

# ON THE BETA PRIME PRIOR FOR SCALE PARAMETERS IN HIGH-DIMENSIONAL BAYESIAN REGRESSION MODELS

Ray Bai and Malay Ghosh

*University of Pennsylvania and University of Florida*

*Abstract:* We study high-dimensional Bayesian linear regression with a general beta prime distribution for the scale parameter. Under the assumption of sparsity, we show that appropriate selection of the hyperparameters in the beta prime prior leads to the (near) minimax posterior contraction rate when  $p \gg n$ . For finite samples, we propose a data-adaptive method for estimating the hyperparameters based on marginal maximum likelihood (MML). This enables our prior to adapt to both sparse and dense settings, and under our proposed empirical Bayes procedure, the MML estimates are never at risk of collapsing to zero. We derive efficient Monte Carlo EM and variational EM algorithms for implementing our model, which are available in the R package `NormalBetaPrime`. Simulations and analysis of a gene expression data set illustrate our model's self-adaptivity to varying levels of sparsity and signal strengths.

*Key words and phrases:* beta prime density, empirical Bayes, high-dimensional data, scale mixtures of normal distributions, posterior contraction

## 1. Introduction

### 1.1 Background

Consider the classical linear regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where  $\mathbf{y}$  is an  $n$ -dimensional response vector,  $\mathbf{X}_{n \times p} = [\mathbf{X}_1, \dots, \mathbf{X}_p]$  is a fixed regression matrix with  $n$  samples and  $p$  covariates,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  is a  $p$ -dimensional vector of unknown regression coefficients, and  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ , where  $\sigma^2$  is the unknown variance. Throughout this paper, we assume that  $\mathbf{y}$  and  $\mathbf{X}$  have been centered at 0 so there is no intercept in our model.

In recent years, the high-dimensional setting where  $p \gg n$  has received considerable attention. This scenario is now routinely encountered in areas as diverse as medicine, astronomy, and finance, just to name a few. In the Bayesian framework, there have been numerous methods proposed to handle the “large  $p$ , small  $n$ ” scenario, including spike-and-slab priors with point masses at zero (e.g., Martin et al. (2017), Castillo et al. (2015), Yang et al. (2016)), continuous spike-and-slab priors (e.g., Narisetty and He (2014), Ročková and George (2018)), nonlocal priors (e.g. Johnson and Rossell (2012), Rossell and Telesca (2017), Shin et al. (2018)), and scale-mixture shrinkage priors (e.g. van der Pas et al. (2016), Song and Liang (2017)).

These priors have been shown to have excellent empirical performance and possess strong theoretical properties, including model selection consistency, (near) minimax posterior contraction, and Bernstein-von Mises theorems. In this paper, we restrict our focus to the scale-mixture shrinkage approach.

Under (1.1), scale-mixture shrinkage priors typically take the form,

$$\begin{aligned}\beta_i | (\sigma^2, \omega_i^2) &\sim \mathcal{N}(0, \sigma^2 \omega_i^2), i = 1, \dots, p, \\ \omega_i^2 &\sim \pi(\omega_i^2), i = 1, \dots, p, \\ \sigma^2 &\sim \mu(\sigma^2),\end{aligned}\tag{1.2}$$

where  $\pi$  and  $\mu$  are densities on the positive reals. Priors of this form (1.2) have been considered by many authors, e.g., Park and Casella (2008), Carvalho et al. (2010), Griffin and Brown (2010), Bhattacharya et al. (2015), Armagan et al. (2011), Armagan et al. (2013).

Computationally, scale-mixture priors are very attractive. Discontinuous spike-and-slab priors require searching over  $2^p$  models, while continuous spike-and-slab priors and nonlocal priors almost always result in multimodal posteriors. As a result, Markov chain Monte Carlo (MCMC) algorithms are prone to being trapped at a local posterior mode, and MCMC can suffer from slow convergence for these models. Scale-mixture shrinkage priors, on the other hand, do not face these drawbacks because they are continuous and typically give rise to unimodal posteriors as long as the signal-to-noise

ratio is not too low. Additionally, there have been recent advances for fast sampling from scale-mixture priors that scale linearly in time with  $p$ , e.g. Bhattacharya et al. (2016), Johndrow et al. (2017).

Scale-mixture priors have been studied primarily under sparsity assumptions. If sparse recovery of  $\beta$  is desired, the prior  $\pi(\cdot)$  can be constructed so that it contains heavy mass around zero and heavy tails. This way, the posterior density  $\pi(\beta|\mathbf{y})$  is heavily concentrated around  $\mathbf{0} \in \mathbb{R}^p$ , while the heavy tails prevent overshrinkage of the true active covariates. While sparsity is often a reasonable assumption, it is not always appropriate. Zou and Hastie (2005) demonstrated an example where this assumption is violated: in microarray experiments with highly correlated predictors, it is often desirable for all genes which lie in the same biological pathway to be selected, even if the final model is not parsimonious. Zou and Hastie (2005) introduced the elastic net to overcome the inability of the LASSO (Tibshirani (1996)) to select more than  $n$  variables. In the Bayesian literature, there seems to be little study of the appropriateness of scale-mixture priors in dense settings. Ideally, we would like our priors on  $\beta$  in (1.1) to be able to handle *both* sparse and non-sparse situations.

Another important issue to consider is the selection of hyperparameters in our priors on  $\beta$ . Many authors, e.g. Narisetty and He (2014), Yang et

al. (2016), Martin et al. (2017), have proposed fixing hyperparameters *a priori* based on asymptotic arguments (such as consistency or minimaxity) or by minimizing some criterion such as Bayesian information criterion (BIC) or deviance information criterion (DIC) (e.g. Song and Liang (2017), Spiegelhalter et al. (2002)). In this paper, we argue in favor of a different approach based on marginal maximum likelihood (MML) estimation, which avoids the need for hyperparameter tuning by the user.

In this manuscript, we consider a scale-mixture prior (1.2) with the beta prime density as the scale parameters. We call our model the normal-beta prime (NBP) model. Our main contributions are summarized as follows:

- We show that for high-dimensional linear regression, the NBP model can serve as both a sparse *and* a non-sparse prior. We prove that under sparsity and appropriate regularity conditions, the NBP prior asymptotically obtains the (near) minimax posterior contraction rate.
- In the absence of prior knowledge about sparsity or non-sparsity, we propose an empirical Bayes variant of the NBP model which is *self-adaptive* and which learns the true sparsity level from the data. Under our procedure, the hyperparameter estimates are never at risk of collapsing to zero. This is not the case for many other choices of priors, where empirical Bayes estimates can often result in degenerate priors.

- We derive efficient Monte Carlo EM and variational EM algorithms for implementing the self-adaptive NBP model. Our algorithms embed the EM algorithm for estimating the hyperparameters into posterior simulation updates, so that they do not need to be tuned separately.

The rest of the paper is structured as follows. In Section 2, we introduce the NBP prior for Bayesian linear regression. In Section 3, we derive posterior contraction rates for the NBP when  $p \gg n$ . In Section 4, we introduce the self-adaptive NBP model, which automatically learns the true sparsity pattern from the data. In Section 5, we introduce the algorithms for implementing the self-adaptive NBP. Section 6 provides simulation studies of our model, and Section 7 gives an application to a gene expression data set.

