Handling Sparsity via the Horseshoe

Carlos M. Carvalho
Booth School of Business
The University of Chicago
Chicago, IL 60637

Nicholas G. Polson
Booth School of Business
The University of Chicago
Chicago, IL 60637

James G. Scott
McCombs School of Business
The University of Texas
Austin, TX 78712

Abstract

This paper presents a general, fully Bayesian framework for sparse supervised-learning problems based on the horseshoe prior. The horseshoe prior is a member of the family of multivariate scale mixtures of normals, and is therefore closely related to widely used approaches for sparse Bayesian learning, including, among others, Laplacian priors (e.g. the LASSO) and Student-t priors (e.g. the relevance vector machine). The advantages of the horseshoe are its robustness at handling unknown sparsity and large outlying signals. These properties are justified theoretically via a representation theorem and accompanied by comprehensive empirical experiments that compare its performance to benchmark alternatives.

1 Introduction

Supervised Learning can be cast as the problem of estimating a set of coefficients $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^p$ that determine some functional relationship between a set of inputs $\{x_i\}_{i=1}^p$ and a target variable y. This framework, while simple, is of central focus in modern statistics and artificial-intelligence research; it encompasses problems of regression, classification, function estimation, covariance regularization, and others still. The main challenges arise in "large-p" problems where, in order to avoid overly complex models that will predict poorly, some form of dimensionality reduction is needed. This entails finding sparse solutions, where some of the elements β_i are zero (or very small).

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

From a Bayesian-learning perspective, there are two main sparse-estimation alternatives: discrete mixtures and shrinkage priors. The first approach (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) models each β_i with a prior comprising both a point mass at $\beta_i = 0$ and an absolutely continuous alternative; the second approach (see, e.g., Tibshirani, 1996 and Tipping, 2001) models the β_i's with absolutely continuous "shrinkage" priors centered at zero.

The choice of one approach or the other involves a series of tradeoffs. Discrete mixtures offer the correct representation of sparse problems by placing positive prior probability on $\beta_i = 0$, but pose several difficulties. These include foundational issues related to the specification of priors for trans-dimensional model comparison, and computational issues related both to the calculation of marginal likelihoods and to the rapid combinatorial growth of the solution set. Shrinkage priors, on the other hand, can be very attractive computationally. But they create their own set of challenges, since the posterior probability mass on $\{\beta_i = 0\}$ (a set of Lebesgue measure zero) is never positive. Truly sparse solutions can therefore be achieved only through artifice.

In this paper we adopt the shrinkage approach, while at the same time acknowledging the discrete-mixture approach as a methodological ideal. Indeed, it is with this ideal in mind that describe the horseshoe prior (Carvalho, Polson and Scott, 2008) as a default choice for shrinkage in the presence of sparsity.

We begin our discussion in the simple situation where $\boldsymbol{\beta}$ is a vector of normal means, since it is here that the lessons drawn from a comparison of different shrinkage approaches for modeling sparsity are most readily understood. In this context, we provide a theoretical characterization of the robustness properties of the horseshoe via a representation theorem for the posterior mean of $\boldsymbol{\beta}$, given data $\boldsymbol{y}$. We then give a handful of examples of the horseshoe's performance in linear models, function estimation, and covariance regular-

ization (a problem of unsupervised learning for which the horseshoe prior is still highly relevant).

Our goal is not to characterize the horseshoe estimator as a "cure-all"—merely a default procedure that is well-behaved, that is computationally tractable, and that seems to outperform its competitors in a wide variety of sparse situations. We also try to provide some intuition as to the nature of this advantage: namely, the horseshoe prior's ability to adapt to different sparsity patterns while simultaneously avoiding the over-shrinkage of large coefficients.

Finally, we will return several times to a happy, and remarkably consistent, fact about the horseshoe's performance: that it quite closely mimics the answers one would get by performing Bayesian model-averaging, or BMA, under a heavy-tailed discrete-mixture model. Bayesian model averaging is clearly the predictive gold standard for such problems (see, e.g., Hoeting *et al*, 1999), and a large part of the horseshoe prior's appeal stems from its ability to provide "BMA-like" performance without the attendant computational fuss.

2 The Horseshoe Prior

We start by introducing our approach to sparsity in the simple, stylized situation where $(\boldsymbol{y}|\boldsymbol{\beta}) \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2 I)$, and where $\boldsymbol{\beta}$ is believed to be sparse.

