Bayesian Analysis (2013) **8**, Number 1, pp. 111–132

Asymptotic Properties of Bayes Risk for the Horseshoe Prior

Jyotishka Datta * and Jayanta. K. Ghosh †

Abstract. In this paper, we establish some optimality properties of the multiple testing rule induced by the horseshoe estimator due to Carvalho, Polson, and Scott (2010, 2009) from a Bayesian decision theoretic viewpoint. We consider the two-groups model for the data and an additive loss structure such that the total loss is equal to the number of misclassified hypotheses. We use the same asymptotic framework as Bogdan, Chakrabarti, Frommlet, and Ghosh (2011) who introduced the Bayes oracle in the context of multiple testing and provided conditions under which the Benjamini-Hochberg and Bonferroni procedures attain the risk of the Bayes oracle. We prove a similar result for the horseshoe decision rule up to $O(1)$ with the constant in the horseshoe risk close to the constant in the oracle. We use the Full Bayes estimate of the tuning parameter τ. It is worth noting that the Full Bayes estimate cannot be replaced by the Empirical Bayes estimate, which tends to be too small.

Keywords: Multiple Testing, Horseshoe Decision Rule, Asymptotic Optimality, Bayes Oracle.

1 Introduction

In the recent past, thanks to microarrays for gene expression as well as examples in other fields (vide Efron (2008)), multiple independent tests have become very popular. A popular model is the two-groups normal model, with sparse signals, due to many people of whom the group at Stanford has become most visible. One of the most popular models that has come out as work of these different groups is as follows:

Suppose our data is modeled as m independent observations $Y_1, Y_2, \ldots, Y_m$ with each $Y_i \sim \mathcal{N}(\theta_i, \sigma_0^2)$, where θ_i's are m unknown parameters. Typically m is large so that we have many tests, each based on a single observation, which is in many cases the test statistic, suitably transformed to have approximate normal distribution. It is remarkable that in the completely classical setting of m unknown θ_i's of Benjamini and Hochberg (1995), one can define a test based on p-values such that the false discovery rate can be controlled at any pre-assigned given value.

It has been realized that since many tests have been performed simultaneously, we should model the data further such that learning via Empirical Bayes or Full Bayes methods becomes possible. We proceed to do that now by introducing the two groups model. We use an indicator $\nu_i, i = 1, \ldots, m$ such that $\nu_i = 0$ indicates $\theta_i = 0$ and $\nu_i = 1$ indicates $\theta_i \neq 0$ and in fact $\theta_i \sim \mathcal{N}(0, \psi^2)$ where ψ^2 is a measure of average signal

*Department of Statistics, Purdue University, West Lafayette, IN, jdatta@stat.purdue.edu
†Department of Statistics, Purdue University, West Lafayette, IN, ghosh@stat.purdue.edu

 DOI:10.1214/13-BA805

magnitude. Unconditionally, θ_i's are independently distributed as:

$$\theta_i \sim (1-p)\delta_{\{0\}} + p\mathcal{N}(0,\psi^2) \tag{1}$$

where $\delta_{\{0\}}$ is the degenerate distribution at zero. The marginal distribution of Y_i is then a mixture of normals, viz.,

$$Y_i \sim (1-p)\mathcal{N}(0,\sigma_0^2) + p\mathcal{N}(0,\sigma_0^2+\psi^2). \tag{2}$$

We are interested in testing the hypothesis:

$$H_{0i} : \nu_i = 0 \text{ vs. } H_{Ai} : \nu_i = 1. \tag{3}$$

Equation (2) defines the two-groups model for Y_i. The two groups being modeled are $\mathcal{N}(0,\sigma_0^2)$ and $\mathcal{N}(0,\sigma_0^2+\psi^2)$. Notice that instead of m unknown parameters $\theta_i, i = 1,\ldots,m$, we have now only three unknown parameters p,σ_0^2,ψ^2, of which σ_0^2 is usually assumed known. Typically, we assume that we have a sparse model, i.e. p, the proportion of non-zero θ_i's, is small. In our assumption, $p \to 0$ as $m \to \infty$. For more details see Bogdan, Ghosh, and Tokdar (2008) and Bogdan, Chakrabarti, Frommlet, and Ghosh (2011).

In this situation, the Empirical Bayes approach has been very popular. Efron (2008, 2004) has shown that the Benjamini-Hochberg rule can be interpreted in the sparse case as basically an Empirical Bayes rule. Major contributions to Empirical Bayes theory are Storey (2003, 2007) and Genovese and Wasserman (2004). A Full Bayes treatment is available in Scott and Berger (2006, 2010). An extension of the Benjamini-Hochberg approach appears in Sarkar (2006). Bogdan, Chakrabarti, Frommlet, and Ghosh (2011) introduced the notion of 'Asymptotic Bayes Optimality under Sparsity' (ABOS) and provided conditions under which the Benjamini-Hochberg procedure is ABOS for a two-groups model. Bogdan et al. (2011) argue that while it is expected that empirical Bayes multiple tests should also have such optimality properties, estimation of the sparsity parameter will need special care, see e.g. Scott and Berger (2010) as well as Bogdan, Ghosh, and Tokdar (2008).

In a series of remarkable papers and technical reports, Carvalho, Polson, and Scott (2010, 2009); Scott (2011) and Polson and Scott (2012) introduced a one group model instead of a two-groups model and what they call the horseshoe prior for the one-group model. The one-group model needs significantly less computational effort than the two-groups model. Moreover, Carvalho et al. (2010) go on to provide strong numerical evidence that it attains the oracle up to $O(1)$ with a constant in horseshoe risk close to the constant in oracle. The one group model and the horseshoe prior is formally introduced in the next section.

Our goal in this paper is to provide a formal theoretical proof for the result of Carvalho et al. (2009) assuming either the value of the tuning parameter is known or it is estimable numerically from the data. We will provide some numerical evidence that the latter assumption seems to be valid. Such a proof, partly based on numerical validation, has appeared in a different context, see Sen, Banerjee, and Woodroofe (2010).

The intuitive reason why the horseshoe prior works so well is that the posterior inclusion probability of the two groups model is well captured in the shrinkage weight of the horseshoe prior (vide Figure 4 presented in Section 4 below). We thank Prof. Jim Berger for suggesting that this might be the case. A similar comparison of the two inclusion probabilities for the 'fixed-k' asymptotic scenario was done by Carvalho et al. (2010) where the number of signals remains fixed, letting the number of noise observations grow arbitrarily.

We have compared the new horseshoe with a much older one arising from multidimensional scaling, studied in depth by Diaconis, Goel, and Holmes (2008). While we do not find any deep link or insight worth sharing, we did find each method can be applied to the other's domain, providing some interesting new twists. We hope to return to a comparison of the two elsewhere.