## 1.2 Notation

For two nonnegative sequences  $\{a_n\}$  and  $\{b_n\}$ , we write  $a_n \asymp b_n$  to denote  $0 < \liminf_{n \rightarrow \infty} a_n/b_n \leq \limsup_{n \rightarrow \infty} a_n/b_n < \infty$ . If  $\lim_{n \rightarrow \infty} a_n/b_n = 0$ , we write  $a_n = o(b_n)$  or  $a_n \prec b_n$ . We use  $a_n \lesssim b_n$  or  $a_n = O(b_n)$  to denote that for sufficiently large  $n$ , there exists a constant  $C > 0$  independent of  $n$  such that  $a_n \leq Cb_n$ . For a vector  $\mathbf{v} \in \mathbb{R}^p$ , we let  $\|\mathbf{v}\|_0 := \sum_i \mathbf{1}(v_i \neq 0)$ ,  $\|\mathbf{v}\|_1 := \sum_i |v_i|$ , and  $\|\mathbf{v}\|_2 := \sqrt{\sum_i v_i^2}$  denote the  $\ell_0$ ,  $\ell_1$ , and  $\ell_2$  norms respectively. For a set  $\mathcal{A}$ , we denote its cardinality as  $|\mathcal{A}|$ .

---

## 2. THE NORMAL-BETA PRIME (NBP) MODEL

### 2. The Normal-Beta Prime (NBP) Model

The beta prime density is given by

$$\pi(\omega_i^2) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} (\omega_i^2)^{a-1} (1+\omega_i^2)^{-a-b}. \quad (2.1)$$

In particular, setting  $a = b = 0.5$  in (2.1) yields the half-Cauchy prior  $\mathcal{C}^+(0,1)$  for  $\omega_i$ . For multivariate normal means estimation, Polson and Scott (2012) conducted numerical experiments for different choices of  $(a, b)$  in (2.1) and argued that the half-Cauchy prior should be a default prior for scale parameters. Pérez et al. (2017) also generalized the beta prime density (2.1) to the Scaled Beta2 family of scale priors by adding an additional scaling parameter to (2.1). These authors did not consider linear regression models under general design matrices.

Under the normal-beta prime (NBP) model, we place a normal-scale mixture prior (1.2) with the beta prime density (2.1) as the scale parameter for each of the individual coefficients in  $\boldsymbol{\beta}$  and an inverse gamma prior  $\mathcal{IG}(c, d)$  prior on  $\sigma^2$ , where  $c, d > 0$ . Letting  $\beta'(a, b)$  denote the beta prime distribution (2.1) with hyperparameters  $a > 0, b > 0$ , our full model is

$$\begin{aligned} \beta_i | \omega_i^2, \sigma^2 &\sim \mathcal{N}(0, \sigma^2 \omega_i^2), \quad i = 1, \dots, p, \\ \omega_i^2 &\sim \beta'(a, b), \quad i = 1, \dots, p, \\ \sigma^2 &\sim \mathcal{IG}(c, d). \end{aligned} \quad (2.2)$$

## 2. THE NORMAL-BETA PRIME (NBP) MODEL

---

For our model (2.2), we can choose very small values of  $c$  and  $d$  in order to make the prior on  $\sigma^2$  relatively noninfluential and noninformative (e.g., a good default choice is  $c = d = 10^{-5}$ ). The most critical hyperparameter choices governing the performance of our model are  $(a, b)$ .

**Proposition 2.1.** *Suppose that we endow  $(\boldsymbol{\beta}, \sigma^2)$  with the priors in (2.2). Then the marginal distribution,  $\pi(\beta_i|\sigma^2), i = 1, \dots, p$ , is unbounded with a singularity at zero for any  $0 < a \leq 1/2$ .*

*Proof.* See Proposition 2.1 in Bai and Ghosh (2019). □

Proposition 2.1 implies that in order to facilitate sparse recovery of  $\boldsymbol{\beta}$ , we should set the hyperparameter  $a$  to be a small value. This would force the NBP prior to place most of its mass near zero, and thus, the posterior  $\pi(\boldsymbol{\beta}|\mathbf{y})$  would also be concentrated near  $\mathbf{0} \in \mathbb{R}^p$ . Figure 1 plots the marginal density,  $\pi(\beta|\sigma^2)$ , for a single  $\beta$ . When  $a = 0.1$ , the marginal density contains a singularity at zero, and the probability mass is heavily concentrated near zero. However, when  $a = 2$ , the marginal density does not contain a pole at zero, and the tails are significantly heavier.

Figure 1 shows that the NBP model can serve as both a sparse and a non-sparse prior. If we have prior knowledge that the true model is sparse with a few large signal values, we can fix  $a$  to be a small value. On the

### 3. POSTERIOR CONTRACTION RATES UNDER THE NBP PRIOR

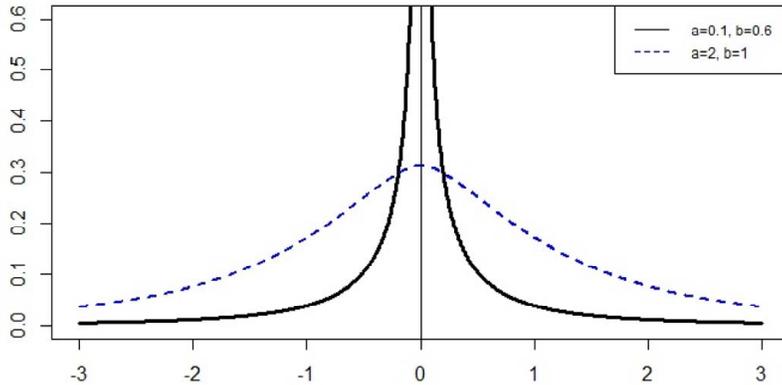


Figure 1: The marginal densities of the NBP prior,  $\pi(\beta|\sigma^2)$ , with  $\sigma^2 = 1$ . A small  $a$  leads to a pole at zero. A large  $a$  removes the singularity.

other hand, if we know that the true model is dense, we can set  $a$  to a larger value, so we have a more diffuse prior. Then there would be less shrinkage of individual covariates in the posterior distribution. In Section 4, we will introduce the *self-adaptive* NBP model, which automatically learns the true sparsity level from the data and avoids the need for tuning by the user.

### 3. Posterior Contraction Rates Under the NBP Prior

For our theoretical analysis, we allow  $p$  to diverge to infinity as sample size  $n$  grows. We write  $p$  as  $p_n$  to emphasize its dependence on  $n$ . We work under the frequentist assumption that there is a true data-generating model, i.e.,

$$\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_n, \quad (3.1)$$

### 3. POSTERIOR CONTRACTION RATES UNDER THE NBP PRIOR

---

where  $\boldsymbol{\epsilon}_n \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$  and  $\sigma_0^2$  is a fixed noise parameter.