The horseshoe prior assumes that each β_i is conditionally independent with density $\pi_{HS}(\beta_i \mid \tau)$, where π_{HS} can be represented as a scale mixture of normals:

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

where $\mathrm{C}^+(0,1)$ is a half-Cauchy distribution for the standard deviation λ_i. We refer to the λ_i's as the *local* shrinkage parameters and to τ as the *global* shrinkage parameter.

Figure 1 plots the densities for the horseshoe, Laplacian and Student-t priors. The density function $\pi_{HS}(\beta_i|\tau)$ lacks a closed-form representation, but it behaves essentially like $\log(1 + 2/\beta_i^2)$, and can be well approximated by elementary functions as detailed in Theorem 1 of Carvalho *et al.* (2008).

The horseshoe prior has two interesting features that make it particularly useful as a shrinkage prior for sparse problems. Its flat, Cauchy-like tails allow strong signals to remain large (that is, un-shrunk) *a posteriori*. Yet its infinitely tall spike at the origin provides severe shrinkage for the zero elements of $\boldsymbol{\beta}$. As we will highlight in the discussion that follows, these are the key elements that make the horseshoe an attractive choice for handling sparse vectors.

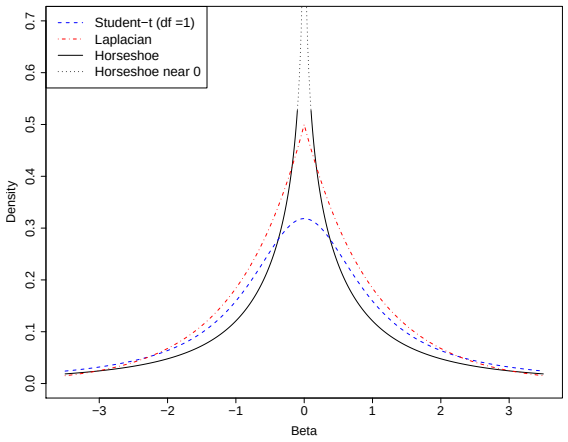

Figure 1: The horseshoe prior and two close cousins: Laplacian and Student-t.

2.1 Relation to other shrinkage priors

The density in (1) is perfectly well defined without reference to the λ_i's, which can be marginalized away. But by writing the horseshoe prior as a scale mixture of normals, we can identify its relationship with commonly used procedures in supervised learning. For example, exponential mixing, with $\lambda_i^2 \sim \mathrm{Exp}(2)$, implies independent Laplacian priors for each β_i; inverse-gamma mixing, with $\lambda_i^2 \sim \mathrm{IG}(a, b)$, leads to Student-t priors. The former represents the underlying model for the LASSO (Tibshirani, 1996), while the latter is the model associated with the relevance vector machine (RVM) of Tipping (2001).

This common framework allows us to compare the appropriateness of the assumptions made by different models. These assumptions can be better understood by representing models in terms of the "shrinkage profiles" associated with their posterior expectations. Assume for now that $\sigma^2 = \tau^2 = 1$, and define $\kappa_i = 1/(1+\lambda_i^2)$. Then κ_i is a random shrinkage coefficient, and can be interpreted as the amount of weight that the posterior mean for β_i places on 0 once the data $\boldsymbol{y}$ have been observed:

$$E(\beta_i \mid y_i, \lambda_i^2) = \left(\frac{\lambda_i^2}{1+\lambda_i^2}\right) y_i + \left(\frac{1}{1+\lambda_i^2}\right) 0 = (1-\kappa_i) y_i\,.$$

Since $\kappa_i \in [0,1]$, this is clearly finite, and so by Fubini's theorem,

$$\begin{aligned} E(\beta_i \mid y) &= \int_0^1 (1-\kappa_i) y_i \; \pi(\kappa_i \mid y) \; \mathrm{d}\kappa_i \\ &= \{1 - E(\kappa_i \mid y_i)\}\; y\,. \end{aligned} \tag{2}$$

By applying this transformation and inspecting the priors on κ_i implied by different choices for $\pi(\lambda_i)$, we

