Large-Scale Multiple Hypothesis Testing with the Normal-Beta Prime Prior

Ray Bai [a] and Malay Ghosh [b]

[a] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA 19104 USA; [b] Department of Statistics, University of Florida, Gainesville, FL 32611 USA

ARTICLE HISTORY
Compiled August 29, 2019

ABSTRACT
We revisit the problem of simultaneously testing the means of n independent normal observations under sparsity. We take a Bayesian approach to this problem by studying a scale-mixture prior known as the normal-beta prime (NBP) prior. To detect signals, we propose a hypothesis test based on thresholding the posterior shrinkage weight under the NBP prior. Taking the loss function to be the expected number of misclassified tests, we show that our test procedure asymptotically attains the optimal Bayes risk when the signal proportion p is known. When p is unknown, we introduce an empirical Bayes variant of our test which also asymptotically attains the Bayes Oracle risk in the entire range of sparsity parameters $p \propto n^{-\epsilon}, \epsilon \in (0, 1)$. Finally, we also consider restricted marginal maximum likelihood (REML) and hierarchical Bayes approaches for estimating a key hyperparameter in the NBP prior and examine multiple testing under these frameworks.

KEYWORDS
Bayes oracle, empirical Bayes, multiple testing, shrinkage prior, sparsity

1. Introduction

1.1. Large-Scale Testing of Normal Means

Suppose we observe an n-component random observation $(X_1, \ldots, X_n) \in \mathbb{R}^n$, such that

$$X_i \sim \mathcal{N}(\theta_i, 1), \;\; i = 1, \ldots, n. \tag{1}$$

This simple framework is the basis for a number of high-dimensional problems, including genetics, wavelet analysis, and image reconstruction [16]. Under model (1), we are primarily interested in identifying the few signals ($\theta_i \neq 0$). This amounts to performing n simultaneous tests, $H_{0i} : \theta_i = 0$ vs. $H_{1i} : \theta_i \neq 0$, $i = 1, \ldots, n$.

In the high-dimensional setting where n is very large, sparsity is a very common phenomenon. In genetics, for example, the X_i's may represent thousands of gene expression data points, but only a few genes are significantly associated with the phenotype of

CONTACT Ray Bai. Email: Ray.Bai@pennmedicine.upenn.edu

interest. For instance, [26] has confirmed that only seven genes have a non-negligible association with Type I diabetes.

1.2. Scale-Mixture Shrinkage Priors

Scale-mixture shrinkage priors are widely used for obtaining (nearly) sparse estimates of $\boldsymbol{\theta}$ in (1). These priors take the form,

$$\theta_i|\sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2), \;\; \sigma_i^2 \sim \pi(\sigma_i^2), \;\; i = 1, \ldots, n, \tag{2}$$

where $\pi : [0, \infty) \to [0, \infty)$ is a density on the positive reals. These priors typically contain heavy mass around zero, so that the posterior density is heavily concentrated around $\mathbf{0} \in \mathbb{R}^n$. However, they also retain heavy enough tails in order to correctly identify and prevent overshrinkage of the true signals. Examples of (2) include the popular horseshoe prior [9] and the Bayesian lasso [17]. Priors of the type (2) have also been considered by numerous other authors: see, e.g. [2,4–6,15,22].

From (2), we see that the posterior mean of θ_i under these priors is given by

$$\mathbb{E}\{\mathbb{E}(\theta_i|X_i, \sigma_i^2)\} = \{\mathbb{E}(1 - \kappa_i)|X_1, \ldots, X_n\}\, X_i, \tag{3}$$

where $\kappa_i = 1/(1 + \sigma_i^2)$. By (3), it is clear that the shrinkage weight κ_i plays a crucial role in the amount of posterior shrinkage under these priors.

1.3. Multiple Testing Under Sparsity

Assuming that the true data-generating model is a two-components mixture density, [7] studied the risk properties of a large number of multiple testing rules. Specifically, [7] considered a symmetric 0-1 loss function taken to be the expected total number of misclassified tests. Under mild conditions, [7] arrived at a simple closed form for the asymptotic Bayes risk under this loss. They termed this as the asymptotically Bayes optimal risk under sparsity (ABOS), or the Bayes Oracle risk. They then provided necessary and sufficient conditions for which a number of classical multiple test procedures (e.g. the Bonferroni correction or the Benjamini-Hochberg [3] procedure) could asymptotically equal the Bayes Oracle risk. A thorough discussion of this decision theoretic framework is presented in Section 3.1.

Testing rules induced by scale-mixture shrinkage priors have also been studied within this decision theoretic framework. Since scale-mixture shrinkage priors of the form (2) are absolutely continuous, they place zero mass at exactly zero. Thus, in order to classify means as either signal or noise, some thresholding rule must be applied. One method of doing this is by thresholding the posterior shrinkage weight κ_i in (3) as follows. For the ith component, the test procedure based on κ_i is:

$$\text{Reject } H_{0i} \text{ if } \mathbb{E}(1 - \kappa_i|X_1, \ldots, X_n) > \frac{1}{2}. \tag{4}$$

Depending on how conservative the test must be, the fraction $1/2$ can be replaced by any $\alpha \in (0, 1)$, and then the final results will depend on α. However, for most practical applications, it seems as though this 'half-thresholding' rule of $1/2$ is sensible [9,11,14].

Assuming that the θ_i's come from a two-components model, [11] showed that rule (4) under the horseshoe prior asymptotically attains the Bayes Oracle risk up to a

multiplicative constant. [14] generalized this result to a general class of shrinkage priors of the form,

$$\theta_i|\tau, \lambda_i \sim \mathcal{N}(0, \lambda_i\tau), \;\; \lambda_i \sim \pi(\lambda_i) = K\lambda_i^{-a-1}L(\lambda_i), \tag{5}$$

where $\tau > 0$ is a variance rescaling parameter, K is the constant of proportionality, $a > 0$, and $L(\cdot)$ is a measurable, nonconstant, slowly varying function. [13] later showed that thresholding rule (4) for this same class of priors (5) could even asymptotically attain the exact Bayes Oracle risk. [5] also extended the same rule for the horseshoe+ prior, showing that rule (4) based on the horseshoe+ prior asymptotically attains the Bayes Oracle risk up to a multiplicative constant.

Recently, [19] studied testing rule (4) under an even broader class of normal scale-mixture shrinkage priors (2) which subsumes priors of the form (5). In this class, the prior on the scale parameter σ_i^2, $\pi(\sigma_i^2)$, satisfies the three properties given in [23]. The properties in [23] are sufficient for scale-mixture priors to obtain the minimax posterior contraction rate under the sparse normal means model (1). For priors satisfying these conditions, [19] derived upper bounds on the asymptotic Bayes risk for both non-adaptive and data-adaptive testing rules. He showed that the upper bound on the Bayes risk for this general class of priors is of the same order as the Bayes Oracle risk up to a multiplicative constant.

The results in this manuscript were developed independently of [19] and give sharper bounds than those of [19]. [19] did not obtain the exact asymptotic Bayes Oracle risk nor did he derive asymptotic lower bounds on the Type I and Type II errors or the Bayes risk. In contrast, our paper establishes tight upper *and* lower bounds. To further highlight the distinction, we refer to testing rules as having the Bayes Oracle property if and only if they can be shown to asymptotically obtain the exact Bayes Oracle risk in [7]. Further, the prior that we propose in this paper departs from the family of priors (5) considered by [13] because it does not require a variance rescaling parameter $\tau > 0$. Therefore, our results also do not automatically follow from those of [13].

In this article, we consider a Bayesian scale-mixture shrinkage prior with the beta prime density as its scale parameter and no variance rescaling parameter τ. We call our model the normal-beta prime (NBP) model. We highlight some of our contributions:

(1) We investigate the properties of the NBP model with *varying* hyperparameters (a, b). Since we allow the hyperparameters to vary with the sample size, the concentration inequalities for the beta prime hierarchical model established in Section 2 are new, and thus, may be of independent interest for Bayesian inference involving the beta prime density as a prior.
(2) We derive both lower and upper bounds on Type I and Type II probabilities under thresholding rules based on the NBP's posterior shrinkage factor. We show that with appropriate choices of (a, b), our method asymptotically achieves the Bayes Oracle risk *exactly*, both when the true number of signals p is known and when it is unknown but is estimated with an appropriate empirical Bayes estimator.
(3) Inspired by the recent work of [25], we introduce two other data-adaptive methods for estimating the hyperparameter a in the NBP model based on restricted marginal maximum likelihood (REML) and hierarchical Bayes estimation. We study multiple testing procedures under these methods for a variety of shrinkage priors and show that they mimic oracle performance.

The organization of this paper is as follows. In Section 2, we introduce the normal-

beta prime (NBP) prior and establish new concentration inequalities for the beta prime density when it is employed as a scale parameter in Bayesian hierarchical models. In Section 3, we consider two different testing rules – one non-adaptive and one data-adaptive – based on thresholding the posterior shrinkage weight and illustrate that they both possess the Bayes Oracle property. In Section 4, we introduce a restricted marginal maximum likelihood approach and a hierarchical Bayes approach for estimating the sparsity parameter in the NBP prior. In Section 5, we present simulation results to validate our theoretical findings. Finally, in Section 6, we utilize the NBP prior to analyse a prostate cancer data set.

Proofs for the propositions and theorems in this article are available in the Supplementary Materials.

1.4. Notation

We use the following notations for the rest of the paper. Let $\{a_n\}$ and $\{b_n\}$ be two non-negative sequences of real numbers indexed by n, where $b_n \neq 0$ for sufficiently large n. If $\lim_{n\to\infty} a_n/b_n = 1$, we write $a_n \sim b_n$. If $|a_n/b_n| \leq M$ for all sufficiently large n where $M > 0$ is a positive constant independent of n, then we write $a_n = O(b_n)$. If $\lim_{n\to\infty} a_n/b_n = 0$, we write $a_n = o(b_n)$. Thus, $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$.