2 Horseshoe prior

The horseshoe model introduced in Carvalho et al. (2009) is given by

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

where $C^+(0,1)$ is a standard half-Cauchy distribution on positive real numbers. The posterior distribution of θ_i is normal with mean and variance given by:

$$\begin{aligned} E(\theta_i|y_i, \lambda_i, \tau, \sigma^2) &= (1 - \frac{1}{1+\lambda_i^2\tau^2})y_i \\ V(\theta_i|y_i, \lambda_i, \tau, \sigma^2) &= (1 - \frac{1}{1+\lambda_i^2\tau^2})\sigma^2. \end{aligned}$$

If we define $\kappa_i = \frac{1}{1+\lambda_i^2\tau^2}$, the posterior mean of θ_i is $E(\theta_i|y_i, \kappa_i, \tau, \sigma^2) = (1-\kappa_i)y_i$ and hence by Fubini's theorem

$$E(\theta_i|y_i, \tau, \sigma^2) = (1 - E(\kappa_i|y_i, \tau, \sigma^2))y_i.$$

For large ψ^2, the posterior mean $\hat{\theta}_i$ under the two-groups model in (1)-(2) is approximately $\omega_i y_i$, where ω_i is the posterior inclusion probability for θ_i. The horseshoe estimator, on the other hand, has the form $\hat{\theta}_i = (1-\hat{\kappa_i})y_i$, which means the shrinkage weight $1-\hat{\kappa_i}$ in the horseshoe model, though not a formal posterior quantity, behaves in the same way as the posterior inclusion probability ω_i (vide Figure 4 as well as Section 3.4 in Carvalho et al. (2010)). Also, κ_i can be interpreted as a random shrinkage coefficient in the sense that the behavior of the posterior density of κ_i near 0 identifies the signals and $\kappa_i \approx 1$ shrinks the noise. The name "horseshoe" was attributed to the fact that the half-Cauchy prior on λ_i imposes a horseshoe-shaped Beta$(\frac{1}{2}, \frac{1}{2})$ prior on the shrinkage coefficients κ_i. The similarity of the shrinkage weights $1-\hat{\kappa_i}$ with the

posterior inclusion probabilities under the two-groups model leads Carvalho et al. (2010) to propose the following natural decision rule under a symmetric 0-1 loss function:

$$\text{Reject } H_{0i} \text{ if } \omega_i = 1 - E(\kappa_i|y_i, \tau, \sigma^2) > \frac{1}{2}.$$

The horseshoe decision rule identifies the signals (resp. the noise) through the simple thresholding rule $1 - \hat{\kappa_i} > \frac{1}{2}$ (resp. $1 - \hat{\kappa_i} < \frac{1}{2}$). Thus, the probabilities for type I and type II error for the horseshoe decision rule are as follows:

$$t_{1i} = P_{H_{0i}}(H_{0i} \text{ is rejected}) = P_{H_{0i}}\left(E(\kappa_i|y_i, \tau, \sigma^2) < \frac{1}{2}\right) \tag{4}$$

$$t_{2i} = P_{H_{Ai}}(H_{0i} \text{ is accepted}) = P_{H_{Ai}}\left(E(\kappa_i|y_i, \tau, \sigma^2) > \frac{1}{2}\right). \tag{5}$$

For the sake of simplicity, we will assume σ^2 is known as in Abramovich et al. (2006) who also provide some discussion on this point. This is a common assumption for deriving the asymptotic properties for a multiple testing problem and has also been used in Bogdan et al. (2008, 2011). Under this assumption, the posterior density for $\sigma = 1$ becomes:

$$p(\kappa_i|y_i, \tau) \propto (1-\kappa_i)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa_i\}^{-1} \exp(-\frac{\kappa_i {y_i}^2}{2})\ \kappa_i \in (0,1).$$

The parameter τ plays a crucial role in controlling the shrinkage behavior of the estimator. We will show in Section 3 that convergence of both the type I and type II error probabilities will depend on the rate of convergence of this parameter. It is called the "global shrinkage parameter" by Carvalho et al. (2010, 2009) as it adjusts to the overall sparsity in the data. In fact, the posterior mass of τ is concentrated near zero when the data is very sparse ($p \to 0$) (vide Figure 3 presented later in Section 4) and signals are identified through smaller values of κ. Polson and Scott (2010) observed that the posterior density of κ has different patterns for different signal strengths when τ is small. For example, the estimator exhibits strong shrinkage through concentration of posterior mass near one for small values of τ and relatively smaller y and the "gravitational pull" of τ is squelched by relatively larger values of y, shifting the mass towards zero (vide Figure 4 of Polson and Scott (2010)). Moderately large y results in a bimodal posterior distribution indicating uncertainty about the decision.

We will follow two routes with respect to τ^2. In the theoretical part of the paper, we will take it as a tuning parameter that we are free to choose. Our choice of τ will depend on the hyper-parameters of the mixture model. In the numerical part, we will discuss how τ can be estimated using a fully Bayesian approach. We will assume the following hyper-priors on τ and σ as in Carvalho et al. (2010, 2009) for the full Bayes treatment:

$$\tau|\sigma \sim \mathrm{C}^+(0, \sigma)$$
$$p(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Simulation from the full joint posterior distribution under these prior specifications can be efficiently done using Markov-chain Monte Carlo updates. A detailed discussion of this strategy is available in Scott (2011).

2.1 Hypergeometric inverted-beta prior

In a technical report, Polson and Scott (2010) have introduced the four parameter hypergeometric inverted-beta prior for κ that has the following form:

$$p(\kappa) = C\kappa^{\alpha-1}(1-\kappa)^{\beta-1}\left\{\frac{1}{\tau^2} + \left(1-\frac{1}{\tau^2}\right)\kappa\right\}^{-1}\exp(-s\kappa); \quad \kappa \in (0,1).$$

Polson and Scott (2010) denote this prior by $\kappa \sim HB(\alpha,\beta,\tau,s)$. They proved that if $\kappa \sim HB(\alpha,\beta,\tau,s)$, the posterior density of κ is proper and it obeys a hypergeometric-beta distribution, in fact, $[\kappa|y] \sim HB(\alpha' = \alpha+\frac{1}{2},\beta,\tau,s+\frac{y^2}{2})$. The effect of these four parameters is discussed in detail in Polson and Scott (2010). This family encompasses a few commonly used shrinkage priors, for example, the hypergeometric-beta prior density yields the horseshoe prior of Carvalho et al. (2010) when $a = b = \frac{1}{2}$ and $s = 0$. The parameter s can be viewed as a second "global shrinkage parameter" with a different effect on the posterior density than τ. Polson and Scott (2010) show through simulation studies that using s has some advantage over τ when the latter converges to zero. We call the prior density $\kappa \sim HB(\alpha = \frac{1}{2},\beta = \frac{1}{2},\tau = 1,s)$ the "modified horseshoe" and compare its misclassification probability with other shrinkage priors in Section 4.

2.2 Double Exponential prior

The double exponential prior with λ_i and τ modeled as

$$\begin{aligned}\lambda_i &\sim \text{Exp}(2\tau^2)\\ \tau &\sim \text{Inverse Gamma}(\xi/2, \xi\text{d}^2/2)\end{aligned}$$

is a popular choice in supervised learning and robust Bayesian inference. This implies an independent Laplacian prior on each θ_i causing noise observations to shrink to zero. The role of the double exponential prior in robust Bayesian inference was established by Pericchi and Smith (1992) before its role as a Bayesian representation of LASSO (Tibshirani (1996)) was made popular by a series of papers by Park and Casella (2008); Hans (2009); Li and Goel (2006). We study briefly the shrinkage properties of this prior in comparison with the original horseshoe and the modified horseshoe.

Now, we will describe the asymptotic framework and show that the simple decision rule for the horseshoe estimator has optimality properties under sparsity under some conditions on the "global shrinkage parameter" τ and the non-null variance of θ_i under the two-groups model. The "optimality" is achieved in the sense of attaining the optimal Bayes risk up to $O(1)$ under an additive loss function, with the constant in horseshoe risk close to the constant in oracle.

3 Theoretical Results

The theoretical results in this section are presented in the following manner. In Section 3.1 we describe the Bayes oracle for multiple testing and the asymptotic framework as introduced in Bogdan et al. (2011). The Bayes oracle provides a comparative basis for different multiple testing procedures considered here. In Section 3.2, we provide two technical facts which may help us to understand what is going on. We derive the asymptotic expressions of type I and type II error rates in Sections 3.3 and 3.4 and show that the Bayes risk for horseshoe decision rule attains the optimal risk up to a multiplicative constant.