Let  $s_n = \|\boldsymbol{\beta}_0\|_0$  denote the size of the true model, and suppose that  $s_n = o(n/\log p_n)$ . Under (3.1) and appropriate regularity conditions, Raskutti et al. (2011) showed that the minimax estimation rate for any point estimator  $\hat{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}_0$  under  $\ell_2$  error loss is  $\sqrt{s_n \log(p_n/s_n)/n}$ . Many frequentist point estimators such as the LASSO (Tibshirani (1996)) have been shown to attain the *near*-minimax rate of  $\sqrt{s_n \log p_n/n}$  under  $\ell_2$  error loss.

In the Bayesian paradigm, we are mainly concerned with the rate at which the *entire* posterior distribution contracts around the true  $\boldsymbol{\beta}_0$ . Letting  $\mathbb{P}_0$  denote the probability measure underlying (3.1) and  $\Pi(\boldsymbol{\beta}|\mathbf{y}_n)$  denote the posterior of  $\boldsymbol{\beta}$ , our aim is to find a positive sequence  $r_n$  such that

$$\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \geq Mr_n | \mathbf{y}_n) \rightarrow 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \rightarrow \infty,$$

for some constant  $M > 0$ . The frequentist minimax convergence rate is a useful benchmark for the speed of contraction  $r_n$ , since the posterior cannot contract faster than the minimax rate (Ghosal et al. (2000)).

We are also interested in posterior *compressibility* (Bhattacharya et al. (2015)), which allows us to quantify how well the NBP posterior captures the true sparsity level  $s_n$ . Since the NBP prior is absolutely continuous, it assigns zero mass to exactly sparse vectors. To approximate the model size for the NBP model, we use the following generalized notion of sparsity

### 3. POSTERIOR CONTRACTION RATES UNDER THE NBP PRIOR

---

(Bhattacharya et al. (2015)). For some  $\delta > 0$ , we define the generalized inclusion indicator and generalized dimensionality, respectively, as

$$\gamma_\delta(\boldsymbol{\beta}) = I(|\boldsymbol{\beta}/\sigma| > \delta) \text{ and } |\boldsymbol{\gamma}_\delta(\boldsymbol{\beta})| = \sum_{i=1}^{p_n} \gamma_\delta(\beta_i). \quad (3.2)$$

The generalized dimensionality counts the number of covariates in  $\boldsymbol{\beta}/\sigma$  that fall outside the interval  $[-\delta, +\delta]$ . With appropriate choice of  $\delta$ , the prior is said to have the posterior compressibility property if the probability that  $|\boldsymbol{\gamma}_\delta(\boldsymbol{\beta})|$  asymptotically exceeds a constant multiple of the true sparsity level  $s_n$  tends to 0 as  $n \rightarrow \infty$ , i.e.

$$\Pi(\boldsymbol{\beta} : |\boldsymbol{\gamma}_\delta(\boldsymbol{\beta})| \geq A s_n | \mathbf{y}_n) \rightarrow 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \rightarrow \infty,$$

for some constant  $A > 0$ .

#### 3.1 Near-Minimax Posterior Contraction Under the NBP Prior

We first introduce the following set of regularity conditions, which come from Song and Liang (2017). Let  $s_n$  denote the size of the true model and  $\lambda_{\min}(\mathbf{A})$  denote the minimum eigenvalue of a symmetric matrix  $\mathbf{A}$ .

(A1) All the covariates are uniformly bounded. For simplicity, we assume they are all bounded by 1.

(A2)  $p_n \gg n$ .

### 3. POSTERIOR CONTRACTION RATES UNDER THE NBP PRIOR

---

(A3) Let  $\xi \subset \{1, \dots, p_n\}$ , and let  $\mathbf{X}_\xi$  denote the submatrix of  $\mathbf{X}_n$  that contains the columns with indices in  $\xi$ . There exists some integer  $\bar{p}$  (depending on  $n$  and  $p_n$ ) and fixed constant  $t_0$  such that  $\bar{p} \succ s$  and  $\lambda_{\min}(\mathbf{X}_\xi^\top \mathbf{X}_\xi) \geq nt_0$  for any model of size  $|\xi| \leq \bar{p}$ .

(A4)  $s_n = o(n/\log p_n)$ .

(A5)  $\max_j \{|\beta_{0j}/\sigma_0|\} \leq \gamma_3 E_n$  for some  $\gamma_3 \in (0, 1)$ , and  $E_n$  is nondecreasing with respect to  $n$ .

Assumption (A3) is a minimum restricted eigenvalue (RE) condition which ensures that  $\mathbf{X}_n^\top \mathbf{X}_n$  is locally invertible over sparse sets. When  $p_n \gg n$ , minimum RE conditions are imposed to render  $\beta_0$  estimable. Assumption (A4) restricts the growth of  $s_n$ , and (A5) constrains the size of the signals in  $\beta_0$  to be  $O(E_n)$  for some nondecreasing sequence  $E_n$ .

As we illustrated in Section 2, the hyperparameter  $a$  in the NBP prior is mainly what controls the amount of posterior mass around zero. Hence, it plays a crucial role in our theory. We rewrite  $a$  as  $a_n$  to emphasize its dependence on  $n$ .

**Theorem 3.1.** *Assume that Assumptions (A1)-(A5) hold, with  $\log(E_n) = O(\log p_n)$  for Assumption (A5). Let  $r_n = M\sqrt{s_n \log p_n/n}$  for some fixed constant  $M > 0$ , and let  $k_n \asymp (\sqrt{s_n \log p_n/n})/p_n$ . Suppose that we place*

### 3. POSTERIOR CONTRACTION RATES UNDER THE NBP PRIOR

the NBP prior (2.2) on  $(\boldsymbol{\beta}, \sigma^2)$ , with  $a_n \lesssim k_n^2 p_n^{-(1+u)}$ , for some  $u > 0$ , and  $b \in (1, \infty)$ . Then under (3.1), the following hold:

$$\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq c_1 \sigma_0 r_n | \mathbf{y}_n) \rightarrow 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \rightarrow \infty, \quad (3.3)$$

$$\Pi(\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \geq c_1 \sigma_0 \sqrt{s} r_n | \mathbf{y}_n) \rightarrow 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \rightarrow \infty, \quad (3.4)$$

$$\Pi(\boldsymbol{\beta} : \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_0\|_2 \geq c_0 \sigma_0 \sqrt{n} r_n | \mathbf{Y}_n) \rightarrow 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \rightarrow \infty, \quad (3.5)$$

$$\Pi(\boldsymbol{\beta} : |\gamma_{k_n}(\boldsymbol{\beta})| \geq \tilde{q}_n | \mathbf{y}_n) \rightarrow 0 \text{ a.s. } \mathbb{P}_0 \text{ as } n \rightarrow \infty, \quad (3.6)$$

where  $c_0 > 0, c_1 > 0$ ,  $|\gamma_{k_n}(\boldsymbol{\beta})| = \sum_i I(|\beta_i/\sigma| > k_n)$ , and  $\tilde{q}_n \asymp s_n$ .