Throughout the paper, we also use Z to denote a standard normal $\mathcal{N}(0,1)$ random variable having cumulative distribution function and probability density function $\Phi(\cdot)$ and $\phi(\cdot)$, respectively.

2. The Normal-Beta Prime (NBP) Prior

Suppose we observe $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{I}_n)$, and our task is to perform signal detection on the n-dimensional vector, $\boldsymbol{\theta}$. Consider putting the normal-beta prime (NBP) prior on each $\theta_i, i = 1, \ldots, n$, as follows:

$$\begin{array}{c} \theta_i|\sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2), i = 1, \ldots, n, \\ \sigma_i^2 \sim \beta'(a, b), i = 1, \ldots, n, \end{array} \tag{6}$$

where $\beta'(a, b)$ denotes the beta prime density,

$$\pi(\sigma_i^2) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}(\sigma_i^2)^{a-1}(1+\sigma_i^2)^{-(a+b)}, i = 1, ..., n, \tag{7}$$

and $a > 0, b > 0$. We point out that [1] also considered the beta prime prior as a prior in a normal scale-mixture model. Specifically, [1] proposed the prior, $\theta_i \sim \mathcal{N}(0, \lambda_i\tau)$ with (7) as the prior for the local scale parameters, $\lambda_i \sim \pi(\lambda_i)$, and an additional variance rescaling parameter $\tau > 0$. They called their model the three parameter beta normal (TPBN) prior. Thus, the NBP model can be thought of as a special case of the TPBN prior with $\tau = 1$. Our work differs from [1] in that [1] recommended fixing the hyperparameters (a, b) *a priori* and controlling the sparsity of the model through the variance rescaling parameter τ. In contrast, we recommend fixing $\tau = 1$ and controlling the sparsity in our model through the hyperparameters (a, b).

Under the NBP model, the priors are *a priori* independent, so the posterior mean

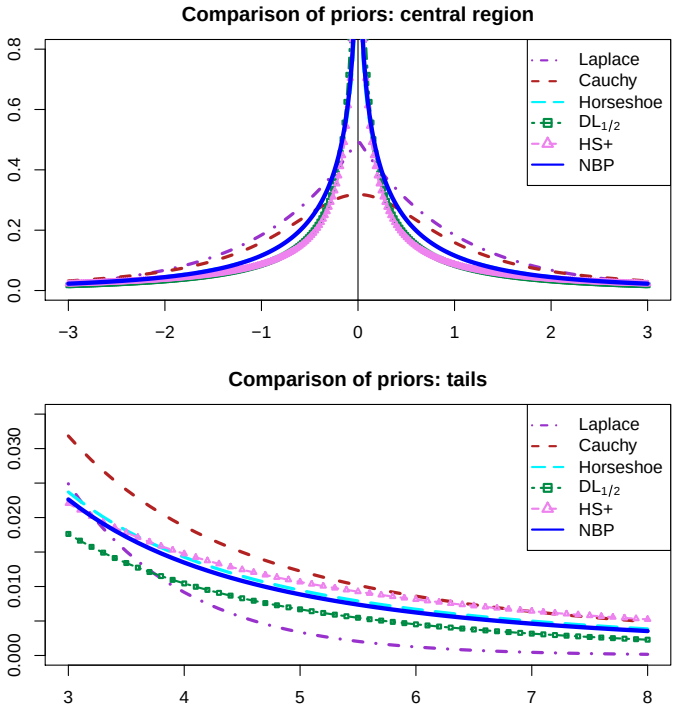

Figure 1. Marginal density of the NBP prior (6) with hyperparameters $a = 0.48, b = 0.52$, in comparison to other shrinkage priors. The HS+ prior is the marginal density of the horseshoe+, and the $\text{DL}_{1/2}$ prior is the marginal density for the Dirichlet-Laplace density with $\mathcal{D}(1/2, ..., 1/2)$ specified as a prior in the Bayesian hierarchy.

of θ_i under (6) is given by

$$\mathbb{E}\{\mathbb{E}(\theta_i|X_i, \sigma_i^2)\} = \{\mathbb{E}(1-\kappa_i)|X_i\}\, X_i, \tag{8}$$

where $\kappa_i = 1/(1+\sigma_i^2)$. Using a simple transformation of variables, we also see that the posterior density of the shrinkage factor κ_i is proportional to

$$\pi(\kappa_i|X_i) \propto \exp\left(-\frac{\kappa_i X_i^2}{2}\right) \kappa_i^{b-1/2}(1-\kappa_i)^{a-1}, \quad \kappa_i \in (0,1). \tag{9}$$

From (8) and (9), it is clear that the amount of posterior shrinkage is controlled by the shrinkage factor κ_i. For example, with $a = b = 0.5$, we obtain the standard half-Cauchy density $\mathcal{C}^+(0,1)$ for σ_i. As noted by [9] and [18], when $\mathcal{C}^+(0,1)$ is used as the prior for σ_i in (2), the marginal density for a single θ is unbounded at zero. In the next proposition, we show that for *any* choice of $a \in (0, 1/2]$, the marginal distribution for θ under the NBP prior also has a singularity at zero.

Proposition 2.1. *Let θ be an individual unknown population mean in (1). If θ is endowed with the NBP prior (6), then the marginal distribution of θ is unbounded with a singularity at zero for any $0 < a \leq 1/2$.*

Proposition 2.1 gives us some insight into how we should choose the hyperparameters in (6). Namely, we see that for small values of a, the NBP prior has a singularity at zero, similar to the horseshoe and the Dirichlet-Laplace [6] priors. Thus, small values of

a enable the NBP to obtain sparse estimates of the θ_i's by shrinking most observations to zero. As we will illustrate in Section 2.1, the tails of the NBP prior are still heavy enough to identify signals that are significantly far away from zero.

Figure 1 gives a plot of the marginal density $\pi(\theta)$ for the NBP prior (6), with $a = 0.48$ and $b = 0.52$. Figure 1 shows that with small values of a and b, the NBP has a singularity at zero, but it maintains the same tail robustness as other well-known shrinkage priors.

2.1. Concentration Properties of the NBP Prior

Consider the NBP prior given in (6), but suppose that we allow the hyperparameter $a \equiv a_n$ to vary with n as $n \to \infty$. Namely, we allow $0 < a_n < 1$ for all n, but $a_n \to 0$ as $n \to \infty$ so that even more mass is placed around zero as $n \to \infty$. We also fix b to lie in the interval $(1/2, \infty)$. To emphasize that the hyperparameter a_n depends on n, we rewrite the prior (6) as

$$\begin{aligned} \theta_i|\sigma_i^2 &\sim \mathcal{N}(0, \sigma_i^2), i = 1, \ldots, n, \\ \sigma_i^2 &\sim \beta'(a_n, b), i = 1, \ldots, n, \end{aligned} \tag{10}$$

where $a_n \in (0, 1)$ with $a_n = o(1)$ and $b \in (1/2, \infty)$. For the rest of the paper, we label this particular variant of the NBP prior as the NBP$_n$ prior.

As described in Section 2, the shrinkage factor $\kappa_i = 1/(1 + \sigma_i^2)$ plays a critical role in the amount of shrinkage of each observation X_i. In this section, we further characterize the tail properties of the posterior distribution $\pi(\kappa_i|X_i)$. Our theoretical results demonstrate that the NBP$_n$ prior (10) shrinks most estimates of θ_i's to zero but still has heavy enough tails to identify true signals. In the following results, we assume the NBP$_n$ prior on θ_i for $X_i \sim \mathcal{N}(\theta_i, 1)$.

Theorem 2.1. *For any $a_n, b \in (0, \infty)$,*

$$\mathbb{E}(1 - \kappa_i|X_i) \leq e^{X_i^2/2} \left(\frac{a_n}{a_n + b + 1/2} \right).$$

Corollary 2.1.1. *If $a_n \to 0$ as $n \to \infty$ and $b > 0$ is fixed, then $\mathbb{E}(1 - \kappa_i|X_i) \to 0$ as $n \to \infty$.*

Theorem 2.2. *Fix $\epsilon \in (0, 1)$. For any $a_n \in (0, 1)$, $b \in (1/2, \infty)$,*

$$\Pr(\kappa_i < \epsilon|X_i) \leq e^{X_i^2/2} \frac{a_n \epsilon}{(b + 1/2)(1 - \epsilon)}.$$

Corollary 2.2.1. *If $a_n \to 0$ as $n \to \infty$ and $b \in (1/2, \infty)$ is fixed, then by Theorem 2.2, $\Pr(\kappa_i \geq \epsilon|X_i) \to 1$ for any fixed $\epsilon \in (0, 1)$.*

Theorem 2.3. *Fix $\eta \in (0, 1), \delta \in (0, 1)$. Then for any $a_n \in (0, 1)$ and $b \in (1/2, \infty)$,*

$$\Pr(\kappa_i > \eta|X_i) \leq \frac{\left(b + \frac{1}{2}\right)(1 - \eta)^{a_n}}{a_n(\eta\delta)^{b+\frac{1}{2}}} \exp\left(-\frac{\eta(1 - \delta)}{2} X_i^2 \right).$$

Corollary 2.3.1. *For any fixed n where $a_n \in (0, 1), b \in (1/2, \infty)$, and for every fixed $\eta \in (0, 1)$, $\Pr(\kappa_i \leq \eta|X_i) \to 1$ as $X_i \to \infty$.*

Corollary 2.3.2. *For any fixed n where $a_n \in (0,1), b \in (1/2, \infty)$, and for every fixed $\eta \in (0,1)$, $\mathbb{E}(1-\kappa_i|X_i) \to 1$ as $X_i \to \infty$.*

Since $\mathbb{E}(\theta_i|X_i) = \{\mathbb{E}(1-\kappa_i)|X_i\}\, X_i$, Corollaries 2.1.1 and 2.2.1 illustrate that all observations will be shrunk towards the origin under the NBP$_n$ prior (10). However, Corollaries 2.3.1 and 2.3.2 demonstrate that if X_i is big enough, then the posterior mean $\{\mathbb{E}(1-\kappa_i)|X_i\}\, X_i \approx X_i$. This ensures the tails of the NBP prior are still sufficiently heavy to detect true signals.