3.1 Bayes Oracle

We consider the two-groups model in equations (1) and (2) and the multiple testing problem (3). We assume that the type 1 error loss (δ_0) and the type 2 error loss (δ_A) are both equal to 1 and the total loss is the sum of losses for individual tests. If we denote by t_1 and t_2 the probability of type I and type II error respectively, the Bayes risk for the 0-1 loss will be given by:

$$BR_{1,1} = (1-p)t_1 + pt_2.$$

Note that, under $0-1$ loss, the Bayes risk $BR_{1,1}$ is equal to the misclassification probability. The Bayes rule, which minimizes the expected value of the total loss, rejects the null hypothesis H_{0i}, if

$$\frac{f_A(Y_i)}{f_0(Y_i)} \geqslant \frac{1-p}{p}$$

where f_A and f_0 are the densities of Y_i under the alternative and the null hypotheses, respectively. Then for known p, ψ^2, the optimal rule depends only on the absolute value of an individual observation $|Y_i|$ and the optimal decision rule is

$$\text{Reject } H_{0i} \text{ if } |Y_i| > C$$

where

$$C^2 = C^2_{\psi,f} = \frac{1+\psi^2}{\psi^2}\left(\log(\psi^2+1) + 2\log f\right)$$

where $f = \frac{1-p}{p}$. We call this rule the *Bayes oracle* as the risk for this is the lower bound of $(1/m)$ times the risk for any multiple testing procedure under the two-groups model in (1-2). If we reparametrize the parameters as $u = \psi^2$ and $v = uf^2$, then the threshold for the Bayes oracle becomes

$$C^2 = (1+\frac{1}{u})(\log v + \log(1+\frac{1}{u})).$$

The asymptotic framework in Bogdan et al. (2011) is then naturally defined as follows:

Assumption 3.1. *The sequence of vectors $\gamma_m = (\psi_m, p_m)$ satisfies the following conditions:*

$$p_m \to 0; u_m \doteq \psi_m^2 \to \infty; v_m \doteq u_m f_m^2 \doteq \psi_m^2 \left(\frac{1-p_m}{p_m}\right)^2 \to \infty$$
$$\frac{\log v_m}{u_m} \to C \in (0,\infty) \text{ as } m \to \infty$$

Bogdan et al. (2011) provide detailed insight on the threshold C. Very briefly, if $C = 0$ then both the errors are zero and for $C = \infty$, the inference is essentially no better than tossing a coin. Under Assumption 3.1, Bogdan et al. (2011) showed that asymptotically type I and type II error rates of the Bayes oracle take a particularly simple form as given below (vide lemma 3.1 of Bogdan et al. (2011)).

$$t_1^{BO} = \mathrm{e}^{-C/2}\sqrt{\tfrac{2}{\pi v \log v}}(1+o_m) \tag{6}$$
$$t_2^{BO} = (2\Phi(\sqrt{C})-1)(1+o_m). \tag{7}$$

We use the notation o_m to denote an infinite sequence of terms, indexed by m (the number of tests), converging to zero as $m \to \infty$. Under the framework of the two-groups model and an additive loss, the Bayes risk for a fixed-threshold multiple testing rule is given by

$$R = m\left((1-p)t_1 + pt_2\right).$$

It can be easily seen from equations (6)-(7) and Assumption 3.1, that the optimum risk for the Bayes oracle is given by:

$$R_{opt} = m\left((1-p)t_1^{BO} + pt_2^{BO}\right) = mp(2\Phi(\sqrt{C})-1)(1+o_m). \tag{8}$$

3.2 Technical facts

We present two concentration inequalities for the posterior distribution of κ to motivate the derivation of type I and type II error probabilities in the next section.

Theorem 3.1. $P(\kappa < \epsilon|y,\tau) \leqslant \frac{1}{2}\epsilon(1-\epsilon)^{-\frac{3}{2}}\exp(\frac{y^2}{2})\tau(1+o(1))$ *for any fixed* $\epsilon \in (0,1)$ *as* $\tau \to 0$ *uniformly in* $y \in \mathcal{R}$.

An immediate upshot of Theorem 3.1 is the convergence of the posterior distribution of κ to a degenerate distribution at 1 as $\tau \to 0$. We state this result in the following corollary.

Corollary 3.1. $P(\kappa > \epsilon|y,\tau) \to 1$ *as* $\tau \to 0$ *for any fixed* $\epsilon \in (0,1)$ *uniformly in* $y \in \mathcal{R}$.

Proof. (Theorem 3.1) The conditional density of κ, $p(\kappa|y,\tau)$, can be written as the product of three functions on $(0,1)$ as follows:

$$p(\kappa|y,\tau) \propto p_1(\kappa)p_2(\kappa|\tau)p_3(\kappa|y)$$

where $p_1(\kappa) = (1-\kappa)^{-\frac{1}{2}}$, $p_2(\kappa|\tau) = \{1-(1-\tau^2)\kappa\}^{-1}$ and $p_3(\kappa|y) = \exp(-\frac{\kappa y^2}{2})$. For the derivation of the bounds to the integral we use the fact that $p_1(\cdot)$ and $p_2(\cdot)$ are increasing and $p_3(\cdot)$ is decreasing in $\kappa \in (0,1)$ when $\tau \in (0,1)$. Also, for any $\tau \in (0,1)$, $\{1-(1-\tau^2)\kappa\}^{-1} \leqslant (1-\kappa)^{-1}$ for $\kappa \in (0,1)$. So, we get

$$\begin{aligned}
\frac{\int_0^\epsilon p(\kappa|y,\tau)d\kappa}{\int_\epsilon^1 p(\kappa|y,\tau)d\kappa} &\leqslant \frac{\int_0^\epsilon (1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa}{\int_\epsilon^1 (1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa} \\
&\leqslant \frac{(1-\epsilon)^{-\frac{3}{2}}\epsilon}{\exp(-\frac{y^2}{2})\int_\epsilon^1\{1-(1-\tau^2)\kappa\}^{-\frac{3}{2}}d\kappa} \\
&= \frac{\epsilon(1-\epsilon)^{-\frac{3}{2}}\exp(\frac{y^2}{2})}{\frac{2}{1-\tau^2}\{\frac{1}{\tau}-\frac{1}{\sqrt{1-(1-\tau^2)\epsilon}}\}} \\
&\leqslant \frac{1}{2}\epsilon(1-\epsilon)^{-\frac{3}{2}}\exp(\frac{y^2}{2})\tau(1+o(1)).
\end{aligned}$$

The proof follows from the fact that $P(\kappa < \epsilon|y,\tau)$ is bounded above by $\{P(\kappa < \epsilon|y,\tau)/P(\kappa > \epsilon|y,\tau)\}$. □

The upper bound in Theorem 3.1 will help us in deriving an asymptotic expression of type I error as we shall see in the next section.

We present a second concentration inequality on the posterior distribution of κ which might give some insight on the other side of the spectrum, i.e. the conditions under which the posterior mass of κ will concentrate near zero.

Theorem 3.2. *$P(\kappa > \eta|y,\tau) \leqslant \frac{2(1-\eta)^{\frac{1}{2}}\exp(-\frac{\eta y^2}{2}(1-\delta))}{\tau^2\eta\delta}$ for any fixed $\tau \in (0,1)$, any fixed $\eta \in (0,1)$, any fixed $\delta \in (0,1)$ and uniformly in $y \in \mathcal{R}$.*

Theorem 3.2 implies that for any fixed $\eta, \tau, \delta \in (0,1)$, $P(\kappa > \eta|y,\tau)$ decays at an exponential rate depending on y. The following corollary formalizes the notion.