The proof of Theorem 3.1 is based on verifying a set of conditions by Song and Liang (2017) and can be found in the Supplementary Material. In particular, (3.3)-(3.5) show that by fixing  $a_n \lesssim p_n^{-(3+u)} \sqrt{s_n \log p_n/n}$ ,  $u > 0$ , and  $b \in (1, \infty)$  as the hyperparameters  $(a_n, b)$  in (2.2), the NBP model's posterior contraction rates under  $\ell_2$ ,  $\ell_1$ , and prediction error loss are the familiar near-optimal rates of  $\sqrt{s_n \log p_n/n}$ ,  $s_n \sqrt{\log p_n/n}$ , and  $\sqrt{s_n \log p_n}$  respectively. By setting  $\delta = k_n \asymp (\sqrt{s_n \log p_n/n})/p_n$  in our generalized inclusion indicator (3.2), (3.6) also shows that the NBP possesses posterior compressibility, i.e. the probability that the generalized dimension size  $|\gamma_{k_n}(\boldsymbol{\beta})|$  is a constant multiple larger than  $s_n$  asymptotically vanishes.

Our result relies on setting the hyperparameter  $a_n$  to be a value dependent upon the sparsity level  $s_n$ . Previous theoretical results for scale-

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

mixture shrinkage priors, e.g. van der Pas et al. (2016), Song and Liang (2017), also rely on fixing hyperparameters to quantities that depend on  $s_n$  in order for these priors obtain minimax posterior contraction. If we want to *a priori* fix the hyperparameters  $(a, b)$  based on asymptotic arguments, we could first obtain an estimate of  $s_n$ ,  $\hat{s}_n$ , and then set  $a_n$  as  $a_n := p_n^{-(3+u)} \sqrt{\hat{s}_n \log p_n/n}$ ,  $u > 0$ . For example, we could take  $\hat{s}_n = \|\hat{\beta}^{ALasso}\|_0$ , where  $\hat{\beta}^{ALasso}$  is an adaptive LASSO solution (Zou (2006)) to (1.1). Fixing  $a_n := p_n^{-(3+u)} \sqrt{\log n/n}$ ,  $u > 0$ , would also satisfy the conditions in our theorem (since  $\log n \prec s_n \log p_n$ ), thus removing the need to estimate  $s_n$ .

#### 4. Empirical Bayes Estimation of Hyperparameters

While fixing  $(a, b)$  *a priori* as  $a = p^{-(3+u)} \sqrt{\log n/n}$ , for some  $u > 0$ , and  $b \in (1, \infty)$  would lead to (near) minimax posterior contraction under conditions (A1)-(A5), this would not allow the NBP prior to adapt to varying patterns of sparsity or signal strengths. The minimum restricted eigenvalue assumption (A3) is also computationally infeasible to verify in practice. Dobriban and Fan (2016) showed that, given an arbitrary design matrix  $\mathbf{X}$ , verifying that the minimum RE condition holds is an NP-hard problem. Finally, there is no practical way of verifying that the model size condition (A4) that  $s = o(n/\log p)$  holds, or that the true model is even sparse.

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

For these reasons, we do not recommend fixing the hyperparameters in the NBP model based on asymptotic arguments. Instead, we prefer to *learn* the true sparsity pattern from the data. One way to do this is to use marginal maximum likelihood (MML). The marginal likelihood,  $f(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}, \sigma^2)d(\boldsymbol{\beta}, \sigma^2)$ , is the probability the model gives to the observed data with respect to the prior (or the “model evidence”). Hence, choosing the prior hyperparameters to maximize  $f(\mathbf{y})$  gives the maximum “model evidence,” and we can learn the most likely sparsity level from the data. One potential shortcoming with MML is that it can lead to degenerate priors. However, this problem is avoided under the NBP prior.

We propose an EM algorithm to obtain the MML estimates of  $(a, b)$ . Henceforth, we refer to this empirical Bayes variant of the NBP model as the *self-adaptive* NBP model. To construct the EM algorithm, we first note that the beta prime density can be rewritten as a product of an independent gamma and inverse gamma densities. Thus, we may reparametrize (2.2) as

$$\begin{aligned}
 \beta_i | (\omega_i^2, \lambda_i^2 \xi_i^2) &\sim \mathcal{N}(0, \sigma^2 \lambda_i^2 \xi_i^2), & i = 1, \dots, p, \\
 \lambda_i^2 &\sim \mathcal{G}(a, 1), & i = 1, \dots, p, \\
 \xi_i^2 &\sim \mathcal{IG}(b, 1), & i = 1, \dots, p, \\
 \sigma^2 &\sim \mathcal{IG}(c, d).
 \end{aligned}
 \tag{4.1}$$

The logarithm of the joint posterior under the reparametrized NBP prior

---

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

(4.1) is given by

$$\begin{aligned}
& - \left( \frac{n+p}{2} \right) \log(2\pi) - \left( \frac{n+p}{2} + c + 1 \right) \log(\sigma^2) - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\
& - \sum_{i=1}^p \frac{\beta_i^2}{2\lambda_i^2 \xi_i^2 \sigma^2} - p \log(\Gamma(a)) + \left( a - \frac{3}{2} \right) \sum_{i=1}^p \log(\lambda_i^2) - \sum_{i=1}^p \lambda_i^2 - p \log(\Gamma(b)) \\
& - \left( b + \frac{3}{2} \right) \sum_{i=1}^p \log(\xi_i^2) - \sum_{i=1}^p \frac{1}{\xi_i^2} + c \log(d) - \log(\Gamma(c)) - \frac{d}{\sigma^2}. \tag{4.2}
\end{aligned}$$

Thus, at the  $k$ th iteration the EM algorithm, the conditional log-likelihood on  $\nu^{(k-1)} = (a^{(k-1)}, b^{(k-1)})$  and  $\mathbf{y}$  in the E-step is given by

$$\begin{aligned}
Q(\nu|\nu^{(k-1)}) &= -p \log(\Gamma(a)) + a \sum_{i=1}^p \mathbb{E}_{a^{(k-1)}} [\log(\lambda_i^2) | \mathbf{y}] - p(\log \Gamma(b)) \\
&\quad - b \sum_{i=1}^p \mathbb{E}_{b^{(k-1)}} [\log(\xi_i^2) | \mathbf{y}] + \text{terms not involving } a \text{ or } b. \tag{4.3}
\end{aligned}$$

The M-step maximizes  $Q(\nu|\nu^{(k-1)})$  over  $\nu = (a, b)$  to produce the next estimate  $\nu^{(k)} = (a^{(k)}, b^{(k)})$ . That is, we find  $(a, b)$ ,  $a \geq 0, b \geq 0$ , such that

$$\begin{aligned}
\frac{\partial Q}{\partial a} &= -p\psi(a) + \sum_{i=1}^p \mathbb{E}_{a^{(k-1)}} [\log(\lambda_i^2) | \mathbf{y}] = 0, \\
\frac{\partial Q}{\partial b} &= -p\psi(b) - \sum_{i=1}^p \mathbb{E}_{b^{(k-1)}} [\log(\xi_i^2) | \mathbf{y}] = 0,
\end{aligned} \tag{4.4}$$

where  $\psi(x) = d/dx(\Gamma(x))$  denotes the digamma function. We can solve for  $(a, b)$  in (4.4) numerically by using a fast root-finding algorithm such as Newton's method. The summands,  $\mathbb{E}_{a^{(k-1)}} [\log(\lambda_i^2) | \mathbf{y}]$  and  $\mathbb{E}_{b^{(k-1)}} [\log(\xi_i^2) | \mathbf{y}]$ ,  $i = 1, \dots, p$ , in (4.4) can be estimated from either the mean of  $M$  Gibbs samples based on  $\nu^{(k-1)}$ , for sufficiently large  $M > 0$  (as in Casella (2011)),

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

or from the  $(k - 1)$ st iteration of the mean field variational Bayes (MFVB) algorithm (as in Leday et al. (2017)).