A referee has pointed out that the conditions on the hyperparameter a_n in Corollaries 2.1.1-2.3.2 closely resemble conditions on the rescaling (or the 'global') parameter $\tau \equiv \tau_n$ in priors of the form (5) in the literature. Indeed, if $\tau_n \in (0,1)$ and $\tau_n \to 0$ in (5), then one obtains analogous results for priors of the form (5). See, e.g. [11,14]. This is because, as seen in Proposition 2.1, the hyperparameter a_n controls the amount of mass around zero for the NBP (with smaller values leading to heavier mass in the neighborhood of zero). At the same time, keeping b fixed in the range $(1/2, \infty)$ ensures that the NBP has heavy enough tails to prevent overshrinkage of large signals. Thus, the hyperparameter b also plays a similar role as the 'local' parameter λ_i in (5).

3. Multiple Testing with the NBP Prior

3.1. Asymptotic Bayes Optimality Under Sparsity

Suppose we observe $\boldsymbol{X} = (X_1, \ldots, X_n)$, such that $X_i \sim \mathcal{N}(\theta_i, 1)$, for $i = 1, \ldots, n$. To identify the true signals in $\boldsymbol{X}$, we conduct n simultaneous tests: $H_{0i} : \theta_i = 0$ against $H_{1i} : \theta_i \neq 0$, for $i = 1, \ldots, n$. For each i, θ_i is assumed to come from the model,

$$\theta_i \overset{i.i.d.}{\sim} (1-p)\delta_{\{0\}} + p\mathcal{N}(0, \psi^2), i = 1, \ldots, n, \tag{11}$$

where $\psi^2 > 0$ represents a diffuse 'slab' density. This point mass mixture model is often considered to be a data generating mechanism for sparse vectors $\boldsymbol{\theta}$ in the statistical literature. [8] referred to model (11) as a 'gold standard' for sparse problems.

Model (11) is equivalent to assuming that for each i, θ_i follows a random variable whose distribution is determined by the latent binary random variable ν_i, where $\nu_i = 0$ denotes the event that H_{0i} is true, while $\nu_i = 1$ corresponds to the event that H_{0i} is false. Here ν_i's are assumed to be i.i.d. Bernoulli(p) random variables, for some p in $(0,1)$. Under H_{0i}, i.e. $\theta_i \sim \delta_{\{0\}}$, the distribution having a mass 1 at 0, while under H_{1i}, $\theta_i \neq 0$ and is assumed to follow an $\mathcal{N}(0, \psi^2)$ distribution with $\psi^2 > 0$. The marginal distributions of the X_i's are then given by the following two-groups model:

$$X_i \overset{i.i.d.}{\sim} (1-p)\mathcal{N}(0,1) + p\mathcal{N}(0, 1+\psi^2), i = 1, \ldots, n. \tag{12}$$

Our testing problem is now equivalent to testing simultaneously

$$H_{0i} : \nu_i = 0 \text{ versus } H_{1i} : \nu_i = 1 \text{ for } i = 1, \ldots, n. \tag{13}$$

We consider a symmetric 0-1 loss for each individual test. The total loss of a multiple testing procedure is assumed to be the sum of the individual losses incurred in each test. Letting t_{1i} and t_{2i} denote the probabilities of type I and type II errors of the ith test respectively, the Bayes risk of a multiple testing procedure under the two-groups

model (12) is then given by

$$R = \sum_{i=1}^{n} \{(1-p)t_{1i} + pt_{2i}\}. \tag{14}$$

[7] showed that the rule which minimizes the Bayes risk in (14) is the test which, for each $i = 1, \ldots, n$, rejects H_{0i} if

$$\frac{f(x_i|\nu_i = 1)}{f(x_i|\nu_i = 0)} > \frac{1-p}{p}, \text{i.e. } X_i^2 > c^2, \tag{15}$$

where $f(x_i|\nu_i = 1)$ denotes the marginal density of X_i under H_{1i}, while $f(x_i|\nu_i = 0)$ denotes that under H_{0i} and

$$c^2 \equiv c^2_{\psi,f} = \frac{1+\psi^2}{\psi^2}(\log(1+\psi^2) + 2\log(f)),$$

with $f = (1-p)/p$. The above rule is known as the Bayes Oracle, because it makes use of unknown parameters ψ and p. By reparametrizing as $u = \psi^2$ and $v = uf^2$, the above threshold becomes

$$c^2 \equiv c^2_{u,v} = \left(1 + \frac{1}{u}\right)\left(\log v + \log\left(1 + \frac{1}{u}\right)\right).$$

[7] considered the following asymptotic scheme.

Assumption 1

The sequences of vectors (ψ_n, p_n) satisfies the following conditions:

(1) $p_n \to 0$ as $n \to \infty$.
(2) $u_n = \psi_n^2 \to \infty$ as $n \to \infty$.
(3) $v_n = u_n f^2 \to \infty$ as $n \to \infty$.
(4) $\frac{\log v_n}{u_n} \to C \in (0, \infty)$ as $n \to \infty$.

The first condition in Assumption 1 assumes that the underlying $\boldsymbol{\theta}$ becomes more sparse as n approaches infinity, while the second condition ensures that true signals can still be identified. [7] provided detailed insight on the threshold C arising from the third and fourth conditions. Summarizing briefly, if $C = 0$, then the probability of a Type I error is one and the probability of a Type II error is zero. If $C = \infty$, then the probability of a Type I error is zero and the probability of a Type II error is one. Under Assumption 1, [7] showed that the corresponding asymptotic Bayes Oracle risk has a particularly simple form, which is given by

$$R^{BO}_{Opt} = n((1-p)t_1^{BO} + pt_2^{BO}) = np(2\Phi(\sqrt{C}) - 1)(1 + o(1)), \tag{16}$$

where the $o(1)$ terms tend to zero as $n \to \infty$. A testing procedure with risk R is said to be asymptotically Bayes optimal under sparsity (ABOS) if

$$\frac{R}{R^{BO}_{Opt}} \to 1 \text{ as } n \to \infty. \tag{17}$$

Remark 3.1. *[7] consider the more general case where under the null hypothesis,* $H_{0i} : \theta_i \sim \mathcal{N}(0, \zeta^2)$ *with* $0 \leq \zeta \ll \psi$. *That is, [7] assumed that the true data-generating mechanism for* $\boldsymbol{\theta}$ *is given by*

$$\theta_i \overset{i.i.d.}{\sim} (1-p)\mathcal{N}(0, \zeta^2) + p\mathcal{N}(0, \psi^2), i = 1, \ldots, n. \tag{18}$$

The point mass mixture (11) is obtained as a special case of (18) by setting $\zeta = 0$. *[7] showed that the asymptotic Bayes Oracle risk under (18) is the same as (16) if we replace* $u = \psi^2$ *in Assumption 1 with* $u = (\psi/(\zeta+1))^2$. *If we assume that the true* $\boldsymbol{\theta}$ *comes from (18) with* $\zeta > 0$ *and we similarly set* $u = (\psi/(\zeta+1))^2$ *in Assumption 1, then all the results in this manuscript will continue to hold. Thus, when* $\boldsymbol{\theta}$ *is 'nearly' (but not exactly) sparse, thresholding rule (19) for classifying signals under the NBP prior is also ABOS.*

In Sections 3.3 and 3.4, we consider two thresholding rules based on the NBP model. In the first case, we assume the sparsity level p under the true data-generating model (11) to be known. For the more realistic scenario where p is unknown, we base our test procedure on a data-driven estimate of p. Since we estimate the unknown proportion of signals from the data, we refer to this latter procedure as a *data-adaptive* testing rule.

3.2. Related Work for Scale-Mixture Shrinkage Priors

We briefly survey related work on multiple testing under normal scale-mixture shrinkage priors (2) to demonstrate the novelty of our results. [13] showed that for priors of the form (5), thresholding rule (19) is ABOS provided that: a) the variance rescaling parameter $\tau > 0$ decays at an appropriate rate or is estimated by an appropriate plug-in estimator, and b) the slowly varying component $L(\cdot)$ in the scale prior, $\pi(\lambda_i) \propto \lambda_i^{-a-1} L(\lambda_i)$, is uniformly bounded above and below on the interval $\lambda_i \in (0, \infty)$. The NBP prior (6) does not require a rescaling parameter $\tau > 0$ in the normal variance in the first level of the Bayes hierarchy. Thus, our results cannot be obtained from those in [13].

Under certain conditions on the prior for the scale parameter σ_i^2 in (2), [19] derived asymptotic upper bounds on Type I and Type II errors and the Bayes risk (14) for both non-adaptive and data-adaptive test procedures induced by scale-mixture shrinkage priors. Under these conditions, the upper bound on the Bayes risk for scale-mixture priors is of the same order as the Bayes Oracle risk. Specifically, [19] showed that the Bayes risk (14) for thresholding rule (4) can be bounded from above by $np[16\sqrt{\pi}C/c + 2\Phi(\sqrt{2K(u_0+1)C}) - 1](1+o(1))$ for known p and by $np[16\sqrt{\pi}CD/c + 2\Phi(\sqrt{2K(u_0+1)(1+\xi)C}) - 1](1+o(1))$ for unknown p, where C is the constant from the fourth condition in Assumption 1 and $c > 0, K \geq 0, u_0 > 0, D > 0, \xi \geq 0$ are appropriate constants that depend on the prior.

One can show that with appropriate conditions on the hyperparameters (a, b), the NBP prior (6) satisfies the conditions in [19]. Therefore, our prior can also obtain the upper bound on the risk derived by [19]. However, the results that we present in this paper do not immediately follow from [19] because: a) we provide *lower* bounds on Type I and Type II errors under our prior, and b) we establish that the Bayes risk under the NBP prior is actually asymptotically the same as that of the Bayes Oracle risk given in (16). Therefore, our bounds are provably sharper than those of [19].

