the exact value of which will depend upon C. (These results are under the fully Bayesian discrete mixture model with a normal prior for each θ_i, but the same logic holds for the heavy-tailed model.)

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

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

$$\sigma_n^2 = \frac{D}{\log\left(\frac{n-k}{k}\right)}, \tag{31}$$

where D is any constant (in this case 0.4), n is the total number of tests being performed, and k is the fixed number of true signals (in this case 10). It is easy to show that as $n \to \infty$, the conditions of Equations (28)–(30) will hold, and the results of Bogdan et al. (2008a) will obtain, if for each sample size the noise variance is σ_n^2.

5.2 Results

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

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

Comprehensive results for Experiments 3 and 4 are given in Tables 4–9. A close inspection of these numbers bears out the story of Figure 7: regardless of the asymptotic regime and regardless of the signal strength, the horseshoe significance weights are a good stand-in for the posterior inclusion probabilities under the discrete mixture. They lead to nearly identifical numerical summaries of the strength of evidence in the data, and nearly identifical classifications of signal versus noise (except in very low-data cases under ideal signal-recovery asymptotics).

What's more, the horseshoe thresholding rule is also quite accurate in its own right. As the tables show, it exhibits very strong control over the number of false

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

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

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

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

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

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

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

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

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

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

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

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

positive declarations while retaining a reasonable amount of power, even under fixed-k asymptotics.

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

6 Estimation of Hyperparameters

6.1 Bayes, empirical Bayes, and cross-validation

This section discusses the estimation of key model hyperparameters.

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

Empirical Bayes is a particularly unattractive option under the horseshoe prior, for two main reasons:

1. The marginal prior for θ_i involves the exponential integral function, which is difficult to evaluate accurately. Hence the marginal likelihood for τ and σ cannot be maximized easily. Note that introducing the λ_i's makes everything conditionally normal, but at the expense of one nuisance parameter for every component of $\boldsymbol{\theta}$.

2. In the horseshoe model, τ acts as a global parameter for multiplicity control. When few signals are present, it is quite common for the posterior mass of τ to concentrate near 0 and for the signals to be flagged via large values of the local shrinkage parameters λ_i. Indeed, strong global shrinkage combined with robust local shrinkage is why the horseshoe works so well. This means that the marginal maximum-likelihood solution is always in danger of collapsing to the degenerate $\hat{\tau} = 0$ (see, for example, Tiao and Tan, 1965).

Cross-validation will not suffer from either of these difficulties, but still involves plugging in a point estimate for the signal-to-noise ratio. This can be perilous, given that σ and τ will typically have an unknown correlation structure that should ideally be averaged over.

Indeed, as the following example serves to illustrate, careful handling of uncertainty in the joint distribution for τ and σ can be crucial.

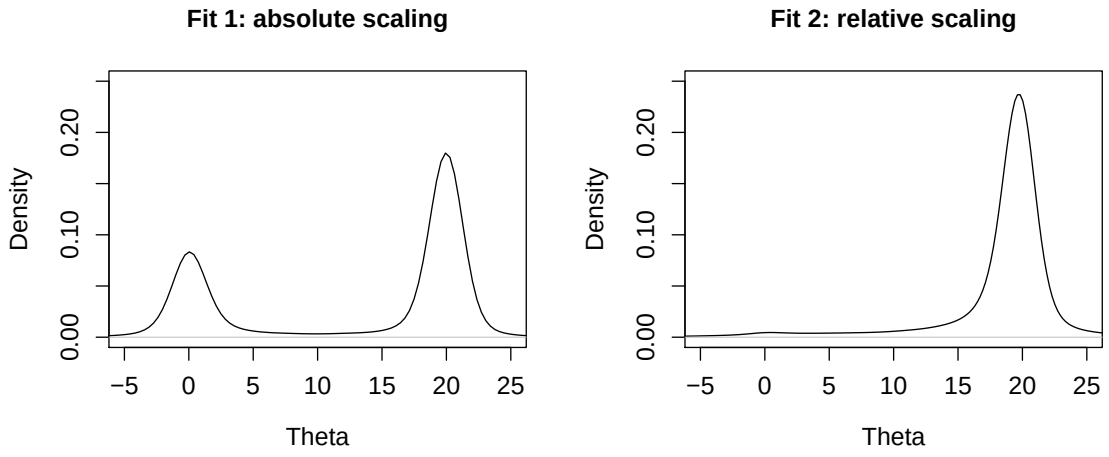

Figure 8: Example 1. Left: the posterior for θ when $\tau \sim C^+(0, 1)$. Right: the posterior when $\tau \sim \mathrm{C}^+(0, \sigma)$.