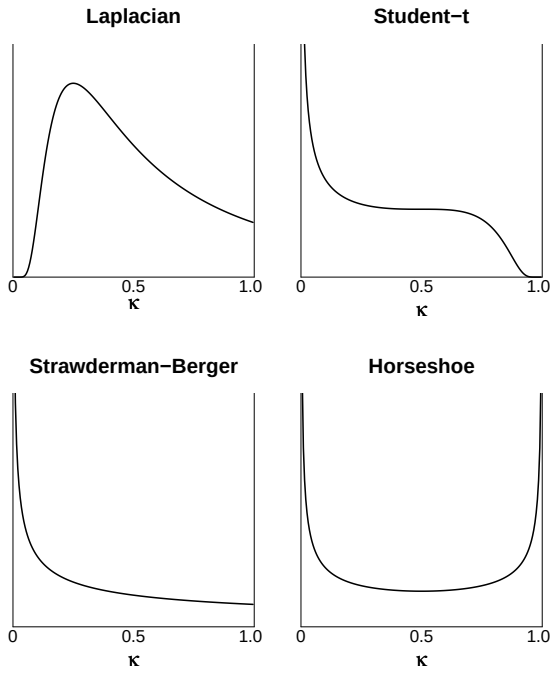

Figure 2: Densities for the shrinkage weights $\kappa_i \in [0, 1]$. $\kappa_i = 0$ means no shrinkage and $\kappa_i = 1$ means total shrinkage to zero.

can develop an understanding of how these models attempt to discern between signal and noise. Figure 2 plots the densities for κ derived from a few important models in this class. Choosing $\lambda_i \sim \mathrm{C}^+(0, 1)$ implies $\kappa_i \sim \mathrm{Be}(1/2, 1/2)$, a density that is symmetric and unbounded at both 0 and 1. This horseshoe-shaped shrinkage profile expects to see two things *a priori*: strong signals ($\kappa \approx 0$, no shrinkage), and zeros ($\kappa \approx 1$, total shrinkage).

No other commonly used shrinkage prior shares these features. The Laplacian prior tends to a fixed constant near $\kappa = 1$, and disappears entirely near $\kappa = 0$. The Student-t prior and the Strawderman–Berger prior (see Section 2.3) are both unbounded near $\kappa = 0$, reflecting their heavy tails. But both are bounded near $\kappa = 1$, limiting these priors in their ability to squelch noise components back to zero.

As an illustration, consider a simple example. Two repeated standard normal observations y_{i1} and y_{i2} were simulated for each of 1000 means: 10 signals with $\beta_i = 10$, 90 signals $\beta_i = 2$ and 900 noise components where $\beta_i = 0$. Based on this data, we estimate the vector $\boldsymbol{\beta}$ under two models: (i) independent horseshoe priors for each β_i, and (ii) independent Laplacian priors. Both models assume $\tau \sim \mathrm{C}^+(0, 1)$, along with Jeffreys' prior $\pi(\sigma) \propto 1/\sigma$.

The shrinkage characteristics of the models are presented in Figure 3, where $\bar{y}_i$ is plotted against $\hat{\beta}_i = \mathrm{E}(\beta_i|\boldsymbol{y})$. The important differences occur when $\bar{y}_i \approx 0$ and when $\bar{y}_i$ is large. Compared to the horseshoe prior, the Laplacian specification tends to over-shrink the large values of $\bar{y}$ and yet under-shrink the noise observations. This is a direct effect of the prior on κ_i, which in the Laplacian case is bounded both at 0 and 1, limiting the ability of each κ_i to approach these values *a posteriori*.

Figure 3 also plots posterior draws for the global shrinkage parameter τ, offering a closer look at the mechanism underlying signal discrimination. Under the horseshoe model, τ is estimated to be much smaller than in the Laplacian model. This is perhaps the single most important characteristic of the horseshoe: the clear separation between the global and local shrinkage effects. The global shrinkage parameter tries to estimate the overall sparsity level, while the local shrinkage parameters are able to flag the non-zero elements of $\boldsymbol{\beta}$. Heavy tails for $\pi(\lambda_i)$ play a key role in this process, allowing the estimates of β_i to escape the strong "gravitational pull" towards zero exercised by τ.

Put another way, the horseshoe has the freedom to shrink globally (via τ) and yet act locally (via λ_i). This is not possible under the Laplacian prior, whose shrinkage profile forces a compromise between shrinking noise and flagging signals. This leads to overestimation of the signal density of underlying vector, combined with under-estimation of larger elements of $\boldsymbol{\beta}$. Performance therefore suffers—in this simple example, the mean squared-error was 25% lower under the horseshoe model.

In fairness, the most commonly used form of the Laplacian model is the LASSO, where estimators are defined by the posterior mode (MAP), thereby producing zeros in the solution set. Our experiments of the next section, however, indicate that the issues we have highlighted about Laplacian priors remain even when the mode is used—the overall estimate of the sparsity level will still be governed by τ, which in turn is heavily influenced by the tail behavior of the prior on λ_i. Robustness here is crucial, which is an issue towards which we now turn.