3.3. A Non-Adaptive Testing Rule Under the NBP Prior

As noted earlier, the posterior mean under the NBP prior depends heavily on the shrinkage factor, $\kappa_i = 1/(1+\sigma_i^2)$. Because of the concentration properties of the NBP prior proven in Section 2.1, a sensible thresholding rule classifies observations as signals or as noise based on the posterior distribution of this shrinkage factor. Consider the following testing rule for the ith observation X_i:

$$\text{Reject } H_{0i} \text{ if } \mathbb{E}(1-\kappa_i|X_i) > \frac{1}{2}, \tag{19}$$

where κ_i is the shrinkage factor based on the NBP$_n$ prior (10). Within the context of multiple testing, a good benchmark for our test procedure (19) should be whether it is ABOS, i.e. whether its optimal risk is asymptotically equal to that of the Bayes Oracle risk (16). Adopting the framework of [7], we let R_{NBP} denote the asymptotic Bayes risk of testing rule (19). In the next four theorems, we derive sharp lower and upper bounds on the Type I and Type II error probabilities for test procedure (19). These error probabilities are given by

$$\begin{aligned} t_{1i} &= \Pr\left[\mathbb{E}(1-\kappa_i|X_i) > \tfrac{1}{2} \middle| H_{0i} \text{ is true}\right], \\ t_{2i} &= \Pr\left[\mathbb{E}(1-\kappa_i|X_i) \leq \tfrac{1}{2} \middle| H_{1i} \text{ is true}\right]. \end{aligned} \tag{20}$$

Theorem 3.1. *Suppose that $X_1, \ldots, X_n$ are i.i.d. observations having distribution (12) where the sequence of vectors (ψ_n^2, p_n) satisfies Assumption 1. Suppose we wish to test (13) using the classification rule (19) under the NBP$_n$ prior. Then for all n, an upper bound for the probability of a Type I error for the ith test is given by*

$$t_{1i} \leq \frac{2\sqrt{2}a_n}{\sqrt{\pi}(a_n+b+1/2)}\left[\log\left(\frac{a_n+b+1/2}{2a_n}\right)\right]^{-1/2}.$$

Theorem 3.2. *Assume the same setup of Theorem 3.1. Suppose we wish to test (13) using the classification rule (19) under the NBP$_n$ prior. Suppose further that $a_n \in (0,1)$, with $a_n \to 0$ as $n \to \infty$ and $b \in (1/2, \infty)$ is fixed. Then for any $\xi \in (0, 1/2)$ and $\delta \in (0,1)$, a lower bound for the probability of a Type I error for the ith test as $n \to \infty$ is given by*

$$t_{1i} \geq 1 - \Phi\left(\sqrt{\frac{2}{\xi(1-\delta)}\left[\log\left(\frac{\left(b+\frac{1}{2}\right)(1-\xi)^{a_n}}{a_n(\xi\delta)^{b+\frac{1}{2}}}\right)\right]}\right).$$

Theorem 3.3. *Assume the same setup as Theorem 3.1. Suppose we wish to test (13) using the classification rule (19) under the NBP$_n$ prior. Suppose further that $a_n \in (0,1)$, with $a_n \to 0$ as $n \to \infty$ in such a way that $\lim_{n\to\infty} a_n/p_n \in (0,\infty)$ and that $b \in (1/2,\infty)$ is fixed. Fix $\eta \in (0,1), \delta \in (0,1)$, and choose any $\rho > 2/(\eta(1-\delta))$. Then as $n \to \infty$, an upper bound for the probability of a Type II error for the ith test*

is given by

$$t_{2i} \leq \left[2\Phi\left(\sqrt{\frac{\rho C}{2}} \right) - 1 \right] (1 + o(1)) \text{ as } n \to \infty,$$

where the $o(1)$ *terms above go to 0 as* $n \to \infty$.

Theorem 3.4. *Assume the same setup as Theorem 3.1. Suppose we wish to test (13) using the classification rule (19) under the NBP*$_n$ *prior. Suppose further that* $a_n \in (0, 1)$*, with* $a_n \to 0$ *as* $n \to \infty$ *in such as way that* $\lim_{n\to\infty} a_n/p_n \in (0, \infty)$ *and that* $b \in (1/2, \infty)$ *is fixed. Then as* $n \to \infty$*, a lower bound for the probability of a Type II error for the* i*th test is given by*

$$t_{2i} \geq \left[2\Phi(\sqrt{C}) - 1 \right] (1 + o(1)) \text{ as } n \to \infty,$$

where the $o(1)$ *terms tend to zero as* $n \to \infty$.

Theorems 3.1-3.2 show that for *any* sequence of hyperparameters a_n such that $a_n \to 0$ as $n \to \infty$, the probability of a Type I error for test (19) is asymptotically vanishing under the NBP$_n$ prior. Meanwhile, Theorems 3.3-3.4 show that if a_n is the same order as the true signal proportion p_n, then the probability of Type II error for test (19) can be bounded from above and below. Notice that in Theorem 3.3, we are free to choose any ρ arbitrarily close to 2 in the upper bound on the probability of a Type II error, with 2 being the infimum for ρ. Thus, the limit inferior of upper bound in Theorem 3.3 is the same as the lower bound established in Theorem 3.4, and so these bounds are sharp. Altogether, Theorems 3.1-3.4 show that asymptotically, the Bayes risk (16) is controlled entirely by the Type II error. If $C \approx 0$ in Assumption 1, then the power of the ith test, $1 - t_{2i}$, under the NBP$_n$ prior (10) will be close to one.

Having obtained appropriate upper and lower bounds on the Type I and Type II probabilities under thresholding rule (19), we are ready to state our main theorem which proves that our method under the NBP$_n$ prior is asymptotically Bayes optimal under sparsity.

Theorem 3.5. *Suppose that* $X_1, \ldots, X_n$ *are i.i.d. observations having distribution (12) where the sequence of vectors* (ψ^2, p) *satisfies Assumption 1. Suppose we wish to test (13) using the classification rule (19). Suppose further that* $a_n \in (0, 1)$*, with* $a_n \to 0$ *as* $n \to \infty$ *in such a way that* $\lim_{n\to\infty} a_n/p_n \in (0, \infty)$ *and that* $b \in (1/2, \infty)$ *is fixed. Then*

$$\lim_{n\to\infty} \frac{R_{NBP}}{R^{BO}_{Opt}} = 1, \tag{21}$$

i.e. rule (19) based on the NBP$_n$ *prior (10) is ABOS.*

We have shown that our thresholding rule based on the NBP$_n$ prior asymptotically attains the Bayes Oracle risk exactly, provided that a_n is of the same order as the sparsity level p_n. Since p_n is typically unknown, it should ideally be estimated from the data, and our theoretical findings suggest how to build adaptive procedures for setting a_n, which we describe in Sections 3.4 and 4.

3.4. A Data-Adaptive Testing Rule Under the NBP Prior

As we found in Theorem 3.5, our test procedure (19) has the Bayes oracle property under the NBP$_n$ prior, provided that a_n is of the same order as the true signal proportion p_n and $b \in (1/2, \infty)$ is fixed. However, p_n is typically unknown, and as a result, we must estimate it from the data. To this end, we use the estimator proposed by [24]:

$$\widehat{a}_n^{ES} := \max\left\{\frac{1}{n}, \frac{1}{c_2 n}\sum_{j=1}^{n} 1\{|X_j| > \sqrt{c_1 \log n}\}\right\}, \tag{22}$$

where $c_1 \geq 2$ and $c_2 \geq 1$ are fixed constants, and we use ES to denote 'estimated sparsity.' This choice is motivated by the so-called 'universal threshold,' $\sqrt{2\log n}$. It is well-known that signals which fall below this threshold are shrunk towards zero, and thus, they may not be detected.

Based on the considerations above, we now introduce a data-adaptive testing rule under the NBP$_n$ prior. Setting $a_n \equiv \widehat{a}_n^{ES}$ and $b \in (1/2, \infty)$ as the hyperparameters in the NBP prior, our test for the ith observation X_i is:

$$\text{Reject } H_{0i} \text{ if } \mathbb{E}(1 - \kappa_i | X_i, \widehat{a}_n^{ES}) > \frac{1}{2}, \tag{23}$$

From a decision theoretic perspective, we now demonstrate that setting the hyperparameter a_n equal to $\widehat{a}_n^{ES}$ is also justified. Letting R_{NBP}^{ES} denote the asymptotic Bayes risk, we first derive sharp lower bounds and upper bounds on the Type I and Type II error probabilities, which we denote as $\widetilde{t}_{1i}$ and $\widetilde{t}_{2i}$ respectively. We then illustrate that testing rule (23) is *also* ABOS.

Following the notation of [14], we denote

$$\alpha_n = \Pr(|X_i| > \sqrt{c_1 \log n}), \text{ and } \beta = 1 - \Phi(c_1 C/2\epsilon), \tag{24}$$

where $\epsilon \in (0,1)$, c_1 is the constant in (22), and C is the constant from Assumption 1. In [14], it was shown that as long as the signal proportion $p_n \propto n^{-\epsilon}$ and Assumption 1 holds, then

$$\alpha_n = 2\beta p_n (1 + o(1)), \tag{25}$$

under the two-groups model (12), where the $o(1)$ terms go to 0 as $n \to \infty$. We will use (24) and (25) to prove Theorems 3.6 and 3.7, which provide asymptotic bounds on the Type I and Type II error probabilities under (23).