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

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

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

6.2 A notable discrepency under the lasso

Scott and Berger (2008) give a detailed comparison of Bayes and empirical-Bayes approaches for handling p, the prior inclusion probability, in the context of discrete

mixture models for variable selection. Since in the horseshoe model, τ plays the role of p in exercising multiplicity control, an analogous set of issues surely arises here.

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

Suppose that in the basic normal-means problem, a lasso-type prior of unknown center and scale is used: $\theta_i = \mu + \tau\eta_i$, where $\eta_i \sim \mathrm{DE}(2)$ has a double-exponential distribution. Hence

$$\pi(\boldsymbol{\theta} \mid \mu, \tau) \propto \tau^{-n} \exp\left(-\frac{1}{\tau}\sum_{i=1}^{n}|\theta_i - \mu|\right), \tag{32}$$

leading to the joint distribution

$$p(\boldsymbol{\theta}, \mathbf{y} \mid \mu, \nu) \propto \nu^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \theta_i)^2 + \nu^{-1}\sum_{i=1}^{p}|\theta_i - \mu|\right)\right\}, \tag{33}$$

where $\nu^{-1} = \sigma/\tau$ is the regularization penalty (for known σ).

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

$$\pi(\boldsymbol{\theta}) = \int \pi(\boldsymbol{\theta}|\mu, \tau)\ \pi(\mu, \tau)\ \mathrm{d}\mu\ \mathrm{d}\tau \tag{34}$$

$$\propto \exp\left\{-\frac{1}{2}Q(\boldsymbol{\theta})\right\}, \tag{35}$$

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

$$\pi(\boldsymbol{\theta}) = (n-2)!\ 2^{-n+1}\sum_{j=1}^{n} w_j^{-1}, \tag{36}$$

where

$$w_j = \begin{cases} 4v_j^{n-1}\left(j - \frac{n}{2}\right)\left(\frac{n}{2} + 1 - j\right), & j \neq \frac{n}{2}, \frac{n}{2}+1 \\ 4v_j^{n-1}\left[1 + (n-1)\left(\theta_{(n/2+1)} - \theta_{(n/2)}\right)v_j^{-1}\right], & j = \frac{n}{2}, \frac{n}{2}+1 \end{cases}. \tag{37}$$

Hence the non-Bayesian estimates $\boldsymbol{\theta}$ using

$$\pi_{EB}(\boldsymbol{\theta} \mid \mathbf{y}, \hat{\nu}, \hat{\mu}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\theta_i)^2 + \hat{\nu}^{-1}\sum_{i=1}^{p}|\theta_i-\hat{\mu}|\right)\right\}, \qquad (38)$$

while the Bayesian estimates $\boldsymbol{\theta}$ using

$$\pi_{FB}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i-\theta_i)^2\right) + \frac{(n-2)!}{2^{n-1}}\log\left(\sum_{i=1}^{n}[w_i(\boldsymbol{\theta})]^{-1}\right)\right\}. \qquad (39)$$

The former is the traditional l^1 penalty, while the latter exhibits a rather complicated dependence upon the order statistics of the θ_i's (which do not appear in the plug-in expression). It is by no means certain that the two procedures will reach similar answers asymptotically, since this difference in functional form persists for all n.

The lasso/l^1 penalty coupled with the noninformative prior on μ and τ is just one example where the marginalization in (36) is analytically tractable. But it serves to convey the essence of the problem, which is quite general. The Bayes and plug-in approaches for estimating τ imply fundamentally different regularization penalties for $\boldsymbol{\theta}$, regardless of whether $\boldsymbol{\theta}$ is estimated by the mean or the mode, and regardless of whether marginal maximum likelihood or cross-validation is used.

Of course, neither penalty is wrong *per se*, but the stark difference between (38) and (39) is interesting in its own right—particularly given the popularity of using cross-validation to fit the lasso—and also calls into question the extent to which the plug-in analysis can approximate the fully Bayesian one. While some practitioners may have different goals for empirical Bayes or cross-validation, such comparison is at least reasonable. Many Bayesians use empirical-Bayes as a computational simplification, and many non-Bayesians appeal to complete-class theorems that rely upon an empirical-Bayes procedure's asymptotic correspondence with a fully Bayesian procedure. Hence questions about where the two approaches agree, and where they disagree, is of interest both to Bayesians and non-Bayesians.