**Theorem 4.1.** *At every  $k$ th iteration of the EM algorithm for the self-adaptive NBP model, there exists a unique solution  $\nu^{(k)} = (a^{(k)}, b^{(k)})$ , which maximizes (4.3) in the  $M$ -step. Moreover,  $a^{(k)} > 0$ ,  $b^{(k)} > 0$  at the  $k$ th iteration.*

The proof of Theorem 4.1 can be found in the Supplementary Material. Theorem 4.1 ensures that we will not encounter the issue of the sparsity parameter  $a$  (or the parameter  $b$ ) collapsing to zero. Empirical Bayes estimates of zero are a major concern for MML approaches to estimating hyperparameters in Bayesian regression models. For example, in  $g$ -priors,

$$\boldsymbol{\beta} | \sigma^2 \sim \mathcal{N}_p(\boldsymbol{\gamma}, g\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}),$$

George and Foster (2000) showed that the MML estimate of the parameter  $g$  could equal zero. In global-local shrinkage priors of the form,

$$\beta_i | (\lambda_i^2, \sigma^2) \sim \mathcal{N}(0, \sigma^2 \tau^2 \lambda_i^2), \quad \lambda_i^2 \sim \pi(\lambda_i^2), \quad i = 1, \dots, p,$$

the variance rescaling parameter  $\tau$  is also at risk of being estimated as zero under MML (Polson and Scott (2010), Tiao and Tan (1965), Carvalho et al. (2009), Datta and Ghosh (2013)). Finally, Scott and Berger (2010) proved

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

that if we endow (1.1) with a binomial model selection prior,

$$\pi(\mathbf{M}_\gamma|\theta) = \theta^{k_\gamma}(1 - \theta)^{p-k_\gamma},$$

where  $\mathbf{M}_\gamma$  is the model indexed by  $\gamma \subset \{1, \dots, p\}$  and  $k_\gamma$  represents the number of included variables in the model, the MML estimate of the mixing proportion  $\theta$  could be estimated as either 0 or 1, leading to a degenerate prior. Clearly, the marginal maximum likelihood approach for tuning hyperparameters is not without problems, as it could potentially lead to degenerate priors in high-dimensional regression. However, with the NBP prior, we can easily incorporate a data-adaptive procedure for estimating the hyperparameters while avoiding this potential pitfall.

In the aforementioned examples, placing priors on  $g$ ,  $\tau$ , or  $\theta$  with strictly positive support or performing cross-validation or *restricted* marginal maximum likelihood estimation over a range of strictly positive values can help to avoid collapse to zero. The hierarchical Bayes approach does not quite address the issue of misspecification of hyperparameters, since these still need to be specified in the additional priors. If we use cross-validation, the “optimal” choice or spacing of grid points is also not clear-cut.

In the general regression setting, it is also unclear what the endpoints should be if we use a truncated range of positive values to estimate hyperparameters from restricted marginal maximum likelihood. Recently, for

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

sparse normal means estimation (i.e.  $\mathbf{X} = \mathbf{I}$ ,  $p = n$ , and  $\sigma^2 = 1$  in (1.1)), van der Pas et al. (2017) advocated using the restricted MML estimator for the sparsity parameter  $\tau$  in the range  $[1/n, 1]$  for the horseshoe prior (Carvalho et al. (2010)). This choice allows the horseshoe model to obtain the (near) minimax posterior contraction rate for multivariate normal means. While this choice gives theoretical guarantees for normal means estimation, it does not seem to be justified for high-dimensional regression (1.1) when  $p \gg n$ . Theorem 3.1 in Song and Liang (2017) shows that the minimax optimal choice for  $\tau$  in the horseshoe under model (1.1) satisfies  $\tau \lesssim (\sqrt{s \log p/n}) p^{-(1+(u+1)/(r-1))}$ , where  $u > 0$ ,  $r > 1$ , and  $s = \|\beta_0\|_0$ . It would thus appear that any  $\tau \in [1/n, 1]$  would lead to a *suboptimal* contraction rate for sparse high-dimensional regression. In our numerical experiments in Section 6, we demonstrate that for the horseshoe prior, endowing  $\tau$  with a  $C^+(0, 1)$  prior fares better than the truncation suggested by van der Pas et al. (2017) under the general linear regression model (1.1).

The self-adaptive NBP prior circumvents these issues by obtaining the MML estimates of  $(a, b)$  over the range  $[0, \infty) \times [0, \infty)$ , while ensuring that  $(a, b)$  are never estimated as zero. Thus, the self-adaptive NBP's automatic selection of hyperparameters provides a practical alternative to hierarchical Bayes or cross-validation approaches for tuning hyperparameters.

## 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

### 4.1 Illustration of the Self-Adaptive NBP Model

To illustrate the self-adaptive NBP prior’s ability to adapt to differing sparsity patterns, we consider two settings: one sparse ( $n = 60$ ,  $p = 100$ , 10 nonzero covariates) and one dense ( $n = 60$ ,  $p = 100$ , and 60 nonzero covariates), where the active covariates are drawn from  $\mathcal{U}([-2, -0.5] \cup [0.5, 2])$ . Our examples come from experiments 1 and 4 in Section 6. We initialize  $(a^{(0)}, b^{(0)}) = (0.01, 0.01)$  and implement the Monte Carlo EM algorithm (described in Section 5.1) for finding the MML estimates of the parameters  $(a, b)$ , which we denote as  $(\hat{a}, \hat{b})$ .

In Figure 2, we plot the iterations from two runs of the EM algorithm. The algorithm terminates at iteration  $k$  when the square of the  $\ell_2$  distance between  $(a^{(k-1)}, b^{(k-1)})$  and  $(a^{(k)}, b^{(k)})$  reaches below  $10^{-6}$ . We then set  $(\hat{a}, \hat{b}) = (a^{(k)}, b^{(k)})$ . The top panel in Figure 2 plots the paths for  $a$  and  $b$  from the sparse model, and the bottom panel plots the paths for  $a$  and  $b$  from the dense model. The final MML estimates of  $a$  are  $\hat{a} = 0.184$  for the sparse model and  $\hat{a} = 1.104$  for the dense model.