Theorem 3.6. *Suppose that $X_1, \ldots, X_n$ are i.i.d. observations having distribution (12) where the sequence of vectors (ψ_n^2, p_n) satisfies Assumption 1, with $p_n \propto n^{-\epsilon}, \epsilon \in (0,1)$. Fix $b \in (1/2, \infty)$, $c_1 \geq 2$, $c_2 \geq 1$, $\xi \in (0, 1/2)$, and $\delta \in (0,1)$. Suppose we wish to test (13) using the classification rule (23). Then as $n \to \infty$, bounds for the probability of a Type I error for the ith test, $\widetilde{t}_{1i}$, are given by*

$$1-\Phi\left(\sqrt{\frac{2}{\xi(1-\delta)}\left[\log\left(\frac{\left(b+\frac{1}{2}\right)(1-\xi)^{2\alpha_n}}{2\alpha_n(\xi\delta)^{b+\frac{1}{2}}}\right)\right]}\right) \leq \widetilde{t}_{1i}$$
$$\leq \frac{4\alpha_n}{\sqrt{\pi}(2\alpha_n+b+1/2)}\left[\log\left(\frac{2\alpha_n+b+1/2}{4\alpha_n}\right)\right]^{-1/2}(1+o(1))$$
$$+\frac{1/\sqrt{\pi}}{n^{c_1/2}\sqrt{\log n}}+e^{-2(2\log 2-1)\beta n p_n(1+o(1))},$$

where α_n and β are as in (24).

Theorem 3.7. *Assume the same setup as Theorem 3.6, and assume that $p_n \propto n^{-\epsilon}, \epsilon \in (0,1)$. Fix $b \in (1/2,\infty)$, $c_1 \geq 2$, $c_2 \geq 1$, $\eta \in (0,1)$, and $\delta \in (0,1)$, and choose any $\rho > 2/(\eta(1-\delta))$. Suppose we wish to test (13) using the classification rule (23). Then as $n \to \infty$, bounds for the probability of a Type II error for the ith test, $\widetilde{t}_{1i}$, are given by*

$$\left[2\Phi(\sqrt{C})-1\right](1+o(1)) \leq \widetilde{t}_{2i} \leq \left[2\Phi\left(\sqrt{\frac{\rho C}{2}}\right)-1\right](1+o(1)) \text{ as } n\to\infty,$$

where the $o(1)$ terms tend to zero as $n \to \infty$.

We pause briefly to compare Theorems 3.6-3.7 with Theorems 3.1-3.4. Theorems 3.1-3.2 demonstrated that for *any* sequence of hyperparameters a_n such that $a_n \to 0$ as $n \to \infty$, the probability of a Type I error under thresholding rule (19) asymptotically vanishes. Theorem 3.6 shows that this will also be the case for plug-in estimator $\widehat{a}_n^{EB}$ as long as $\alpha_n := \Pr(|X_i| > \sqrt{c_1 \log n})$, goes to 0 as $n \to \infty$. This condition holds for any $p_n \propto n^{-\epsilon}, \epsilon \in (0,1)$. In replacing the generic sequence a_n with a specific plug-in value $\widehat{a}_n^{EB}$, the bounds in Theorem 3.6 differ in constants from the bounds derived in Theorem 3.1. However, the bounds in Theorems 3.1 and 3.6 are ultimately of the same order if $a_n \to 0$ and $\alpha_n \to 0$. In addition, Theorem 3.7 shows that if $p_n \propto n^{-\epsilon}$ and we utilize the EB estimator $\widehat{a}_n^{ES}$ (22) in place of a_n, then the upper and lower bounds on probability of Type II error are the same as those in Theorems 3.3-3.4.

Having obtained appropriate upper and lower bounds on the Type I and Type II probabilities under thresholding rule (23), we are ready to state our main theorem which proves that our data-adaptive testing procedure (23) also possesses the Bayes Oracle property in the entire range of sparsity parameters $p \propto n^{-\epsilon}, \epsilon \in (0,1)$.

Theorem 3.8. *Suppose that $X_1,\ldots,X_n$ are i.i.d. observations having distribution (12) where the sequence of vectors (ψ^2, p) satisfies Assumption 1. Further assume that $p \propto n^{-\epsilon}, \epsilon \in (0,1)$. For the NBP prior (6), fix $b \in (1/2,\infty)$ and set $a = \widehat{a}_n^{ES}$, where $\widehat{a}_n^{ES}$ is as in (22), with fixed (c_1, c_2) satisfying $c_1 \geq 2$ and $c_2 \geq 1$. Suppose that we wish to test (13) using the classification rule (23). Then*

$$\lim_{n\to\infty} \frac{R_{NBP}^{ES}}{R_{Opt}^{BO}} = 1, \tag{26}$$

i.e. data-adaptive thresholding rule (23) is ABOS.

Proof. This follows the same reasoning as the proof for Theorem 3.5, except we replace the bounds for t_{1i} and t_{2i} with those of $\widetilde{t}_{1i}$ and $\widetilde{t}_{2i}$ from Theorems 3.6 and 3.7. To prove that the bounds for $\widetilde{t}_{1i}$ in Theorem 3.6 tend to zero, note that $p_n \propto n^{-\epsilon}, \epsilon \in (0,1)$, and therefore, by (25), $\alpha_n \to 0$ as $n \to \infty$. □

Note that this condition on p is quite mild. For comparison, [7] showed that the widely used Benjamini-Hochberg (BH) [3] procedure for controlling the false discovery rate (FDR) is ABOS if and only if $p_n \propto n^{-\epsilon}, \epsilon \in (0,1]$. Unlike the BH procedure, however, the NBP requires an estimate of the unknown sparsity level p_n in order to achieve the Bayes Oracle property and is *not* ABOS if $p_n = n^{-1}$ (in this case, the probability of a Type I error is not asymptotically vanishing). Nevertheless, there are several advantages of the NBP model over BH. The BH procedure cannot be used for estimation or uncertainty quantification of $\boldsymbol{\theta}$. In contrast, the NBP model not only admits a testing procedure that is ABOS, but the NBP posterior can also be used to obtain estimates and credible intervals for $\boldsymbol{\theta}$. Further, obtaining an estimate of unknown sparsity p_n, such as the one in (22) or the ones described in Section 4, is not computationally expensive; this adds only a single preprocessing step or an extra iteration in the Markov chain Monte Carlo (MCMC) algorithm. The assumption that the true sparsity level satisfies $p \propto n^{-\epsilon}, \epsilon \in (0,1)$ (i.e. that there is more than one signal in the data) is also very likely to be satisfied in practice.

4. Two Other Data-Adaptive Approaches for Estimating the Sparsity Parameter

As we demonstrated in Section 3.3 and 3.4, we can construct hypothesis tests based on the NBP prior which have the Bayes Oracle property by fixing $b \in (1/2, \infty)$ and by choosing a to be comparable to the proportion of true signals. By Proposition 2.1, a also controls the amount of mass around zero. Thus, a can be interpreted as the sparsity parameter, and the ideal choice of a should lie in the range $[1/n, 1]$.

In [25], the variance rescaling parameter τ in the horseshoe prior is estimated through restricted marginal maximum likelihood (REML) on the interval $[1/n, 1]$ or by placing a prior on τ with its support truncated to lie in the interval $[1/n, 1]$. The methods in [25] enable the horseshoe to achieve near-minimax posterior contraction.

4.1. A Restricted Marginal Maximum Likelihood (REML) Approach

Inspired by [25]'s work, we first propose a REML approach to estimating a. We take our estimate $\widehat{a}_n^{REML}$ to be the marginal maximum likelihood estimate of a restricted on the interval $[1/n, 1]$. That is, for a fixed b, we define $\widehat{a}_n^{REML}$ as

$$\widehat{a}_n^{REML} = \underset{a \in [1/n,1]}{\arg\max} \prod_{i=1}^{n} m(X_i), \tag{27}$$

where $m(X_i)$ denotes the marginal density for a single observation X_i, i.e.,

$$m(X_i) = \int_{-\infty}^{\infty} \int_{0}^{\infty} \phi(X_i - \theta_i)\phi(\theta_i/\sigma_i)\pi(\sigma_i^2) d\sigma_i^2 d\theta_i, \tag{28}$$

and $\pi(\sigma_i^2)$ is the prior for beta prime density given in (7). A closed form solution to (27) is unavailable, but it can be computed using numerical integration and optimization.

We now introduce yet another data-adaptive testing rule under the NBP prior. Suppose that we set $(\widehat{a}_n^{REML}, b)$ as our hyperparameters in the NBP prior (6), where $b \in (1/2, \infty)$. Then our test for the ith observation X_i is:

$$\text{Reject } H_{0i} \text{ if } \mathbb{E}(1 - \kappa_i | X_i, \widehat{a}_n^{REML}) > \frac{1}{2}. \tag{29}$$

4.2. A Hierarchical Bayes Approach

Our results also suggest that if we adopt a fully Bayes approach for estimating the sparsity parameter a, the prior on a should have its support truncated to $[1/n, 1]$. Suppose that we fix $b \in (1/2, \infty)$. Our hierarchical model is defined as

$$\begin{aligned} \theta_i | \sigma_i^2 &\sim \mathcal{N}(0, \sigma_i^2), i = 1, \ldots, n, \\ \sigma_i^2 &\sim \beta'(a, b), i = 1, \ldots, n, \\ a &\sim \pi(a), \end{aligned} \tag{30}$$

where the support of $\pi(a)$ is $[1/n, 1]$. Under (30), our thresholding rule now becomes

$$\text{Reject } H_{0i} \text{ if } \mathbb{E}(1 - \kappa_i | X_1, \ldots, X_n) > \frac{1}{2}. \tag{31}$$

Note that because we have placed a prior on a, the priors for the θ_i's are no longer *a priori* independent. Thus, the posterior densities of the θ_i's (and hence the κ_i's) also depend on *all* the data. For our simulation studies, we consider both a uniform prior for a, i.e. $a \sim \mathcal{U}(1/n, 1)$, and a standard Cauchy prior for a truncated to $[1/n, 1]$, i.e. $\pi(a) = [\arctan(1) - \arctan(1/n)]^{-1}(1 + a)^{-1}\mathbb{I}\{1/n < a < 1\}$.