These questions are particularly important for the models considered in this paper, which rely upon only a handle of parameters such as τ and p to do the hard work of adapting to unknown sparsity in high-dimensional data, and where small differences can disproportionately impact results. While we do not wish to argue that plugging in point estimates for these parameters is necessarily wrong, we feel that in light of these issues, it should be done only with great care.

7 Discussion

We do not claim that the horseshoe is a panacea for sparse problems—merely a good default option. It is both surprising and interesting that its answers coincide nearly

perfectly with the answers from the two-group "gold-standard" of a Bayesian mixture model. Equally surprising is that the answers under double-exponential priors do not.

To be sure, all Bayes estimators are necessarily optimal with respect to their assumed priors, and so different procedures will work best in different situations of sparsity. But it is one thing to use a prior like the lasso when its assumptions are likely to be accurate, and quite another to recommend it as an automatic procedure. To put it bluntly, the double-exponential prior simply isn't a good model for sparsity. It lacks both ingredients that make the horseshoe effective: strong global shrinkage through τ, and robust local shrinkage through the λ_i's. This reflects a fundamental deficit in the model that simply cannot be repaired by using to the mode to get zeros, or by cross-validating to choose the amount of shrinkage.

References

J. Angers and J. Berger. Robust hierarchical bayes estimation of exchangeable means. *The Canadian Journal of Statistics*, 19(1):39–56, 1991.

J. O. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8(4):716–761, 1980.

J. O. Berger and M. Delampady. Testing precise hypotheses. *Statistical Science*, 2 (3):317–52, 1987.

D. Berry. Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian Statistics 3*, pages 79–94. Oxford University Press, 1988.

M. Bogdan, A. Chakrabarti, and J. K. Ghosh. Optimal rules for multiple testing and sparse multiple regression. Technical report, Purdue University, 2008a.

M. Bogdan, J. K. Ghosh, and S. T. Tokdar. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1, pages 211–30. Institute of Mathematical Statistics, 2008b.

L. Brown. Admissible estimators, recurrent diffusions and insoluble boundary problems. *The Annals of Mathematical Statistics*, 42:855–903, 1971.

J. Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society, Series B*, 45(3):311–54, 1983.

D. Denison and E. George. Bayesian prediction using adaptive ridge estimators. Technical report, Imperial College, London, 2000.

B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008.

B. Efron and C. Morris. Limiting the risk of Bayes and empirical Bayes—part i: the Bayes case. *Journal of the American Statistical Association*, 66:807–815, 1971.

T. Fan and J. O. Berger. Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions. *Stat. Decisions*, 10:99–120, 1992.

W. J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–33, 2006.

E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

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.

W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–79, 1961.

I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–36, 1988.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–6, 2008.

J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2), 1983.

J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Discussion Paper 2008-10, Duke University Department of Statistical Science, 2008.

C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–51, 1981.

W. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Statistics*, 42:385–8, 1971.

G. C. Tiao and W. Tan. Bayesian analysis fo random-effect models in the analysis of variance. i. Posterior distribution of variance components. *Biometrika*, 51:37–53, 1965.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–88, 1996.

V. Uthoff. The most powerful scale and location invariant test of the normal versus the double exponential. *The Annals of Statistics*, 1(1):170–4, 1973.

M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–8, 1987.

A Proofs

Proof of Theorem 1.1. Clearly,

$$\pi_H(\theta) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left\{\frac{-\theta^2}{2\lambda^2}\right\} \frac{2}{\pi(1+\lambda^2)}\, \mathrm{d}\lambda\,. \tag{40}$$

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

$$\pi_H(\theta) = K \int_0^\infty \frac{1}{1+u} \exp\left\{-\frac{\theta^2 u}{2}\right\}\, \mathrm{d}u\,, \tag{41}$$

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

$$\pi_H(\theta) = K e^{\theta^2/2} \int_1^\infty \frac{1}{z} e^{-z\theta^2/2}\, \mathrm{d}z \tag{42}$$

$$= K e^{\theta^2/2} \cdot E_1(\theta^2/2)\,, \tag{43}$$

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:

$$\frac{e^{-t}}{2} \log\left(1 + \frac{2}{t}\right) < E_1(t) < e^{-t} \log\left(1 + \frac{1}{t}\right) \tag{44}$$

for all $t > 0$, which proves Part (b). Part (a) then follows from the lower bound in Equation (3), which approaches ∞ as $\theta \to 0$. □