2.2 Robust Shrinkage

The robust behavior of the horseshoe can be formalized using the following representation of the posterior mean of β when $(y|\beta) \sim \mathrm{N}(\beta, 1)$. Conditional on one sample y^*,

$$\mathrm{E}(\beta|y^*) = y^* + \frac{\mathrm{d}}{\mathrm{d}y^*} \ln m(y^*), \tag{3}$$

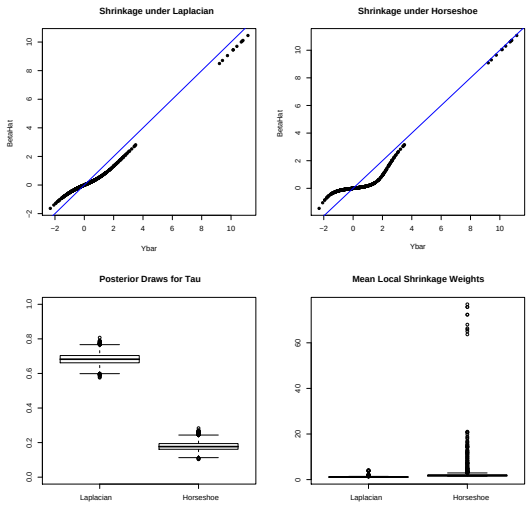

Figure 3: Plots of $\bar{y}_i$ versus $\hat{\theta}_i$ for Laplacian (left) and horseshoe (right) priors on data where most of the means are zero. The diagonal lines are where $\hat{\theta}_i = \bar{y}_i$.

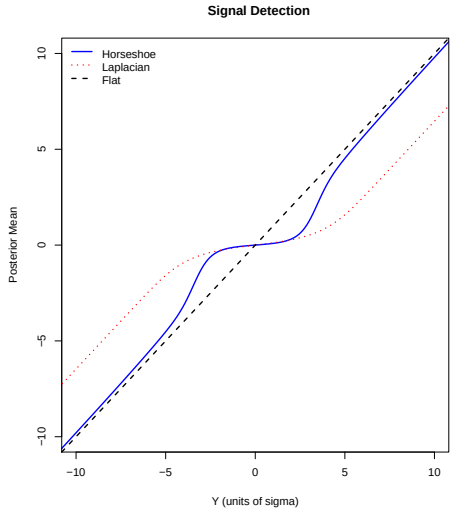

Figure 4: A comparison of the posterior mean versus y for horseshoe and Laplacian priors.

where $m(y^*) = \int p(y^*|\beta)\,\pi(\beta)\,\mathrm{d}\beta$ is the marginal density for y^* (see Polson, 1991).

From (3) we get an essential insight about the behavior of an estimator in situations where y^* is very different from the prior mean. In particular, robustness is achieved by using priors having the "bounded influence" property—i.e. those giving rise to a score function that is bounded as a function of y^*. If such a bound exists, then for large values of $|y^*|$, $\mathrm{E}(\beta|y^*) \approx y^*$, implying that the estimator never misses too badly in the tails of the prior.

Theorem 3 of Carvalho *et al.* (2008) shows that the horseshoe prior is indeed of bounded influence, and furthermore that

$$\lim_{|y^*|\to\infty} \frac{\mathrm{d}}{\mathrm{d}y^*} \ln m_H(y^*) = 0\,. \tag{4}$$

The Laplacian prior is also of bounded influence, but crucially, this bound does not decay to zero in the tails. Instead,

$$\lim_{|y^*|\to\infty} \frac{\mathrm{d}}{\mathrm{d}y^*} \ln m_L(y^*) = \pm a\,, \tag{5}$$

where a varies inversely with the global shrinkage parameter τ (Pericchi and Smith, 1992). Unfortunately, when the vector $\boldsymbol{\beta}$ is sparse, τ will be estimated to be small, and this "nonrobustness bias" a will be quite large. Figure 4 illustrates these results by showing the relationship between y^* and the posterior mean under both the horseshoe and the Laplacian priors. These are available analytically for fixed values of τ, which for the sake of illustration were chosen to yield near-identical shrinkage within roughly 2σ of the origin. Both models can "bow" near the origin to accommodate sparse vectors by changing τ; only the horseshoe can simultaneously perform well in the tails, even when τ is very small.

This effect can be confirmed by inspecting the joint distribution of the data and parameters under the horseshoe prior,

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

from which it is clear that the marginal density for κ_i is always unbounded at 1, regardless of τ. (This is one reason why the posterior mode is inappropriate here.) Hence the horseshoe prior, its tail robustness notwithstanding, will always have the ability to severely shrink elements of $\boldsymbol{\beta}$ when needed.