In Section 5, we demonstrate that test procedures (29) and (31) both mimic the Bayes Oracle performance in simulations. We hope to provide theoretical justification for (29) and (31) in the future. Following the work of [25] for the horseshoe prior, we believe that useful bounds on (28) and on the posterior $\pi(a|X_1, \ldots, X_n)$ under (30) can be derived to facilitate theoretical analysis of the NBP prior when a is estimated by REML or by a truncated prior.

5. Simulation Studies

5.1. Implementation and Selection of the Hyperparameter b

In the case where a is fixed *a priori* or estimated with a plug-in estimator, the NBP model (6) can be implemented straightforwardly using Gibbs sampling. If a prior is placed on the hyperparameter a, as in (30), then we use Metropolis-Hastings to update a. In the Supplementary Materials, we provide the full details on how to sample from models (6) and (30). For the hierarchical Bayes approach, we saw that the MCMC chains mixed well and converged very quickly (in less than 100 iterations), even if we initialized the values to be far away from the truth. This is also illustrated in the Supplementary Materials. We provide the implementation of the NBP model and the multiple testing procedures (23), (29), and (31) in a comprehensive R package, `NormalBetaPrime`.

In order to use the NBP prior (6) for multiple testing, we recommend setting b to lie in the interval $(1/2, 1/2+\delta]$, for some small $\delta > 0$, and estimating a from the data. We could also estimate b from the data, but our theoretical results in Theorems 3.5 and Theorem 3.8 demonstrate that asymptotically, the specific choice of b plays no role. As pointed out by [18], smaller values of b correspond to heavier tails, with values of b close to $1/2$ giving Cauchy-like tails. Based on these considerations, we suggest the default choice of $b = 1/2 + 1/n$, so that the theoretical results established earlier hold, while the tails are still quite heavy.

5.2. Simulation Study

We adopt the simulation framework of [11] and [14] and fix sparsity levels at $p \in \{0.01, 0.05, 0.10, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$, for a total of 11 simulation settings. For sample size $n = 500$ and each p, we generate data from the two-groups model (12), with $\psi = \sqrt{2 \log n} = 3.53$. We fix $b = 1/2 + 1/n$ and implement the NBP model with each of the following estimates for a:

(1) NBP-ES: the estimated sparsity (ES) estimator $\widehat{a}^{ES}$, as in (22), with fixed constants $c_1 = 2, c_2 = 1$.
(2) NBP-REML: the REML estimator $\widehat{a}^{REML}$, as in (27),
(3) NBP-UNIF: a uniform prior on a, i.e. $a \sim \mathcal{U}(1/n, 1)$ in (30),
(4) NBP-TC: a truncated standard Cauchy prior on a, i.e. $\pi(a) = [\arctan(1) - \arctan(1/n)]^{-1}(1+a)^{-1}\mathbb{I}\{1/n < a < 1\}$ in (30). For shorthand notation, we denote this prior as $a \sim \mathcal{TC}(0, 1; 1/n, 1)$.

For each of these models, we apply the appropriate thresholding rule: (23) for NBP-ES, (29) for NBP-REML, and (31) for NBP-UNIF and NBP-TC to classify θ_i's in our model as either signals ($\theta_i \neq 0$) or noise ($\theta_i = 0$). We estimate the average misclassification probability (MP) for these thresholding rules from 100 replicates.

We compare the performance of our testing procedures to those under the horseshoe (HS), the horseshoe+ (HS+), and the Dirichlet-Laplace (DL) priors. In the HS and HS+ models, the sparsity parameter τ is the variance rescaling parameter in (5), while in the DL model, the sparsity parameter τ is the hyperparameter in the Dirichlet prior, $\mathcal{D}(\tau, \ldots, \tau)$. For each of these models, we estimate τ using either the ES estimator (22), $\widehat{\tau}^{ES}$, the REML estimator (27), $\widehat{\tau}^{REML}$, or by placing priors on τ, $\tau \sim \mathcal{U}(1/n, 1)$ or $\tau \sim \mathcal{TC}(0, 1; 1/n, 1)$. Implementation for the HS prior is available in the R package `horseshoe` [1], while the methods for the HS+ and DL priors are available in our package `NormalBetaPrime`.

Figure 2 plots the estimated misclassification probabilities (MP) against the true sparsity level p for each of the models, along with the MP's for the Bayes Oracle (BO) and the Benjamini-Hochberg procedure (BH). Recall that the Bayes Oracle rule, defined in (15), is the decision rule that minimizes the expected number of misclassified signals (14) when (p, ψ) are known. The Bayes Oracle therefore serves as the lower bound to the MP. For the Benjamini-Hochberg rule, we use $\alpha = 1/\log n = 0.1887$. [7] theoretically established for this choice of α, the BH procedure is ABOS.

Figure 2 illustrates that all the different models perform very similarly to the Bayes Oracle in sparse situations (p in the range of 0.01 to 0.30), regardless of whether the sparsity parameter is estimated by empirical Bayes or by hierarchical Bayes. Our

[1] For the method $\tau \sim \mathcal{U}(1/n, 1)$, we slightly modify the code in the `HS.normal.means` function in the `horseshoe` R package.

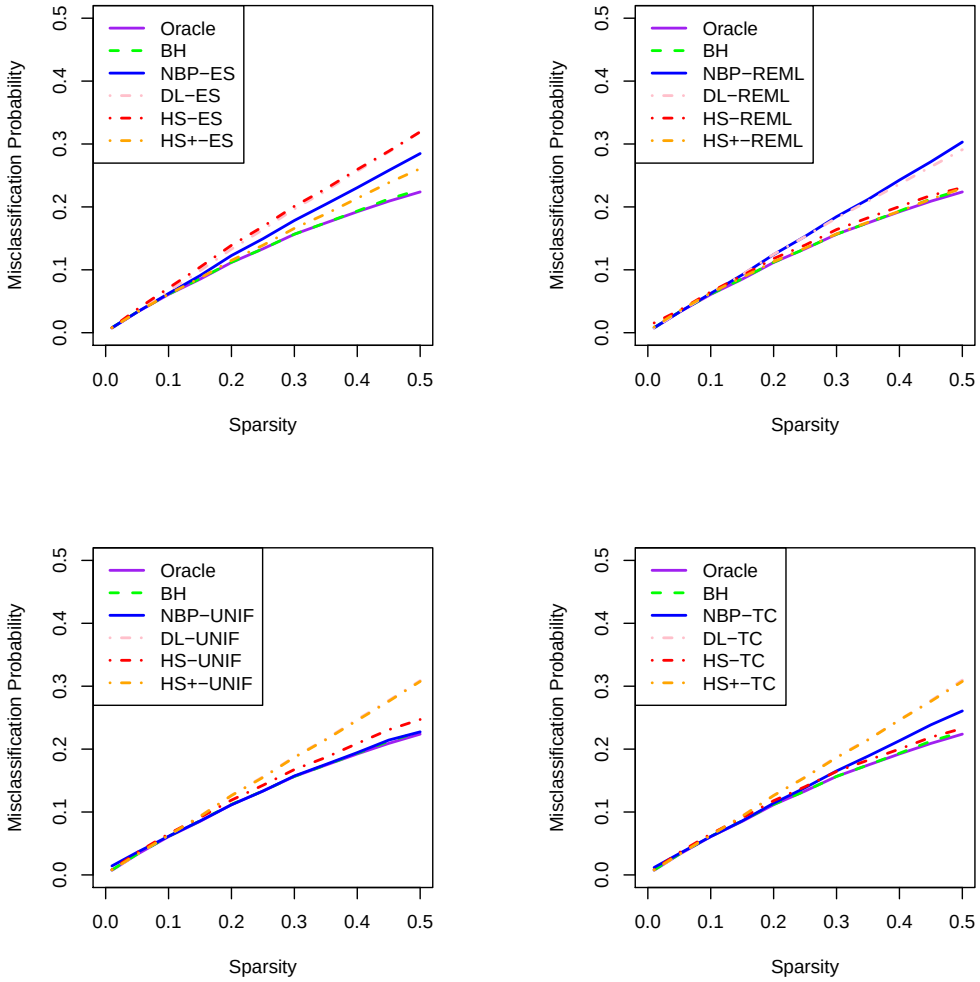

Figure 2. Estimated misclassification probabilities for the NBP, HS, HS+, and DL models when the different estimators for the sparsity parameter are used: estimated sparsity (ES), REML, $\mathcal{U}(1/n, 1)$, and $\mathcal{TC}(0, 1; 1/n, 1)$. The different models are compared to the Bayes Oracle (BO) and Benjamini-Hochberg (BH) procedures.

numerical experiments thus corroborate our theoretical findings that the NBP prior (6) is well-behaved under sparsity. If the ES estimator (22) is used, then the HS+-ES prior performs the best, with the NBP-ES model following closely behind. If the REML estimator (27) is used, then the NBP-REML model performs well under sparsity but not as well as the other methods in more dense situations. Under the truncated Cauchy prior, the NBP-TC model performs the second best behind HS-TC in dense situations. Finally, under the uniform prior, the NBP-UNIF outperforms all the other models and behaves very similarly to the Bayes Oracle across *all* sparsity levels. Based on our empirical results, the NBP prior displays the best overall performance when a is endowed with a uniform prior, $a \sim \mathcal{U}(1/n, 1)$.