Corollary 3.2. *$P(\kappa < \eta|y,\tau) \to 1$ as $y \to \infty$, for any fixed τ, any fixed δ and any fixed $\eta \in (0,1)$.*

Proof. (Theorem 3.2)

$$\begin{aligned}
\int_0^\eta p(\kappa|y,\tau)d\kappa &= c(y,\tau)\int_0^\eta (1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa \\
&\geqslant c(y,\tau)\int_0^{\eta\delta}\exp(-\frac{\kappa y^2}{2})d\kappa \quad \text{for any } \delta \in (0,1) \\
&\geqslant c(y,\tau)\exp(-\frac{\eta\delta y^2}{2})\eta\delta
\end{aligned}$$

and

$$\begin{aligned}\int_\eta^1 p(\kappa|y,\tau)d\kappa &= c(y,\tau)\int_\eta^1 (1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa \\ &\leqslant c(y,\tau)\exp(-\frac{\eta y^2}{2})\frac{1}{\tau^2}\int_\eta^1 \frac{d\kappa}{(1-\kappa)^{\frac{1}{2}}} \\ &= c(y,\tau)\frac{1}{\tau^2}\exp(-\frac{\eta y^2}{2})2(1-\eta)^{\frac{1}{2}}.\end{aligned}$$

So, we have the ratio of the two integrals

$$\frac{P(\kappa>\eta|y,\tau)}{P(\kappa<\eta|y,\tau)} = \frac{\int_\eta^1 p(\kappa|y,\tau)d\kappa}{\int_0^\eta p(\kappa|y,\tau)d\kappa} \leqslant \frac{2(1-\eta)^{\frac{1}{2}}\exp(-\frac{\eta y^2}{2}(1-\delta))}{\tau^2\eta\delta}.$$

The proof follows from the fact that $P(\kappa > \eta|y,\tau)$ is bounded above by $\{P(\kappa > \eta|y,\tau)/P(\kappa<\eta|y,\tau)\}$. □

3.3 Type I error

By (4) the type I error is given by

$$t_1 = P_{H_0:Y\sim\mathcal{N}(0,1)}\left(E(\kappa|Y,\tau) < \frac{1}{2}\right).$$

Theorem 3.3. *The probability of type I error for the horseshoe decision rule is given by:*

$$t_1 = \frac{2\tau}{\sqrt{\ln(1/\tau)}}(1+o(1)) \text{ as } \tau\to 0.$$

Proof. Since the posterior mass of κ concentrates near 1 when τ is very small and y is not too large, the posterior mean of κ should be well approximated by taking the average with respect to the posterior mass near 1. Towards proving this, we first write the expectation as the sum of two different integrals, one near zero and the other near 1, and show that the sum is dominated by the second term, i.e.

$$\begin{aligned}E(\kappa|y,\tau) &= \int_0^{\frac{1}{2}}\kappa p(\kappa|y,\tau)d\kappa + \int_{\frac{1}{2}}^1 \kappa p(\kappa|y,\tau)d\kappa &(9)\\ &= \int_{\frac{1}{2}}^1 \kappa p(\kappa|y,\tau)d\kappa(1+o(1)) \text{ as } \tau\to 0. &(10)\end{aligned}$$

To prove (10) we take the ratio of the two integrals in (9) and show that it converges

to zero as τ converges to zero:

$$\begin{aligned}\frac{\int_0^{\frac{1}{2}} \kappa p(\kappa|y,\tau)d\kappa}{\int_{\frac{1}{2}}^1 \kappa p(\kappa|y,\tau)d\kappa} &= \frac{\int_0^{\frac{1}{2}} \kappa(1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa}{\int_{\frac{1}{2}}^1 \kappa(1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa} \\ &\leqslant \frac{2^{-5/2}\frac{1}{\frac{1}{2}+\frac{1}{2}\tau^2}}{\exp(-\frac{y^2}{2})\int_{\frac{1}{2}}^1 \kappa(1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}d\kappa} \\ &\leqslant \frac{2^{-3/2}}{\exp(-\frac{y^2}{2})\frac{1}{2}\int_{\frac{1}{2}}^1 \{1-(1-\tau^2)\kappa\}^{-3/2}d\kappa} \\ &\leqslant \frac{2^{-3/2}}{\exp(-\frac{y^2}{2})\frac{1}{1-\tau^2}(\frac{1}{\tau}-\frac{\sqrt{2}}{\sqrt{1+\tau^2}})} = 2^{-\frac{3}{2}}\mathrm{e}^{\frac{y^2}{2}}\tau(1+o(1)).\end{aligned}$$

The right hand side of the last equation goes to zero as $\tau \to 0$, thus proving the approximation in (10). Now, we will show that the integral in (10) can be well approximated by the posterior mass in the same region as the posterior density of κ is concentrated near 1, i.e. we prove the following:

$$\int_{\frac{1}{2}}^1 \kappa p(\kappa|y,\tau)d\kappa = \int_{\frac{1}{2}}^1 p(\kappa|y,\tau)d\kappa(1-o(1)) \tag{11}$$

$$\equiv \frac{\int_{\frac{1}{2}}^1 (1-\kappa)p(\kappa|y,\tau)d\kappa}{\int_{\frac{1}{2}}^1 p(\kappa|y,\tau)d\kappa} \to 0 \text{ as } \tau \to 0. \tag{12}$$

To prove (11) we note that the numerator in the ratio of the two integrals in (12) remains bounded whereas the integral in the denominator diverges near $\kappa = 1$ as the integrand $(1-(1-\tau^2)\kappa)^{-3/2}$ behaves like $(1-\kappa)^{-3/2}$ when $\tau \to 0$. The proof is as follows:

$$\begin{aligned}\frac{\int_{\frac{1}{2}}^1 (1-\kappa)p(\kappa|y,\tau)d\kappa}{\int_{\frac{1}{2}}^1 p(\kappa|y,\tau)d\kappa} &= \frac{\int_{\frac{1}{2}}^1 (1-\kappa)^{\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa}{\int_{\frac{1}{2}}^1 (1-\kappa)^{-\frac{1}{2}}\{1-(1-\tau^2)\kappa\}^{-1}\exp(-\frac{\kappa y^2}{2})d\kappa} \\ &\leqslant \frac{\exp(\frac{y^2}{4})\int_{\frac{1}{2}}^1 (1-\kappa)^{-\frac{1}{2}}d\kappa}{\int_{\frac{1}{2}}^1 \{1-(1-\tau^2)\kappa\}^{-\frac{3}{2}}d\kappa}.\end{aligned}$$

The integral in the denominator will converge to the divergent integral $\int_{\frac{1}{2}}^1 (1-\kappa)^{-\frac{3}{2}}d\kappa$ as $\tau \to 0$. Finally, from (10) and (11) we get $E(\kappa|y,\tau) = \int_{\frac{1}{2}}^1 p(\kappa|y,\tau)d\kappa(1-o(1))$. Therefore, from (4), the type 1 error for the horseshoe estimator can be written as,

$$\begin{aligned}t_1 &= P_{H_0}(\textstyle\int_{\frac{1}{2}}^1 p(\kappa|Y,\tau)d\kappa < \frac{1}{2})(1+o(1)) \\ &= P_{H_0}(\textstyle\int_0^{\frac{1}{2}} p(\kappa|Y,\tau)d\kappa > \frac{1}{2})(1+o(1)).\end{aligned} \tag{13}$$

We use Theorem 3.1 with $\epsilon = \frac{1}{2}$ to derive a tight upper bound to the integral in (13):

$$\int_0^{\frac{1}{2}} p(\kappa|y,\tau)d\kappa \leqslant \frac{\mathrm{e}^{\frac{y^2}{2}}\tau}{\sqrt{2}}(1+o(1)) \quad \text{uniformly in } y \in \mathcal{R}.$$