2.3 Relation to Bayesian model averaging

As we have mentioned, one alternative approach for handling sparsity is the use of discrete mixtures priors, where

$$\beta_i \sim (1-w)\delta_0 + w \cdot \pi(\beta_i)\,. \tag{7}$$

Here, w is the prior inclusion probability, and δ_0 is a degenerate distribution at zero, so that each β_i is assigned probability $(1-w)$ of being zero *a priori*.

Crucial to the good performance of the model in (7) are the choice of $\pi(\beta)$ and the careful estimation of w. The former allows large signals to be accommodated, while the latter allows the model to adapt to the overall level of sparsity in $\boldsymbol{\beta}$, automatically handling the

implied multiple-testing problem (Scott and Berger, 2006). By carefully choosing $\pi(\beta)$ and accounting for the uncertainty in w, this model can be considered the "gold standard" for sparse problems, both theoretically and empirically. This is extensively discussed by, for example, Hoeting *et al* (1999) and Johnstone and Silverman (2004).

The discrete-mixture model is therefore an important benchmark for any shrinkage prior. Here, we will focus on a version of the discrete mixture where the nonzero β_i's follow independent Strawderman–Berger priors (Strawderman, 1971; Berger, 1980), which have Cauchy-like tails and and yet still allow closed-form convolution with the normal likelihood. Figure 2 displays the shrinkage profile of the Strawderman-Berger prior on the κ scale, where it is seen to yield a Beta(1, 1/2) distribution. Here, the point mass at $\beta_i = 0$ can be equivalently be construed as a point mass at $\kappa_i = 1$. If Strawderman–Berger priors are assumed for the nonzero β_i's, the discrete mixture model will yield a shrinkage profile with the desired unboundedness both at $\kappa_i \approx 0$ (signal) and $\kappa_i \approx 1$ (noise).

Notice that both the horseshoe prior and the discrete mixture have mechanisms for controlling the overall signal density in $\boldsymbol{\beta}$. In the discrete mixture model, this parameter is clearly w, the prior inclusion probability. But under the horseshoe prior, this role is played by τ, the common variance parameter. This is easily seen from the joint distribution in (6), since one can approximate the conditional posterior for τ by

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

where $\bar{\kappa} = p^{-1} \sum_{i=1}^{p} \kappa_i$. This is essentially a $\mathrm{Ga}\{(p+2)/2, (p-\bar{\kappa})/\bar{\kappa}\}$ distribution for τ^{-2}, with posterior mean equal to $2(1-\bar{\kappa})/\bar{\kappa}$. When $\bar{\kappa}$ gets close to 1, implying that most observations are shrunk to zero, then τ^2 is estimated to be very small.

2.4 Hyperparameters

Much of the above discussion focused on the behavior implied by different choices of priors for the local shrinkage parameters λ_i's. Yet the estimation of the global parameters τ and σ plays a large role in separating signal from noise, as seen in the example depicted in Figure 2.

So far, we have focused on a fully Bayesian specification where weakly informative priors were used both for τ and σ (as well as w in the discrete mixture). There is a vast literature on choosing priors for variance components in general hierarchical models, and justifications for our choices of $\tau \sim C^+(0, 1)$ and $\pi(\sigma) \propto 1/\sigma$ appear in Gelman (2006). Alternatives to a fully Bayesian analysis include cross validation and empirical-Bayes, often called Type-II maximum likelihood. These "plug-in" analysis are, in fact, the standard choices in many applications of shrinkage estimation in both machine learning and statistics.

While we certainly do not intend to argue that "plug-in" alternatives are wrong *per se*, we do recommend, as a conservative and more robust route, the use of the fully Bayesian approach. The full Bayes analysis is quite simple computationally using MCMC, and will avoid at least three potential problems:

1. Plug-in approaches will ignore the unknown correlation structure between τ and σ (or τ, σ and w in the discrete mixture model). This can potentially give misleading results in situations where the correlation is severe, while the full Bayes analysis will automatically average over this joint uncertainty.

2. The marginal maximum-likelihood solution is always in danger of collapsing to the degenerate $\hat{\tau} = 0$. The issue is exacerbated when very few signals are present, in which case the posterior mass of τ will concentrate near 0 and signals will be flagged via large values of the local shrinkage parameters λ_i.

3. Plug-in methods may fail to correspond to any kind of Bayesian analysis even asymptotically, when there no longer is any uncertainty about the relevant hyperparameters. See Scott and Berger (2008) for an extensive discussion of this phenomenon.

One may ask, of course, whether a global scale parameter τ is even necessary, and whether the local parameters λ_i can be counted upon to do all the work. (This is the tactic used in, for example, the relevance vector machine.) But this is equivalent to choosing $\tau = 1$, and we feel that Figure 3 is enough to call this practice into question, given how far away the posterior distribution is from $\tau = 1$.