Figure 3 provides further justification for using the NBP-UNIF model. By Theorem 3.3, the sparsity parameter a in the NBP prior (6) may be interpreted as the true signal proportion p. In Figure 3, we plot the true sparsity level p against the ES,

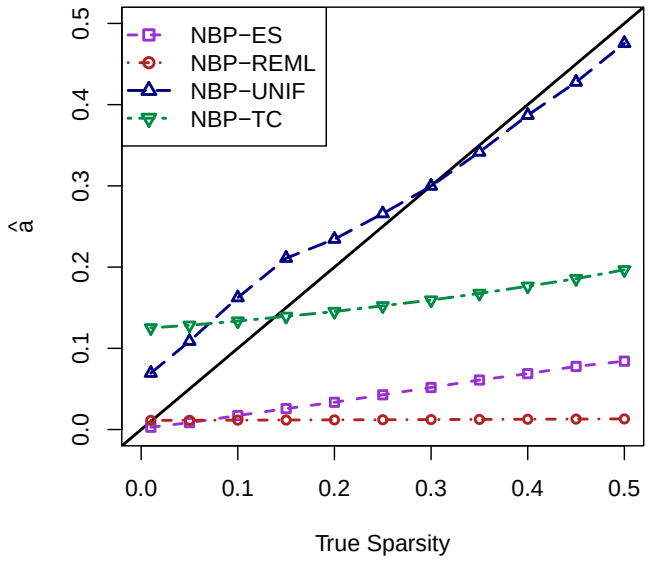

Figure 3. Plot of the true sparsity level p against the estimate for the sparsity parameter $\widehat{a}$ under the NBP-ES, NBP-REML, NBP-UNIF, and NBP-TC models averaged across 100 replications.

REML, $\mathcal{U}(1/n, 1)$, and $\mathcal{TC}(0, 1; 1/n, 1)$ estimators for a averaged across 100 replications (for the hierarchical Bayes methods, we take the posterior mean of $\pi(a|X_1, \ldots, X_n)$ as $\widehat{a}$). The diagonal solid line represents the true p, so the closer an estimate of p is to this diagonal, the more accurate it is. Our plot shows that the ES, REML, and $\mathcal{TC}(0, 1; 1/n, 1)$ estimators for a all tend to *underestimate* the true sparsity as p increases. Meanwhile, the $\mathcal{U}(1/n, 1)$ estimator for a is the closest to the true p for all $p \geq 0.05$. This may explain why NBP-UNIF performs empirically better than the other variants of the NBP model.

5.3. Estimation and False Discovery Rate (FDR) Control

While our focus has been on designing a test procedure with the NBP prior which has the Bayes Oracle property, practitioners may also be interested in estimation of the underlying $\boldsymbol{\theta}$ or in false discovery rate (FDR) control. Let Δ_i and Ω_i be defined as

$$\begin{aligned} \Delta_i &\equiv \{H_{0i} \text{ is rejected when } H_{0i} \text{ is true}\}, \\ \Omega_i &\equiv \{H_{0i} \text{ is rejected when } H_{1i} \text{ is true}\}. \end{aligned}$$

The (empirical) FDR is defined as

$$\text{FDR} = \frac{\sum_{i=1}^{n} I(\Delta_i)}{\max\{1, \sum_{i=1}^{n} I(\Delta_i) + \sum_{i=1}^{n} I(\Omega_i)\}}, \tag{32}$$

and the goal of frequentist FDR control is to design a test such that $\mathbb{E}(FDR) \leq \alpha$ for a prespecified $\alpha \in (0, 1)$. Both estimation and FDR control are separate procedures

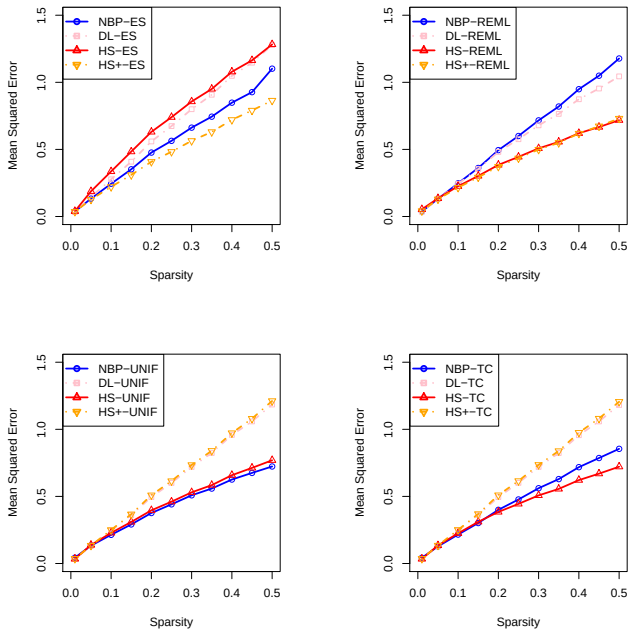

Figure 4. Mean squared error for the NBP, HS, HS+, and DL models when the different estimators for the sparsity parameter are used: estimated sparsity (ES), REML, $\mathcal{U}(1/n, 1)$, and $\mathcal{TC}(0, 1; 1/n, 1)$. The different models are compared to the Bayes Oracle (BO) and Benjamini-Hochberg (BH) procedures.

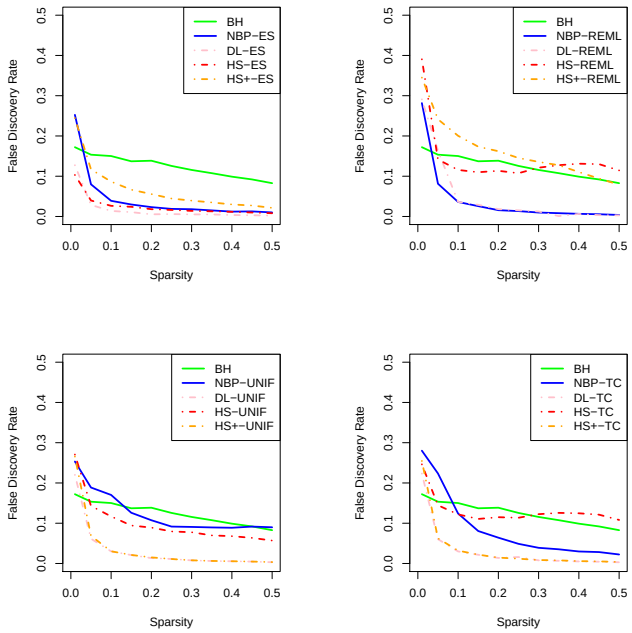

Figure 5. False discovery rates for the NBP, HS, HS+, and DL models when the different estimators for the sparsity parameter are used: estimated sparsity (ES), REML, $\mathcal{U}(1/n, 1)$, and $\mathcal{TC}(0, 1; 1/n, 1)$. The different models are compared to the Bayes Oracle (BO) and Benjamini-Hochberg (BH) procedures.

than the ones considered in this paper and indeed may give conflicting results in terms of 'optimality.' For example, [21] proved that any estimator $\widehat{\boldsymbol{\theta}}$ which asymptotically has FDR of zero cannot simultaneously obtain the minimax estimation rate. Similarly, a procedure which has the Bayes Oracle property is intended to minimize the *total* expected number of misclassified tests (false positives plus false negatives). It is thus conceivable that a test which has high FDR could still be ABOS, provided that the number of false negatives is very low. Conversely, a test which has very low FDR may still have a very high misclassification probability (MP) if the test results in a high number of false negatives.

Thresholding rules (19), (23), (29), and (31) are explicitly designed to minimize the expected *total* number of misclassified tests. Nevertheless, it is worth investigating the estimation quality under the NBP prior and the extent to which these tests control the FDR in our simulation study. To assess the estimation of $\boldsymbol{\theta}$ under the NBP prior, we compute the mean squared error (MSE) about the posterior median, i.e. MSE $= (1/n)\sum_{i=1}^{n}(\widehat{\theta}_i^{\text{med}} - \theta_{0i})^2$, for the NBP-ES, NBP-REML, NBP-UNIF, and NBP-TC models, averaged across 100 replications. We compare the performance to the respective DL, HS, and HS+ models. Our results are plotted in Figure 4. Figure 4 shows that the hierarchical Bayes approaches give the best estimation quality for the NBP prior, with the NBP-UNIF prior outperforming all other methods.

We also plot the FDR (32) for all our models in Figure 5. Figure 5 shows that testing rules (23), (29), and (31) under the NBP, DL, HS, and HS+ priors all control FDR well. For most of the sparsity levels, the FDRs under these shrinkage priors are lower than the FDR under BH. In particular, Figure 5 shows that tests under the plug-in ES and REML estimators give FDR close to zero in dense settings. However, as illustrated in Figures 2 and 4, NBP-REML has the highest *total* misclassification rate and estimation error, indicating that NBP-REML misses a large proportion of actual signals in dense settings. Based on our numerical studies, we recommend the hierarchical Bayes NBP prior with $a \sim \mathcal{U}(1/n, 1)$ as the 'default' implementation for the NBP model. Figure 5 shows that the FDR under the NBP-UNIF model compares favorably to that of the BH procedure. In addition, NBP-UNIF mimics the Bayes Oracle performance the closest and has the lowest estimation error.

If a more conservative test is desired, then we recommend using the NBP-ES model. The NBP-ES model performs slightly worse than NBP-UNIF in terms of MP and MSE, but it has lower FDR. At present, designing theoretically rigorous tests with frequentist FDR control using scale-mixture shrinkage priors (2) is still an open problem.

6. Analysis of a Prostate Cancer Data Set

We demonstrate practical application of the NBP prior using a popular prostate cancer data set introduced by [20]. In this data set, there are gene expression values for $n = 6033$ genes for $m = 102$ subjects, with $m_1 = 50$ normal control subjects and $m_2 = 52$ prostate cancer patients. We aim to identify genes that are significantly different between control and cancer patients. We first conduct a two-sample t-test for each gene and then transform the test statistics $(t_1, ..., t_n)$ to z-scores using the inverse normal cumulative distribution function (CDF) transform $\Phi^{-1}(F_{t_{100}}(t_i))$, where $F_{t_{100}}$ denotes the CDF for the Student's t distribution with 100 degrees of freedom.

With z-scores $(z_1, ..., z_n)$, it is clear that z_i follows a standard normal distribution under the null hypothesis, i.e. $H_{0i} : z_i \sim \mathcal{N}(0, 1), i = 1, \ldots, n$. This allows us to implement the NBP prior on the z-scores to conduct simultaneous testing of H_{0i} :

Table 1. The z-scores and the effect size estimates for the top 10 genes selected by [12] by the NBP-UNIF, DL-UNIF, HS-UNIF, and HS+-UNIF models and the two-groups empirical Bayes model by [12].