Plugging in this upper bound in (13) we get,

$$\begin{aligned} t_1 &= P_{H_0}\left[\int_0^{\frac{1}{2}} p(\kappa|Y,\tau)d\kappa > \frac{1}{2}\right](1+o(1)) \\ &= P_{H_0}\left[\frac{\mathrm{e}^{\frac{Y^2}{2}}\tau}{\sqrt{2}} > \frac{1}{2}\right](1+o(1)) \\ &= P_{H_0}\left[|Y| > \sqrt{2\ln\frac{1}{\sqrt{2}\tau}}\right](1+o(1)) \\ &\approx \frac{\phi(\sqrt{2\ln\frac{1}{\sqrt{2}\tau}})}{\sqrt{2\ln\frac{1}{\sqrt{2}\tau}}}(1+o(1)) = \frac{2\tau}{\sqrt{\ln(1/\tau)}}(1+o(1)). \end{aligned}$$

We use the well-known normal tail approximation $\frac{1-\Phi(x)}{\phi(x)} = \frac{1}{x} + O(\frac{1}{x^3})$ for deriving the final approximation. □

3.4 Type II error

By (5) the type II error is given by

$$t_2 = P_{Y\sim\mathcal{N}(0,1+\psi^2)}(E(\kappa|Y,\tau) > \frac{1}{2}).$$

Choice of τ. *We choose the free tuning parameter τ to be of the same order of p, this makes the type II error probability for the horseshoe comparable to the oracle. In numerical work we do not assume p known and use the Full Bayes estimate of τ, which seems to adapt to the level of sparsity in the data (vide Figure 3).*

For such a choice of $\tau = \tau_m = O(p_m)$, it can be easily seen from Assumption 3.1 that $\frac{\log(1/\tau_m^2)}{\psi_m^2} \to C \in (0,\infty), \tau_m \to 0, \psi_m^2 \to \infty$ as $m \to \infty$, where C is the constant appearing in Assumption 3.1 and the formula for the Bayes risk for the oracle (vide equation (8)).

Theorem 3.4. *If the "global shrinkage parameter" τ of the horseshoe prior is chosen to be of the same order as the proportion of signals p, then the type II error rate of the horseshoe decision rule will be given by*

$$t_2 = (2\Phi(3\sqrt{C}) - 1)(1+o_m).$$

As before, o_m denotes a sequence of terms indexed by m, converging to zero as $m \to \infty$. Also, for ease of notation we will suppress the index 'm' from the parameters for the rest of the paper.

Proof. We prove Theorem 3.4 in three steps as described below:

1. Corollary 3.2 and the posterior density plots in Polson and Scott (2010) suggest that the posterior distribution of κ converges to a degenerate distribution at zero as $y \to \infty$ for a fixed τ. Thus, using the concentration inequalities in Theorem 3.2, we can find $\delta(y, \tau)$ such that,

$$\int_0^{\frac{1}{4}} p(\kappa|y, \tau)d\kappa \geqslant 1 - \delta(y, \tau) \text{ for a fixed } y \text{ and } \tau.$$

2. Now, under $H_A : Y \sim \mathcal{N}(0, 1 + \psi^2)$, we find $\upsilon(\psi, \tau)$ such that

$$P_{H_A}(\delta(Y, \tau) > 1/4) = \upsilon(\psi, \tau).$$

   Then we use the inequality

$$\kappa < \mathrm{I}\{\frac{1}{4} \leqslant \kappa \leqslant 1\} + \frac{1}{4}$$

   where $\mathrm{I}\{\kappa \in [\frac{1}{4}, 1]\}$ is the indicator function taking value 1 for all values of $\kappa \in [\frac{1}{4}, 1]$ and 0 otherwise, to deduce that,

$$\begin{aligned} &E(\kappa|Y = y, \tau) < \delta(y, \tau) + \frac{1}{4} \\ \Rightarrow\ &P_{H_A}(E(\kappa|Y, \tau) > \frac{1}{2}) \leqslant P_{H_A}(\delta(Y, \tau) > \frac{1}{4}) = \upsilon(\psi, \tau). \end{aligned}$$

3. Then we show that $\upsilon(\psi, \tau) = (2\Phi(3\sqrt{C}) - 1)(1 + o(1))$ as $\frac{\log(1/\tau^2)}{\psi^2} \to C \in (0, \infty)$ for our choice of the free tuning parameter τ.

Step 1: Theorem 3.2 for $\eta = \frac{1}{4}$ and $\delta = \frac{1}{9}$ yields an explicit expression for $\delta(y, \tau)$ as follows:

$$\begin{aligned} &\int_{\frac{1}{4}}^{1} p(\kappa|y, \tau)d\kappa \leqslant \frac{72\exp(-\frac{y^2}{9})}{\tau^2} \\ \iff &\int_0^{\frac{1}{4}} p(\kappa|y, \tau)d\kappa \geqslant 1 - \frac{72\exp(-\frac{y^2}{9})}{\tau^2}. \end{aligned} \tag{14}$$

Step 2: From (14), $\delta(y,\tau) = \frac{72\exp(-\frac{y^2}{9})}{\tau^2}$. Then $\upsilon(\psi,\tau)$ is immediately obtained as:

$$\begin{aligned} P_{H_1}(\delta(Y,\tau) > \frac{1}{4}) &= P_{H_A}\left[\frac{72\exp(-\frac{y^2}{9})}{\tau^2} > \frac{1}{4}\right] \\ &= P_{H_A}\left[|Y| < 3\sqrt{\ln 288 + 2\ln(1/\tau)}\right] \\ &= P_{H_A}\left[|Z| < 3\sqrt{\frac{\ln 288 + 2\ln(1/\tau)}{1+\psi^2}}\right] \\ &\text{where, } Z = \tfrac{Y}{\sqrt{1+\psi^2}} \sim \mathcal{N}(0,1) \\ &= \left[2\Phi\left(3\sqrt{\frac{\ln(1/\tau^2)}{\psi^2}}\right) - 1\right](1+o(1)) = \upsilon(\psi,\tau). \end{aligned} \tag{15}$$

Step 3: Proof of step 3 follows immediately from (15) and the fact that under the choice of τ as $\tau = O(p)$, $\frac{\log(1/\tau^2)}{\psi^2} \to C \in (0,\infty)$. □

3.5 Optimality of the horseshoe decision rule

The Bayes risk for a multiple testing procedure for the two-groups model under an additive 0-1 loss function is given by

$$R = m\{(1-p)t_1 + pt_2\}.$$

Therefore, under our choice of τ, the type I and type II error probabilities and the Bayes risk for the horseshoe decision rule (R_{HS}) are given by:

$$\begin{aligned} t_1 &= \frac{2\tau}{\sqrt{\ln(1/\tau)}}(1+o_m) \\ t_2 &= (2\Phi(3\sqrt{C}) - 1)(1+o_m) \\ R_{HS} &= mp(2\Phi(3\sqrt{C}) - 1)(1+o_m). \end{aligned} \tag{16}$$

Under the asymptotic framework defined by Assumption 3.1, we saw that the constants for both the horseshoe risk and the Bayes oracle are the same provided the global shrinkage parameter τ in horseshoe prior is chosen to be of the same order as the sparsity parameter p in the two-groups model. Also, from equations (8) and (16), the ratio of the risks for the horseshoe decision rule and the Bayes oracle is given by:

$$\begin{aligned} \frac{R_{HS}}{R_{opt}} &= \frac{(2\Phi(3\sqrt{C}) - 1)}{(2\Phi(\sqrt{C}) - 1)}(1+o_m) \\ R_{HS} &= O(R_{opt}) \text{ as } m \to \infty. \end{aligned}$$

4 Numerical Results

We compared the performance of the decision rule imposed by the horseshoe and the modified horseshoe prior (vide Section 2.1) with other shrinkage priors, viz. the double exponential prior (DE) (vide Section 2.2), the Benjamini-Hochberg rule (BH) and the Bayes oracle (BO) in terms of the misclassification probability (MP). We simulated data of size $m = 200, \psi_m = \sqrt{2 \log m} = 3.26$. This plot corresponds to the misclassification probability plot given in Figure 1 of Bogdan et al. (2008).