3 Examples

Carvalho, Polson, Scott and Yae (2009) provide an extensive discussion of the use of the horseshoe in traditional supervised-learning situations, including linear regression, generalized linear models, and function estimation through basis expansions. We now focus on a few examples that highlight the effectiveness of the horseshoe as a good default procedure.

Loss		$\sigma^2 = 1$			$\sigma^2 = 9$		
		LP	HS	DM	LP	HS	DM
ℓ^2	LP	**209**	1.62	1.62	**850**	1.47	1.51
	HS		**77**	0.95		**416**	0.99
	DM			**93**			**440**
ℓ^1	DE	**178**	1.50	1.60	**341**	1.56	1.75
	HS		**80**	1.02		**142**	1.10
	DM			**83**			**123**

Table 1: Risk under squared-error (ℓ^2) loss and absolute-error (ℓ^1) loss in Experiment 1. Bold diagonal entries in the top and bottom halves are median sum of squared-errors and absolute errors, respectively, in 1000 simulated data sets. Off-diagonal entries are average risk ratios, risk of row divided by risk of column, in units of σ. LP: Laplacian. HS: horseshoe. DM: discrete mixture, fully Bayes.

These situations all involve an n-dimensional vector $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$, where $\boldsymbol{X}$ is a $n \times p$ design matrix. In regression, the rows of $\boldsymbol{X}$ are the predictors for each subject; in basis models, they are the bases evaluated at the points in predictor space to which each entry of $\boldsymbol{y}$ corresponds. The horseshoe prior for the p-dimensional vector $\boldsymbol{\beta}$ takes the form in (1).

3.1 Exchangeable means

Experiment 1 demonstrates the operational similarities between the horseshoe and a heavy-tailed discrete mixture. We still focus on the problem of estimating a p-dimensional sparse mean ($\boldsymbol{\beta}$) of a multivariate normal distribution (implying that $\boldsymbol{X}$ is the identity). We simulated 1000 data sets with different sparsity configurations, and 20% non-zero entries on average. Nonzero β_i's were generated randomly from a Student-t distribution with scale $\tau = 3$ and degrees of freedom equal to 3. Data $\boldsymbol{y}$ was simulated under two possibilities for the noise variables: $\sigma^2 = 1$ and $\sigma^2 = 9$. In each data set we estimate $\boldsymbol{\beta}$ by the posterior mean under three different models: horseshoe, Laplacian and discrete mixtures. The results for estimation risk are reported in Table 1.

Regardless of the situation, the Laplacian loses quite significantly both to the horseshoe prior and the discrete-mixture model. Yet neither of these two options enjoys a systematic advantage; their similarities in shrinkage profiles seem to translate quite directly to similar empirical results. Meanwhile, the nonrobustness of the Laplacian prior is quite apparent.

3.2 Regression

In Experiment 2, we chose two fixed vectors of ten nonzero coefficients: $\boldsymbol{\beta}_{1:10} = (2, 2, 2, 2, 2, 2, 2, 2, 5, 20)$ and $\boldsymbol{\beta}_{1:10} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. We then "padded" these with $(p - 10)$ zeros for several different choices of p, simulated random design matrices with moderately correlated entries, and simulated $\boldsymbol{y}$ by adding standard normal errors to the true linear predictor $\boldsymbol{X}\boldsymbol{\beta}$. In all cases, n scaled linearly with p.

Case 1: $\boldsymbol{\beta}_{1:10} = (2, 2, 2, 2, 2, 2, 2, 2, 5, 20)$

p	20	50	100	200	400
n	24	60	120	240	480
Lasso	1.86	0.78	0.34	0.13	0.12
HS	1.28	0.33	0.11	0.06	0.07

Case 2: $\boldsymbol{\beta}_{1:10} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$

p	20	50	100	200	400
n	25	55	105	205	405
Lasso	0.61	0.40	0.48	0.21	0.23
HS	0.31	0.23	0.12	0.09	0.08

Table 2: Mean-squared error in estimating $\boldsymbol{\beta}$ in Experiment 2.

For this example, we evaluated the horseshoe using the LASSO (i.e. the posterior mode under Laplacian priors) as a benchmark, with τ chosen through cross-validation. Results are presented in Table 2.