Gene	z-score	$\widehat{\theta}_i^{NBP}$	$\widehat{\theta}_i^{DL}$	$\widehat{\theta}_i^{HS}$	$\widehat{\theta}_i^{HS+}$	$\widehat{\theta}_i^{Efron}$
610	5.29	4.87	4.61	4.87	4.87	4.11
1720	4.83	4.39	4.09	4.30	4.37	3.65
332	4.47	3.97	3.62	3.85	3.73	3.24
364	-4.42	-3.94	-3.56	-3.81	-3.85	-3.57
914	4.40	3.85	3.54	3.74	3.71	3.16
3940	-4.33	-3.80	-3.49	-3.53	-3.68	-3.52
4546	-4.29	-3.74	-3.39	-3.58	-3.70	-3.47
1068	4.25	3.69	3.31	3.41	3.35	2.99
579	4.19	3.60	3.32	3.38	3.43	2.92
4331	-4.14	-3.54	-3.14	-3.23	-3.19	-3.30

$\theta_i = 0$ vs. $H_{1i} : \theta_i \neq 0$, $i = 1, ..., n$, to identify genes that are significantly associated with prostate cancer. Additionally, we can also estimate $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ under model (1) using the posterior mean. As argued by [12], $|\theta_i|$ can be interpreted as the effect size of the ith gene for prostate cancer. [12] first analysed this model for this particular data set by obtaining empirical Bayes estimates $\widehat{\theta}_i^{Efron}$, $i = 1, ..., n$, based on the two-groups model (11). In our analysis, we use the posterior means $\widehat{\theta}_i, i = 1, ..., n$, to estimate the strength of association.

We implement the NBP-UNIF model and use classification rule (31) to identify significant genes. For comparison, we also fit this model for the DL-UNIF, HS-UNIF, and HS+-UNIF priors, and benchmark these models to the Benjamini-Hochberg (BH) procedure with FDR $\alpha = 0.10$. The NBP-UNIF model selects 165 out of the 6033 genes as significant, in comparison to 60 genes under the BH procedure. All 60 genes selected by the BH procedure are included in the 166 genes determined to be significant by the NBP prior. The HS-UNIF and HS+-UNIF priors select 55 and 38 genes respectively as significant, while the DL prior selects 102 genes as significant.

Table 1 shows the top 10 genes selected by [12] and their estimated effect size on prostate cancer. We compare [12]'s empirical Bayes posterior mean estimates with the posterior mean estimates under the NBP-UNIF, DL-UNIF, HS-UNIF, and HS+-UNIF priors. Our results confirm the tail robustness of the NBP prior. All of the scale-mixture shrinkage priors shrink the estimated effect size for significant genes less aggressively than Efron's procedure. On this particular dataset, the NBP model shrinks large signals the least of all the methods considered when the sparsity parameter a is endowed with a prior, $a \sim \mathcal{U}(1/n, 1)$.

Figure 5 illustrated that the hierarchical Bayes model with a uniform prior tends to give higher FDR than the models where the sparsity parameter is estimated with the estimated sparsity (ES) plug-in estimator $\widehat{a}^{ES}$ (22). In some applications, it may be better to have tests which are more conservative. With this in mind, we repeat our analysis using (22) as the sparsity parameter and classification rule (23). In this case, the NBP-ES model selected just 72 genes, including all 60 genes selected by the BH procedure. This can partly be explained by the fact that the sparsity parameter estimate was $\widehat{a}^{ES} = 0.0005$ under the NBP-ES model, while the posterior mean estimate for a under the NBP-UNIF model was $\mathbb{E}(a|z_1, \ldots, z_n) = 0.1748$. The DL-ES and HS-ES models were also more conservative, selecting 39 and 4 genes respectively. The HS+-ES model selected 50 genes as significant.

7. Concluding Remarks and Future Work

In this paper, we have studied a scale-mixture shrinkage prior with the beta prime prior (7) as the scale parameters for multiple testing under sparsity. By appropriately estimating the sparsity parameter in the normal-beta prime prior and thresholding the posterior shrinkage weight, the NBP can be used to identify signals in sparse normal mean vectors. We have investigated these testing rules within the decision theoretic framework of [7] and established that the NBP prior has the Bayes Oracle property.

Our results also suggest that scale-mixture shrinkage priors of the most general form (2) can asymptotically attain the exact optimal Bayes risk for multiple testing. In the future, we hope to derive general sufficient conditions under which shrinkage priors (2) are asymptotically Bayes optimal under sparsity. We would also like to provide theoretical justification for the use of the restricted marginal maximum likelihood (REML) and hierarchical Bayes methods presented in Section 4. Previously, [25] showed that these adaptive methods lead to near-minimax *estimation* under the horseshoe prior. Our results suggest that these methods are also optimal for multiple testing and that they are appropriate to use for general shrinkage priors besides the horseshoe.

Finally, there has been a rapid growth in the 'frequentist Bayes' theory field in recent years, but the literature on frequentist assessment of Bayesian multiple testing procedures is only now emerging. In a recent preprint, [10] show that thresholding the posterior under a point-mass spike-and-slab prior at level $\alpha \in (0, 1)$ asymptotically gives frequentist false discovery rate (FDR) control of level α (up to a multiplicative constant) for sparse normal means. We conjecture that thresholding rules based on the posterior shrinkage weight under the NBP prior (6) – and under general scale-mixture shrinkage priors (2) – can also be constructed for frequentist FDR control.

Acknowledgments

The authors would like to thank Dr. Anirban Bhattacharya and Dr. Xueying Tang for sharing their codes, which were modified to generate Figures 1-5. We are grateful to two anonymous reviewers, the Associate Editor, and the Editors whose thoughtful comments and suggestions helped to greatly improve this paper.

Disclosure statement

The authors declare that we have no conflicts of interest in the authorship or publication of this contribution.

Supplementary Data

The Supplementary Materials document contains the proofs for the propositions and theorems in Sections 2.1, 3.3, and 3.4, as well as the technical details for implementing our model and a comparison of the posterior shrinkage weights with the theoretical posterior inclusion probabilities under the true model (11).

Code to implement our model is available in the R package `NormalBetaPrime`, which also contains the prostate cancer data set analysed in Section 6.

References

[1] Armagan A, Clyde M, Dunson DB. Generalized beta mixtures of gaussians. NeurIPS 2011; 24:523-531.

[2] Armagan A, Dunson DB, Lee J. Generalized double pareto shrinkage. Statist. Sinica. 2013;23:119-143.

[3] Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 1995;57:289-300.

[4] Berger J. A robust generalized bayes estimator and confidence region for a multivariate normal mean. Ann. Statist. 1980;8:716-761.

[5] Bhadra A, Datta J, Polson NG, Willard B. The horseshoe+ estimator of ultra-sparse signals. Bayesian Anal. 2017;12:1105-1131.

[6] Bhattacharya A, Pati D, Pillai NS, Dunson DB. Dirichlet-laplace priors for optimal shrinkage. J. Amer. Statist. Assoc. 2015;110:1479-1490.

[7] Bogdan M, Chakrabarti A, Frommlet F, Ghosh JK. Asymptotic bayes-optimality under sparsity of some multiple testing procedures. Ann. Statist. 2011;39:1551-1579.

[8] Carvalho CM, Polson NG, Scott JG. Handling sparsity via the horseshoe. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, PMLR. 2009;5:73-80.

[9] Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika. 2010;97:465-480.

[10] Castillo I, Roquain E. On spike and slab empirical bayes multiple testing. arXiv pre-print arXiv: 1808.09748. 2018.

[11] Datta J, Ghosh JK. Asymptotic properties of bayes risk for the horseshoe prior. Bayesian Anal. 2013;8:111-132.

[12] Efron B. The future of indirect evidence. Statist. Sci. 2010;25:145-157.

[13] Ghosh P, Chakrabarti A. Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. Bayesian Anal. 2017;12:1133-1161.

[14] Ghosh P, Tang X, Ghosh M, Chakrabarti A. Asymptotic properties of bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. Bayesian Anal. 2016;11:753-796.

[15] Griffin JE, Brown PJ. Some priors for sparse regression modeling. Bayesian Anal. 2013;8:691-702.

[16] Johnstone IM, Silverman BW. Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. Ann. Statist. 2004;32:1594-1649.

[17] Park T, Casella G. The bayesian lasso. J. Amer. Statist. Assoc. 2008;103:681-686.

[18] Polson NG, Scott JG. On the half-cauchy prior for a global scale parameter. Bayesian Anal. 2012;7:887-902.

[19] Salomond JB. Risk quantification for the thresholding rule for multiple testing using gaussian scale mixtures. arXiv pre-print arXiv: 1711.08705. 2017.

[20] Singh D, Febbo PG, Ross K, Jackson DG, Maonla J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell. 2002;2:203-209.

[21] Song Q, Cheng G. Optimal false discovery control of minimax estimator. arXiv pre-print arXiv: 1812.10013. 2018.

[22] Strawderman WE. Proper bayes minimax estimators of the multivariate normal mean. Ann. Math. Statist. 1971;42:385-388.

[23] van der Pas S, Salomond JB, Schmidt-Hieber J. Conditions for posterior contraction in the sparse normal means problem. Eletron. J. Statist. 2016;10:976-1000.

[24] van der Pas SL, Kleijn BJK, van der Vaart AW. The horseshoe estimator: Posterior concentration around nearly black vectors. Electron. J. Statist. 2014;8:2585-2618.

[25] van der Pas S, Szabó B, van der Vaart A. Adaptive posterior contraction rates for the horseshoe. Electron. J. Statist. 2017;11:3196-3225.

[26] Wellcome Trust. Genome-wide association study of 14,000 cases of seven common diseases

and 3000 shared controls. Nature. 2007;447:661-678.