Figure 1 shows the misclassification probabilities (henceforth abbreviated as MP) for different shrinkage priors considered for ten equispaced values of $p \in [0.01, 0.5]$ along with the oracle and the straight line (MP $= p$). The oracle serves as the lower bound and the $MP = p$ line is basically the misclassification probability for the decision rule that rejects all the null hypotheses without looking at the data. The right pane of the figure shows that for both double exponential and normal priors, the MP plot hugs the MP $= p$ line. The performance of the horseshoe and its modified version is shown in the left pane, which shows that the horseshoe priors attain an MP inferior to the oracle, but it is much better than the other candidates we considered, viz. double exponential and normal. We have also plotted the MP for the Benjamini-Hochberg rule, for $\alpha = 1/\log m = 0.1887$, along with the one-group shrinkage priors. Our plots show that the Benjamini-Hochberg rule achieves the same MP as the oracle under this setting. This is in concordance with the theoretical results for optimality of BH in Bogdan et al. (2011).

As discussed in Polson and Scott (2010), the modified horseshoe prior with posterior density of κ as

$$p(\kappa|y, s) \propto (1-\kappa)^{-\frac{1}{2}} \exp\{-(s + \frac{y^2}{2})\kappa\} \quad \kappa \in (0, 1)$$

is an alternative to the original horseshoe prior where the "global shrinkage" takes place through the parameter s instead of τ (vide Section 2.1). In our experiment, the modified horseshoe performs as well as the original horseshoe for the extremely sparse case and the original horseshoe has a slight advantage over the modified version for non-sparse data with moderately large values of $p \approx 0.5$.

We used the full Bayes estimates for the hyperparameters for both the horseshoe prior and the double exponential prior. For estimating τ, we assumed half-Cauchy prior on τ and Jeffreys prior for the variance $p(\sigma^2) \propto \frac{1}{\sigma^2}$ for deriving the full conditionals using a Gibbs sampler. As pointed out by Carvalho et al. (2009); Scott and Berger (2006), the fully Bayesian approach for estimating τ has a few adavantages over its alternatives, viz. empirical Bayes and cross-validation. In the extremely sparse case, the empirical Bayes estimate of τ might collapse to 0. This phenomenon is well-known, see Scott and Berger (2010) and Bogdan et al. (2008). Cross-validation, though free of this problem, uses plug-in estimates for the signal-to-noise ratio and hence does not take into account the potential correlation structure between τ and σ. Carvalho et al. (2009) argue that the plug-in estimates are not necessarily wrong, but caution should be exercised while using them for extremely sparse problems.

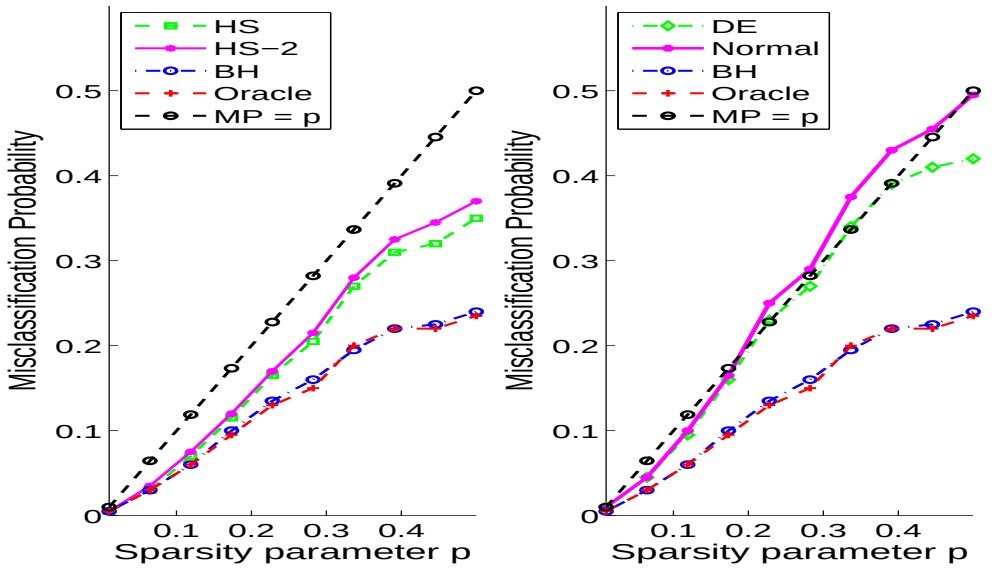

Figure 1: Misclassification probability plot for the Horseshoe, Laplacian and Normal shrinkage priors, Benjamini-Hochberg and the Bayes oracle for $p \in (0.1, 0.5)$.

It is worth noticing that the double exponential prior, though a good default candidate for shrinkage in the sparse signal situation, has higher MP than any other multiple testing procedure for the sparse case (vide Figure 1). Carvalho et al. (2009, 2010) pointed out the inability of the double exponential prior to effectively shrink noise through a comparison of squared error loss under different estimators. The results in Carvalho et al. (2010) show that the double exponential prior is consistently outperformed by the horseshoe in terms of squared error loss. The relatively poor performance by the double exponential may be attributed to insufficiently heavy exponential tails that tend to over-shrink the large signals and under-shrink the noisy observations at the same time. The $E(\theta|y, \hat{\tau})$ vs. y plot supports this observation (vide Figure 2). The authors argue that while the horsehoe density is symmetric and unbounded at both 0 and 1, the double exponential prior density is bounded at both 0 and 1, implying lack of total shrinkgae *a priori*. The horseshoe prior, on the other hand, does not face the same difficulty as the full Bayes estimate of τ is much smaller under the horseshoe than under the double exponential prior (vide Figure 3 of Carvalho et al. (2009)), in which case τ adapts to the sparsity level p, and heavy tails of $\pi(\lambda_i)$ segregate the signals from the noise.

We show the adaptivity of the global shrinkage parameter τ to the sparsity level of the data in Figure 3. We plot the posterior samples for τ for both sparse and non-sparse cases showing that the estimate changes with changing sparsity levels. It supports the claim of Scott and Berger (2006); Carvalho et al. (2010) that the global shared parameters control the error rates in multiple testing by estimating the overall

sparsity level, particularly when the parameters are estimated from the data.