In Experiment 3, we fixed $p = 50$, but rather than fixing the non-zero values of $\boldsymbol{\beta}$, we simulated 1000 data sets with varying levels of sparsity, where nonzero β_i's were generated from a standard Student-t with 2 degrees of freedom. (The coefficients were 80% sparse on average, with nonzero status decided by a weighted coin flip.) We again compared the horseshoe against the LASSO, but also included Bayesian model-averaging using Zellner-Siow priors as a second benchmark. Results for both estimation error and out-of-sample prediction error are displayed in Figure 5.As these results show, both BMA and the horseshoe prior systematically outperform the LASSO in sparse regression problems, without either one enjoying a noticeable advantage over the other.

3.3 Basis expansion with kernels

In Experiment 4, we used the sine test function described in Tipping (2001) to assess the ability of the horseshoe prior to handle regularized kernel regression. For each of 100 different simulated data sets, 100 random points t_i were simulated uniformly between -20 and 20. The response y_i was then set to $\sin(t_i)/t_i + \epsilon_i$, with $\epsilon_i \sim \mathrm{N}(0, \sigma = 0.15)$.

The goal was to estimate the underlying function $f(t)$ using kernel methods. As a benchmark, we use the relevance vector machine, corresponding to independent Student-t priors with zero degrees of freedom. Gaussian kernels were centered at each of the 100 observed points, with the kernel bandwidth chosen as the de-

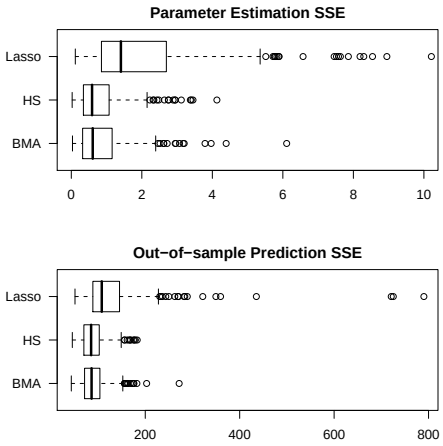

Figure 5: Results for Experiment 3. "BMA" refers to the model-averaged results under Zellner-Siow priors. "Lasso" refers to the posterior mode under Laplacian priors.

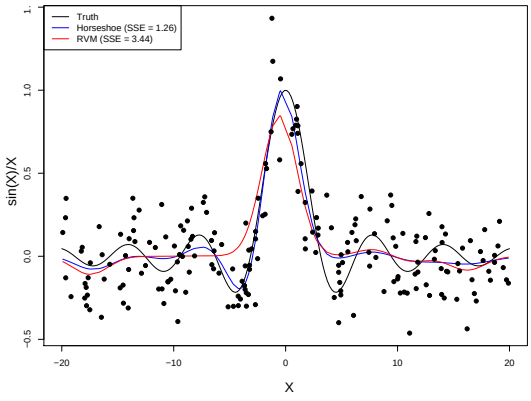

Figure 6: One example data set in Experiment 5 involving the $\sin(t)/t$ test function, showing the true function, data, and horseshoe/RVM estimates.

fault in the "rvm" function in the R package "kernlab." These kernel basis functions, evaluated at the observed values of t_i, formed the 100×100 design matrix, with $\boldsymbol{\beta}$ representing the vector of kernel weights.

In these 100 simulated data sets, the average sum of squared errors in estimating $f(t)$ at 100 out-of-sample t points was 7.55 using the horseshoe prior, and 8.19 using the relevance vector machine. In 91 cases of 100, the horseshoe prior yielded lower risk. An example of one simulated data set is in Figure 6.

3.4 Unsupervised covariance estimation

Suppose we observe a matrix Y whose rows are n realizations of a p-dimensional vector $\boldsymbol{y} \sim \mathrm{N}(0, \Sigma)$, and that the goal is to estimate Σ. This is an important problem in portfolio allocation, where one must assess the variance of a weighted portfolio of assets, and where regularized estimates of Σ are known to offer substantial improvements over the straight estimator $\hat{\Sigma} = Y'Y$.

L-OR	L-AND	L-Chol	HS	BMA
791	729	520	372	347

Table 3: Sum of squared errors in predicting missing return values in the 59-dimensional mutual fund example. "L-OR" and "L-AND" refer to estimates based on Lasso regressions in the full conditionals of each asset. "L-Chol" and "HS" refer to Lasso and horseshoe models on the triangular system of linear regressions from the Cholesky decomposition of Σ^{-1}. Finally, "BMA" is based on Bayesian model averaging using the FINCS.

A useful way of regularizing Σ is by introducing off-diagonal zeros in its inverse Ω. This can be done by searching for undirected graphs that characterize the Markov structure of $\boldsymbol{y}$, a process known as Gaussian graphical modeling (see Jones *et. al*, 2005). While quite potent as a tool for regularization, Bayesian model averaging across different graphical models poses the same difficulties as it does in linear models: marginal likelihoods are difficult to compute, and the model space is enormously difficult to search.