Figure 3 shows the NBP’s marginal density,  $\pi(\beta|\hat{a}, \hat{b}, \sigma^2)$ , for a single coefficient  $\beta$  using the MML estimates of  $(a, b)$  obtained in sparse and the dense settings respectively. The left panel depicts the marginal density under the sparse setting (10 active predictors,  $(\hat{a}, \hat{b}) = (0.184, 1.124)$ ). Here,

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

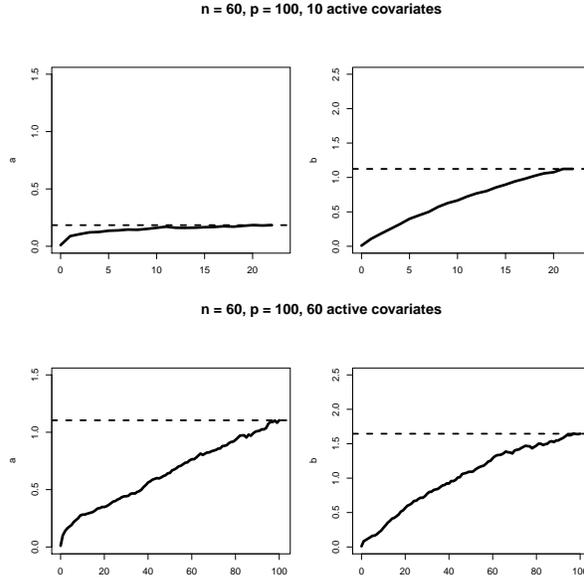


Figure 2: Paths of the Monte Carlo EM algorithm for obtaining  $(\hat{a}, \hat{b})$ . The dashed line indicates the final MML estimate at convergence.

the marginal density for  $\beta$  contains a singularity at zero, *and* most of the probability mass is around zero. We thus recover a sparse model for  $\pi(\beta|\mathbf{y})$  under these MML hyperparameters. Meanwhile, the right panel depicts the marginal density in the dense setting (60 active predictors,  $(\hat{a}, \hat{b}) = (1.104, 1.645)$ ). Here, the marginal density for  $\beta$  does *not* contain a pole, and more mass is placed in neighborhoods away from zero. Thus, we recover a more dense model. Figures 2 and 3 illustrate that in both cases, the EM algorithm was able to correctly learn the true sparsity (or non-sparsity) from the data and incorporate this into its estimates of the hyperparameters.

#### 4. EMPIRICAL BAYES ESTIMATION OF HYPERPARAMETERS

---

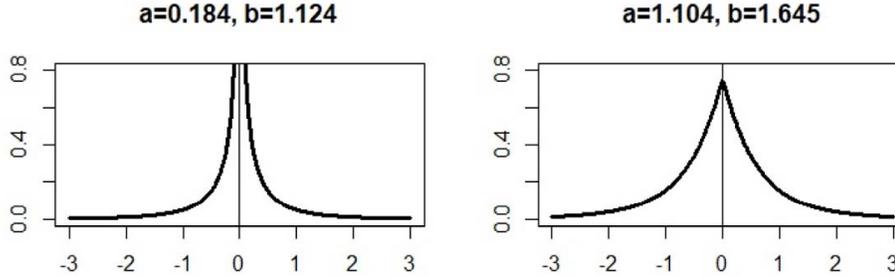


Figure 3: The marginal densities of the NBP prior,  $\pi(\beta|a, b, \sigma^2)$ , with different MML estimates of  $(a, b)$ .

A referee has pointed out that placing a mixture prior of beta prime densities as the prior for  $\omega_i^2$ 's in (1.2) could also accommodate dense situations. While we recognize this fact, we believe that it is better to use marginal maximum likelihood (MML). First, putting a mixture of beta primes as the prior on  $\omega_i^2, i = 1, \dots, p$ , would make the posteriors for  $\beta_i, i = 1, \dots, p$ , multimodal. The quality of our posterior approximation algorithms in Section 5 is dependent on the assumption that the approximate posterior is unimodal (especially if we use a variational density to approximate  $\pi(\boldsymbol{\beta}|\mathbf{y})$ ). Second, if we used a mixture prior, we would then need to tune both the mixture weight(s) and the hyperparameters in each mixture component. As we demonstrate in Sections 4.1 and 6, utilizing a single beta prime prior as the scale with MML estimates for hyperparameters performs quite well.

## 5. Computation for the NBP Model

### 5.1 Posterior Approximation

Using the reparametrization (4.1), the NBP model admits fully closed form conditional densities for the parameters  $(\boldsymbol{\beta}, \lambda_1^2, \dots, \lambda_p^2, \xi_1^2, \dots, \xi_p^2, \sigma^2)$ . Thus, the NBP model can be implemented using either Markov chain Monte Carlo or mean field variational Bayes. Meanwhile, the EM algorithm of Section 4 is easily embedded into either MCMC or MFVB updates, thus negating the need to estimate hyperparameters  $(a, b)$  separately. The complete details for these algorithms are given in the Supplementary Material.

Both the Monte Carlo EM and variational EM algorithms for the self-adaptive NBP model are implemented in the R package, `NormalBetaPrime`. In our experience, the Monte Carlo EM algorithm tends to be slower than the variational EM algorithm, but Monte Carlo EM is more accurate. The Monte Carlo EM algorithm is also relatively immune to the initialization of parameters, whereas the variational EM algorithm is very sensitive to this. This is not a problem with our model, but an inherent shortcoming of MFVB; since MFVB optimizes a highly non-convex objective function over  $O(p^2)$  parameters, it can become “trapped” at a suboptimal local solution. We leave the issues of deriving more efficient sampling algorithms and more

accurate variational algorithms for the NBP model as problems for future research.

## 5.2 Variable Selection

Since the NBP model assigns zero mass to exactly sparse vectors, selection must be performed using some posthoc method. We propose using the “decoupled shrinkage and selection” (DSS) method proposed by Hahn and Carvalho (2015). Letting  $\widehat{\boldsymbol{\beta}}$  denote the posterior mean of  $\boldsymbol{\beta}$ , the DSS method performs variable selection by finding the “nearest” exactly sparse vector to  $\widehat{\boldsymbol{\beta}}$ . DSS solves the optimization,

$$\widehat{\boldsymbol{\gamma}} = \arg \min_{\boldsymbol{\gamma}} n^{-1} \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}\| + \lambda \|\boldsymbol{\gamma}\|_0, \quad (5.1)$$

and chooses the nonzero entries in  $\widehat{\boldsymbol{\gamma}}$  as the active set. Since (5.1) is an NP-hard combinatorial problem, Hahn and Carvalho (2015) propose using local linear approximation, i.e. solving the following surrogate optimization:

$$\widehat{\boldsymbol{\gamma}} = \arg \min_{\boldsymbol{\gamma}} n^{-1} \|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}\| + \lambda \sum_{i=1}^p \frac{|\gamma_i|}{|\widehat{\beta}_i|}, \quad (5.2)$$

where  $\widehat{\beta}_i$ 's are the components in the posterior mean  $\widehat{\boldsymbol{\beta}}$ , and  $\lambda$  is chosen through 10-fold cross-validation to minimize the mean squared error (MSE). Solving this optimization is not computationally expensive, because (5.2) is essentially an adaptive LASSO regression (Zou (2006)) with weights

$1/|\widehat{\beta}_i|, i = 1, \dots, p$ , and there exist very efficient gradient descent algorithms to find LASSO solutions, e.g. Friedman et al. (2010). We use the R package `glmnet`, developed by Friedman et al. (2010), to solve (5.2). We select the nonzero entries in  $\widehat{\gamma}$  from (5.2) as the active set of covariates. The DSS method is available for the NBP prior in the R package, `NormalBetaPrime`.