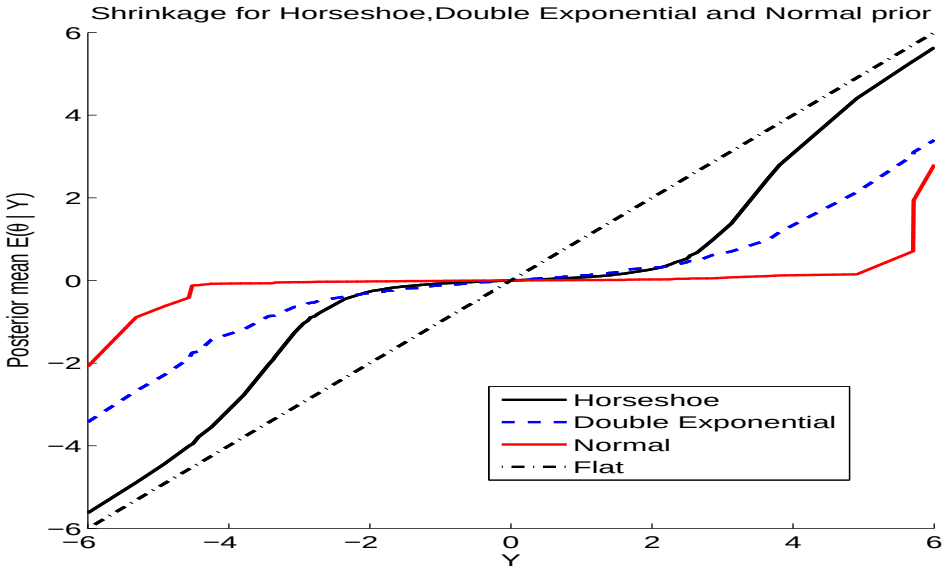

Figure 2: Posterior Mean $E(\theta|y)$ versus y plot for $p = 0.25$.

Finally, we show how well the posterior inclusion probability for the more complex two groups model can be approximated by the proxy posterior inclusion probability of the one group horseshoe model for $p = 0.1$ and $p = 0.5$. We calculated the shrinkage weights $1 - \hat{\kappa_i}, i = 1, \ldots, n$ for all the observations generated from a two-groups model with the level of sparsity $p = 0.1$ and compared them against the theoretical posterior inclusion probability for the two-groups model (1-2) given by:

$$\omega_i = P(\theta_i \neq 0|y_i) = \left\{ (\frac{1-p}{p})\sqrt{1+\psi^2}\exp(-\frac{y_i^2}{2}\frac{\psi^2}{1+\psi^2}) + 1 \right\}^{-1}.$$

The resulting plot in Figure 4 shows that for small $p = 0.1$, the theoretical posterior inclusion probability is well captured by the proxy posterior inclusion probability imposed by the horseshoe decision rule and the approximation is not so good for larger $p = 0.5$. This might give some insight to the success of horseshoe priors in achieving similar risk as a two-groups model in the sparse case.

5 Discussion

We have examined the asymptotic risk properties of the horseshoe estimator due to Carvalho et al. (2009) from a Bayesian decision theoretic viewpoint. We prove that under the asymptotic framework of Bogdan et al. (2011), the natural decision rule

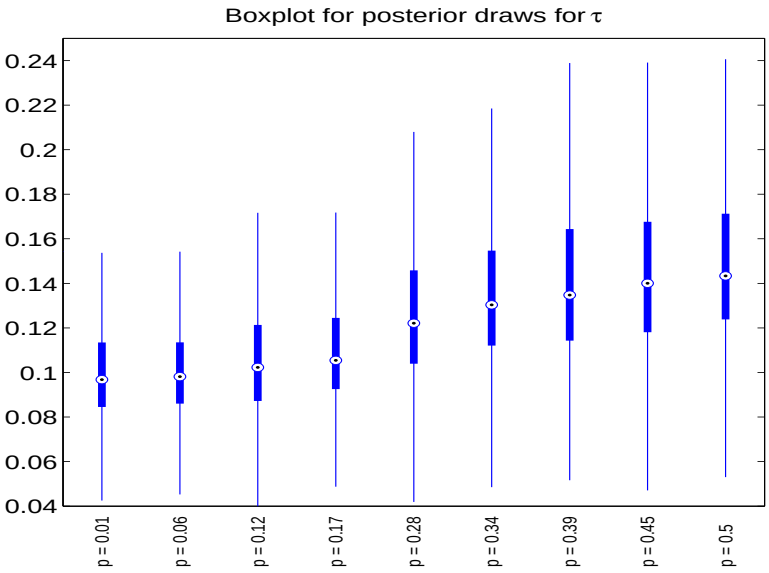

Figure 3: Posterior draws for τ at different levels of the sparsity parameter p.

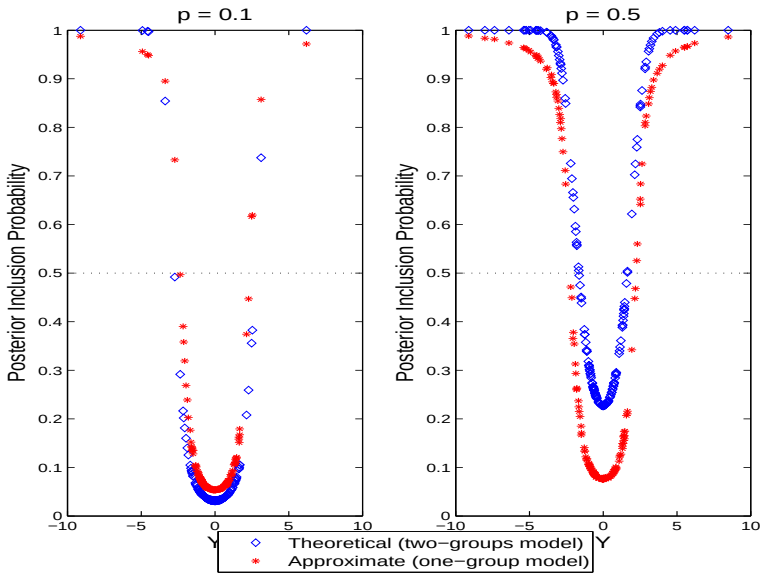

Figure 4: Comparison of the theoretical posterior inclusion probability for the two-groups model and the shrinkage weight $(1 - \hat{\kappa})$ for the one-group horseshoe model for different values of Y.

induced by the horseshoe prior attains the risk of the Bayes oracle up to $O(1)$ with a constant close to the constant in the oracle if the shrinkage parameter τ is chosen to be of the same order as the proportion p of signals. We provide numerical evidence that the optimality holds if we use the full Bayes estimate of τ, which seems to adapt to the unknown sparsity in the data. Our theoretical results provide some insights into the asymptotic behavior of the continuous one-group model in the context of multiple testing. The theoretical as well as the numerical results support the observation of Carvalho et al. (2010) that the one-group model can closely mimic the results from Bayesian methods on the two-groups model when the global shrinkage parameter is suitably tuned to handle the sparsity in the data.

We will briefly discuss why the horseshoe prior doesn't do as well as the Benjamini-Hochberg rule. We feel that this is due to the fact that the former is based on an approximate one group formulation whereas the latter is an optimal procedure in a two-groups setting. This is why we can prove optimality up to a constant and argue that the two groups oracle is almost attained by the horseshoe decision rule when the constant C is moderately large. Here, C is the constant appearing in Assumption 3.1 and also in the formula for horseshoe risk and oracle risk (vide equations (8) and (16)). We provide a comparison of the misclassification probability for the horseshoe and the oracle along with their ratio, in Table (1). Based on the numerical study reported in Section 4, we show that the horseshoe decision rule closely attains the oracle risk for smaller p and doesn't perform too badly for moderate $p \approx 0.5$. Note that smaller p implies bigger C, when ψ^2, the non-null variance, is held fixed.

Sparsity p	Horseshoe MP	Bayes oracle MP	$\frac{\text{Horseshoe MP}}{\text{Oracle MP}}$	$\frac{\log(1/p^2)}{\psi^2} \approx C$
0.01	0.005	0.005	1.00	0.87
0.0644	0.03	0.03	1.00	0.52
0.1189	0.07	0.06	1.17	0.40
0.1733	0.115	0.095	1.21	0.33
0.2278	0.165	0.13	1.27	0.28
0.2822	0.205	0.15	1.37	0.24
0.3367	0.27	0.2	1.35	0.21
0.3911	0.31	0.22	1.41	0.18
0.4456	0.32	0.22	1.45	0.15
0.5	0.35	0.235	1.49	0.13

Table 1: Misclassification probability for the horseshoe and the Bayes oracle for ten equidistant values of $p \in [0.01, 0.5]$. It shows that in practice the horseshoe decision rule closely attains the oracle for small p and does reasonably well even for moderate $p \approx 0.5$.