Luckily, Gaussian graphical modeling can also be done indirectly, either by fitting a series of sparse self-on-self regression models for $(y_j \mid \boldsymbol{y}_{-j})$, $j = 1, \ldots, p$, or by representing the Cholesky decomposition of Ω as a triangular system of sparse regressions. The first option is done using the LASSO by Meinshausen and Buhlmann (2006). We now present similar results using the horseshoe.

Our test data set is the Vanguard mutual-fund data set ($p = 59$, $n = 86$) of Carvalho and Scott (2009). We recapitulate their out-of-sample prediction exercise, which involves estimating Σ using the first 60 observations, and then attempting to impute random subsets of missing values among the remaining 26 observations. We use that paper's full BMA results as a benchmark (which required many hours of computing using the FINCS algorithm of Scott and Carvalho, 2008).

Results from this prediction exercise are presented in Table 3, where it is clear that the horseshoe, despite being a much simpler computational strategy, performs almost as well as the benchmark (Bayesian model averaging). Once again, both BMA and the horseshoe outperform alternatives based on the Laplacian prior.

4 Discussion

We have introduced and discussed the use of the horseshoe prior in the estimation of sparse vectors in supervised learning problems. The horseshoe prior is based on a novel multivariate-normal scale mixture; it yields estimates that are robust both to unknown sparsity patterns and to large outlying signals, making it an attractive default option.

It is reassuring that the theoretical insights of Section 2 regarding sparsity and robustness can be observed in practice, as we have demonstrated through a variety of experiments. Moreover, it is surprising that in all situations where we have investigated the matter, the answers obtained by the horseshoe closely mimic those arising from the gold standard for sparse estimation and prediction: Bayesian model averaging across discrete mixture models. This is an interesting (and as yet under-explored) fact that may prove very useful in ultra-high-dimensional situations, where the computational challenges associated with BMA may be very cumbersome indeed.

Additional detail concerning these issues can be found in working papers available from the authors' websites.

Acknowledgements

The first author acknowledges the support of the IBM Corporation Scholar Fund at the University of Chicago Booth School of Business, and the third author that of a graduate research fellowship from the U.S. National Science Foundation.

References

J. Berger (1980). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean. *The Annals of Statistics, 8* 716–761.

C. Carvalho, N. Polson and J. G. Scott (2008). The Horseshoe Estimator for Sparse Signals. Discussion Paper 2008-31. Duke University Department of Statistical Science.

C. Carvalho, N. Polson, J. G. Scott and S. Yae (2009). Bayesian Regularized Regression and Basis Expansion via the Horseshoe. *Working Paper.*

C. Carvalho and J. G. Scott (2009). Objective Bayesian Model Selection in Gaussian Graphical Models. *Biometrika* (to appear).

A. Gelman (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis, 1.* 515–533.

E. George and R. McCulloch (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association 88*, 881–889.

J. Hoeting, D. Madigan, A. E. Raftery, and C. Volinsky (1999). Bayesian model averaging: a tutorial. *Statist. Sci. 14*, 382–417.

I. Johnstone and B. Silverman (2004). Needles and Straw in Haystacks: Empirical-Bayes Estimates of Possibly Sparse Sequences. *The Annals of Statistics, 32*, 1594–1649.

B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter and M. West (2005). Experiments in Stochastic Computation for High-dimensional Graphical Models. *Statistical Science, 20*, 388-400.

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics 34*, 1436–1462.

T. Mitchell and J. Beauchamp (1988). Bayesian Variable Selection in Linear Regression (with discussion). *Journal of the American Statistical Association, 83*, 1023–1036.

L. Pericchi and A. Smith (1992). Exact and Appropriate Posterior Moments for a Normal Location Parameter. *Journal of the Royal Statistical Society B 54*, 793–804.

N. Polson (1991). A Representation of the Posterior Mean for a Location Model. *Biometrika, 78*, 426–430.

J. G. Scott and J. Berger (2006). An Exploration of Aspects of Bayesian Multiple Testing. *Journal of Statistical Planning and Inference, 136*, 2144-2162.

J. G. Scott and J. Berger (2008). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. Discussion Paper 2008-10. Duke University Department of Statistical Science.

J. G. Scott and C. Carvalho (2008). Feature-inclusion Stochastic Search for Gaussian Graphical Models. *Journal of Computational and Graphical Statistics, 17*, 790-808.

W. Strawderman (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Statistics, 42*, 385–388.

R. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society B, 58*, 267-288.

M. Tipping (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research, 1*, 211-244.