## 6. Simulation Studies

For our simulation studies, we implement the self-adaptive NBP model (2.2) for model (1.1) using the Monte Carlo EM algorithm described in Section 5. We set  $c = d = 10^{-5}$  in the  $\mathcal{IG}(c, d)$  prior on  $\sigma^2$ . We run the Gibbs samplers for 15,000 iterations, discarding the first 10,000 as burn-in. We use the posterior median estimator  $\widehat{\beta}$  as our point estimator and deploy the DSS strategy described in Section 5.2 for variable selection.

### 6.1 Adaptivity to Different Sparsity Levels

In the first simulation study, we evaluate the self-adaptive NBP model's performance under a variety of sparsity levels. Under model (1.1), we generate a design matrix  $\mathbf{X}$  where the  $n$  rows are independently drawn from  $\mathcal{N}_p(\mathbf{0}, \mathbf{\Gamma})$ ,  $\mathbf{\Gamma} = (\Gamma_{ij})_{p \times p}$  with  $\Gamma_{ij} = 0.5^{|i-j|}$ , and then centered and scaled. The nonzero predictors in  $\beta_0$  are generated from  $\mathcal{U}([-2, -0.5] \cup [0.5, 2])$ .

## 6. SIMULATION STUDIES

---

We fix  $\sigma^2 = 2$  and set  $n = 60, p = 100$ , with varying levels of sparsity:

- Experiment 1: 10 active predictors (sparse model)
- Experiment 2: 20 active predictors (fairly sparse model)
- Experiment 3: 40 active predictors (fairly dense model)
- Experiment 4: 60 active predictors (dense model)

We compare the self-adaptive NBP prior with several other popular Bayesian and frequentist methods. For the competing Bayesian methods, we use the horseshoe (Carvalho et al. (2010)) and the spike-and-slab LASSO (SSL) (Ročková and George (2018)). For the horseshoe, we consider two ways of tuning the global shrinkage parameter  $\tau$ : 1) endowing  $\tau$  with a standard half-Cauchy prior  $\mathcal{C}^+(0, 1)$ , and 2) estimating  $\tau$  from restricted marginal maximum likelihood on the interval  $[1/n, 1]$ , as advocated by van der Pas et al. (2017). These methods are denoted as HS-HC and HS-REML respectively. For the SSL model, the beta prior on the mixture weight  $\theta$  controls the sparsity of the model. We consider two scenarios: 1) endowing  $\theta$  with a  $\mathcal{B}(1, p)$  prior, which induces strong sparsity, and 2) endowing  $\theta$  with a  $\mathcal{B}(1, 1)$  prior, which does not strongly favor sparsity. Finally, we consider the following frequentist methods: minimax concave penalty (MCP) (Zhang (2010)), smoothly clipped absolute deviation (SCAD) (Fan and Li

## 6. SIMULATION STUDIES

---

(2001)), and elastic net (ENet) (Zou and Hastie (2005)). These methods are available in the R packages: `horseshoe`, `SSLASSO`, `picasso`, and `glmnet`.

For each of our methods, we compute the mean squared error (MSE), false discovery rate (FDR), false negative rate (FNR), and overall misclassification probability (MP) averaged across 100 replications:

$$\begin{aligned} \text{MSE} &= \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2/p, & \text{FDR} &= \text{FP} / (\text{TP} + \text{FP}), \\ \text{FNR} &= \text{FN} / (\text{TN} + \text{FN}), & \text{MP} &= (\text{FP} + \text{FN})/p, \end{aligned}$$

where FP, TP, FN, and TN denote the number of false positives, true positives, false negatives, and true negatives respectively.

Tables 1 and 2 in the Supplementary Material show our results averaged across 100 replications for the NBP, HS-HC, HS-REML,  $\text{SSL-}\mathcal{B}(1, p)$ ,  $\text{SSL-}\mathcal{B}(1, 1)$ , MCP, SCAD, and ENet methods. Across all of the sparsity settings, the NBP has the lowest mean squared error, showing that it performs consistently well for estimation. In Experiments 2, 3, and 4, the NBP model also achieves either the lowest or the second lowest misclassification probability, demonstrating that it is also robust for variable selection.

The HS, SSL, MCP, and SCAD methods all perform worse as the true model becomes more dense. The truncation of  $\tau \in [1/n, 1]$  in the HS-REML

---

For the HS-REML method, we slightly modified the code in the `horseshoe` function in the `horseshoe` R package.

model lowers the FDR for the horseshoe, but this also tends to overshrink large signals, leading to greater estimation error than the HS-HC model. For the SSL model, endowing the sparsity parameter  $\theta$  with a  $\mathcal{B}(1, 1)$  prior improves the SSL's performance under dense settings, but not enough to be competitive with the NBP. Finally, the ENet performs the worst under sparsity, but its performance improves as the model becomes more dense. However, the NBP still outperforms the ENet in terms of estimation.

## 6.2 More Numerical Experiments with Large $p$

In the following experiments, the design matrix  $\mathbf{X}$  is generated the same way that it was in Section 6.1. The active predictors are randomly selected and fixed at a certain level, and the remaining covariates are set to zero.

- Experiment 5: ultra-sparse model with a few large signals ( $n = 100, p = 500$ , 8 active predictors set equal to 5)
- Experiment 6: dense model with many small signals ( $n = 200, p = 400$ , 200 active predictors set equal to 0.6)

We implement Experiments 5 and 6 for the self-adaptive NBP, HS-HC, HS-REML,  $\text{SSL-}\mathcal{B}(1, p)$ ,  $\text{SSL-}\mathcal{B}(1, 1)$ , MCP, SCAD, and ENet models. Table 3 in the Supplementary Material shows our results averaged across

## 6. SIMULATION STUDIES

---

100 replications. In Experiment 5, the NBP, HS, and SSL models all significantly outperform their frequentist competitors, with the HS and SSL performing slightly better than NBP. In Experiment 5, the NBP model gives 0 for FDR, FNR, and MP, which illustrates that the self-adaptive NBP is resilient against overfitting if the true model is indeed very sparse. In Experiment 6, the NBP model gives the lowest MSE and lowest MP of all the methods, demonstrating that the self-adaptive NBP model can also effectively adapt to non-sparse situations.

It seems as though the horseshoe, spike-and-slab lasso, MCP, and SCAD are well-suited for sparse estimation but cannot accommodate non-sparse situations as well. Meanwhile, the elastic net seems to be a suboptimal estimator under sparsity (e.g., in Experiment 5, its misclassification rate was 0.104, much higher than the other methods), but it greatly improves in dense settings.

In contrast, the self-adaptive NBP prior is the most robust estimator across *all* the different sparsity patterns. If the true model is sparse, the sparsity parameter  $a$  will be estimated to be small and hence place heavier mass around zero. But if the true model is dense, the sparsity parameter  $a$  will be large, so the singularity at zero disappears and the prior becomes more diffuse.