It is indeed true that for many priors a result like Theorem 3.4 will hold, but characterizing these priors may not be easy. We expect that the priors for which such a theorem holds will be shrinkage prior with similar shrinkage as for the horseshoe. However, our proof can be applied only if we can prove similar contraction properties of the posterior densities near 0 and 1. Though the double exponential prior is known to squelch noise observations to zero (vide Pericchi and Smith (1992); Tibshirani (1996);

Park and Casella (2008); Hans (2009); Li and Goel (2006)), our method of proof fails to show any contraction property for the double exponential prior. The misclassification probability is much higher for the double exponential, in fact it is close to the MP for the trivial test that rejects all the null hypotheses without looking at the data. This might be attributed to the over-shrinkage of the large signals by the double exponential prior, as shown in Figure 2. For the double pareto even the algebra for computing the posterior given the global parameter τ is too formidable for any theoretical or numerical computation, at least as far as we could see. So it remains an interesting open question.

We would like to shed some light on the shrinkage profile for the normal prior and argue that this might be an example of a near non-optimal shrinkage. We put a standard half-normal prior on λ_i, i.e.

$$\lambda_i \sim N^+(0,1). \tag{17}$$

This allows us to work with the same shrinkage coefficent $\kappa_i = \frac{1}{1+\lambda_i^2\tau^2}$ which gives

$$E(\theta_i|y_i,\tau,\sigma^2) = (1 - E(\kappa_i|y_i,\tau,\sigma^2))y_i.$$

For the normal prior, like the horseshoe prior, posterior shrinkage comes through the quantity $\hat{\kappa_i} = E(\kappa_i|y_i,\tau,\sigma^2)$. The quantity $\hat{\kappa}$ can be seen as the posterior shrinkage towards zero, and a higher value of this near the tails means over-shrinking of the large coefficients. We show the posterior mean $E(\theta|Y)$ vs. Y plots for the horseshoe, the double exponential and the normal prior (vide Figure 2) for data originating from a two-groups model with $m = 200$ and $p = 0.25$. It shows that the normal prior has the strongest shrinkage and shrinks even the larger signals to zero, unlike the horseshoe which performs well near the origin as well as away from it.

We have examined the very powerful contraction inequalities obtained in Armagan, Dunson, Lee, and Bajwa (2011); Pati, Bhattacharya, Pillai, and Dunson (2012); Strawn, Armagan, Saab, Carin, and Dunson (2012). We believe these results will have many applications in Bayesian analysis and are indebted to the referees for pointing out these new articles. However, these do not seem to apply to our setup because the asymptotics of the contraction are very different. In the case of Strawn et al. (2012), contraction of the posterior comes by increasing the sample size, though there are other factors including choice of prior. However, in Bogdan et al. (2011) or the present paper, posterior contraction comes from the hyper-parameter p going to 0 and hyper-parameter ψ^2 going to infinity. Note that ψ^2 is the non-null variance, which acts like a non-centrality parameter, measuring the strength of the signal. The asymptotics for the two groups model is in Bogdan et al. (2011). The asymptotic calculations of the present paper are different and based on the two results on the concentration of the posterior, but the framework of asymptotics is the same, coming through p and ψ^2. Ours is a modest contribution to asymptotics about multiple testing of means in the one group problem. The corresponding asymptotic results in the two groups model (vide Bogdan et al. (2011)) have a very different proof.

There is some similarity with Armagan et al. (2011) but their assumptions neither imply ours, nor are implied by our assumptions. In a rather different direction, Pati et al.

(2012) have considered similar reults in the case of inference on covariance matrices, but unlike our problems, the means are assumed to be zero.

References

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2006). "Adapting to unknown sparsity by controlling the false discovery rate." *The Annals of Statistics*, 34(2): 584–653. 114

Armagan, A., Dunson, D., Lee, J., and Bajwa, W. (2011). "Posterior consistency in linear models under shrinkage priors." *Arxiv preprint arXiv:1104.4135*. 129

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300. 111

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). "Asymptotic Bayes-optimality under sparsity of some multiple testing procedures." *The Annals of Statistics*, 39(3): 1551–1579. 111, 112, 114, 116, 117, 124, 126, 129

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). "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 of *Institute of Mathematical Statistics Collections*, 211–230. 112, 114, 124

Carvalho, C., Polson, N., and Scott, J. (2009). "Handling sparsity via the horseshoe." *Journal of Machine Learning Research W&CP*, 5(73-80). 111, 112, 113, 114, 124, 125, 126

— (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. 111, 112, 113, 114, 115, 125, 128

Diaconis, P., Goel, S., and Holmes, S. (2008). "Horseshoes in multidimensional scaling and local kernel methods." *The Annals of Applied Statistics*, 2(3): 777–807. 113

Efron, B. (2004). "Large-scale simultaneous hypothesis testing." *Journal of the American Statistical Association*, 99(465): 96–104. 112

— (2008). "Microarrays, empirical Bayes and the two-groups model." *Statistical Science*, 23(1): 1–22. 111, 112

Genovese, C. and Wasserman, L. (2004). "A stochastic process approach to false discovery control." *The Annals of Statistics*, 32(3): 1035–1061. 112

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 115, 129

Li, B. and Goel, P. K. (2006). "Regularized optimization in statistical learning: a Bayesian perspective." *Statistica Sinica*, 16(2): 411–424. 115, 129

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 115, 129

Pati, D., Bhattacharya, A., Pillai, N., and Dunson, D. (2012). "Posterior contraction in sparse Bayesian factor models for massive covariance matrices." *Arxiv preprint arXiv:1206.3627*. 129

Pericchi, L. R. and Smith, A. F. M. (1992). "Exact and approximate posterior moments for a normal location parameter." *Journal of the Royal Statistical Society. Series B. Methodological*, 54(3): 793–804. 115, 128

Polson, N. and Scott, J. (2010). "Large-scale simultaneous testing with hypergeometric inverted-beta priors." *Arxiv preprint arXiv:1010.5223*. 114, 115, 122, 124

Polson, N. G. and Scott, J. G. (2012). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis*, 7(2): 1–16. 112

Sarkar, S. (2006). "False discovery and false nondiscovery rates in single-step multiple testing procedures." *The Annals of Statistics*, 34(1): 394–415. 112

Scott, J. and Berger, J. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. 112, 124, 125

— (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. 112, 124

Scott, J. G. (2011). "Bayesian estimation of intensity surfaces on the sphere via needlet shrinkage and selection." *Bayesian Analysis*, 6(2): 307–327. 112, 115

Sen, B., Banerjee, M., and Woodroofe, M. (2010). "Inconsistency of bootstrap: The Grenander estimator." *The Annals of Statistics*, 38(4): 1953–1977. 112

Storey, J. (2007). "The optimal discovery procedure: a new approach to simultaneous significance testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3): 347–368. 112

Storey, J. D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value." *The Annals of Statistics*, 31(6): 2013–2035. 112

Strawn, N., Armagan, A., Saab, R., Carin, L., and Dunson, D. (2012). "Finite sample posterior concentration in high-dimensional regression." *Arxiv preprint arXiv:1207.4854*. 129

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B. Methodological*, 58(1): 267–288. 115, 128

Acknowledgments

The authors would like to thank an anonymous referee, the editor, the associate editor, and the editor-in-chief, whose many constructive comments have led to substantial improvement in our presentation and some improvement in our results.