## 7. Analysis of a Gene Expression Data Set

We analyze a real data set from a study on Bardet-Biedl syndrome (BBS) (Scheetz et al. (2006)), an autosomal recessive disorder which leads to progressive vision loss and which is caused by a mutation in the TRIM32 gene. This data set, which is available in the R package `flare`, contains  $n = 120$  samples with TRIM32 as the response variable and the expression levels of  $p = 200$  other genes as the covariates.

To determine TRIM32's association with these other genes, we implement the self-adaptive NBP, HS-HC, HS-REML, SSL- $\mathcal{B}(1, p)$ , SSL- $\mathcal{B}(1, 1)$ , MCP, SCAD, and ENet models on this data set after centering and scaling  $\mathbf{X}$  and  $\mathbf{y}$ . To assess these methods' predictive performance, we perform five-fold cross validation, using 80 percent of the data as our training set to obtain an estimate of  $\boldsymbol{\beta}$ ,  $\hat{\boldsymbol{\beta}}_{\text{train}}$ . We then use  $\hat{\boldsymbol{\beta}}_{\text{train}}$  to compute the mean squared error of the residuals on the remaining 20 percent of left-out data. We repeat this five times, using different training and test sets each time, and take the average MSE as our mean squared prediction error (MSPE).

Table 4 in the Supplementary Material shows the results of our analysis. The NBP and ENet models give the best predictive performance of all the methods, with 31 genes and 26 genes selected as significantly associated with TRIM32, respectively. The ENet has slightly lower MSPE, but the

## 8. CONCLUDING REMARKS AND FUTURE WORK

---

NBP model’s performance is very similar to the ENet’s. The HS, SSL, MCP, and SCAD methods result in parsimonious models, with 6 or fewer genes selected, but their average prediction errors are all higher.

Figure 4 plots the posterior medians and 95 percent posterior credible intervals for the 31 genes that the NBP model selected as significant. Figure 4 shows that the self-adaptive NBP prior is able to detect small gene expression values that are very close to zero. On this particular data set, the slightly more dense models had much better prediction performance than the most parsimonious models, suggesting that there may be a number of small signals in our data.

### 8. Concluding Remarks and Future Work

In this paper, we have introduced the normal-beta prime (NBP) model for high-dimensional Bayesian linear regression. We proved that the NBP prior obtains the (near) minimax posterior contraction rate in the asymptotic regime where  $p \gg n$  and the underlying model is sparse. To make our prior self-adaptive in finite samples, we introduced an empirical Bayes approach for estimating the NBP’s hyperparameters based on maximum marginal likelihood (MML). Our MML approach for estimating hyperparameters affords the NBP a great deal of flexibility and adaptivity to different levels

## 8. CONCLUDING REMARKS AND FUTURE WORK

---

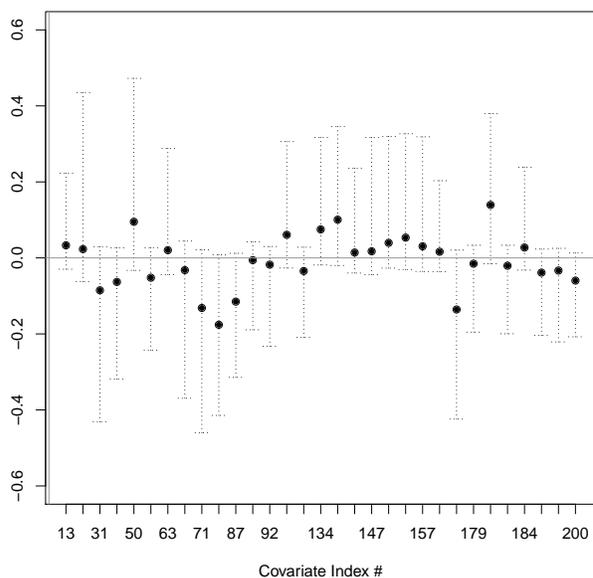


Figure 4: Point estimates and credible intervals for the 31 genes that were selected as significantly associated with TRIM32 by the self-adaptive NBP. of sparsity and different signal strengths, while avoiding degeneracy.

Future work will be to extend the NBP prior to more complex and more flexible models, such as nonparametric regression or semiparametric regression with unknown error distribution. The NBP prior can also be employed for other statistical problems like density estimation or classification. Due to its flexibility, we anticipate that the NBP prior would retain its strong empirical and theoretical properties in these other settings.

Additionally, we would like to provide further theoretical support for the marginal maximum likelihood approach described in Section 4. Al-

## 8. CONCLUDING REMARKS AND FUTURE WORK

---

though there are philosophical reasons for MML (i.e., that it maximizes the “model evidence”), it would be interesting to see if the MML estimates of  $(a, b)$  also lead to (near) minimax posterior contraction under the same conditions as those in Section 3.1. Currently, theoretical justifications for MML under model (1.1) have been confined to either the simple normal means model ( $\mathbf{X} = \mathbf{I}$ ,  $n = p$ ) or the scenario where  $p \leq n$  and the MML estimate can be explicitly calculated in closed form (as is the case for the hyperparameter  $g$  in  $g$ -priors). See, e.g., van der Pas et al. (2017), Johnstone and Silverman (2004), George and Foster (2000), Sparks et al. (2015). Recently, Rousseau and Szabó (2017) extended the class of models for which the posterior contraction rate can be obtained under MML estimates of a hyperparameter in the prior, but unfortunately, their framework does not seem to be applicable to the high-dimensional linear regression model (1.1), which is complicated by the presence of a high-dimensional design matrix  $\mathbf{X}$ . We hope to address the theoretical aspects of the self-adaptive NBP model with MML-estimated hyperparameters in future work.

### Supplementary Material

The Online Supplement contains the tables reporting simulation and data analysis results from Sections 6 and 7, proofs for Theorems 3.1 and 4.1, and technical details for the Monte Carlo EM and variational EM

algorithms from Section 5 for implementing the self-adaptive NBP model.

## Acknowledgements

The authors thank the Editor, the Associate Editor, and two anonymous referees for their helpful and insightful comments.

## References

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized beta mixtures of gaussians.

*Advances in Neural Information Processing Systems 24*, pp. 523-531.

Armagan, A., Clyde, M., and Dunson, D. B. (2013). Generalized double pareto shrinkage.

*Statistica Sinica 23*, pp. 119-143.

Bai, R. and Ghosh, M. (2019). Large-scale multiple hypothesis testing with the normal-beta

prime prior. *ArXiv e-prints, 2019. arXiv:1807.02421*.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian

scale mixture priors in high-dimensional regression. *Biometrika 103*, pp. 985-991.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-laplace priors for

optimal shrinkage. *Journal of the American Statistical Association 110*, pp. 1479-1490.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe.

*Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics,*

*PMLR 5*, pp. 73-80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse

## REFERENCES

---

- signals. *Biometrika* 97, pp. 465-480.
- Casella, G. (2001). Empirical bayes gibbs sampling. *Biostatistics* 2, pp. 485-500.
- Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. Bayesian linear regression with sparse priors. *The Annals of Statistics* 43, pp. 1986-2018.
- Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis* 8, pp. 111-132.
- Dobriban, E. and Fan, J. Regularity properties for sparse regression. *Communications in Mathematics and Statistics* 4, pp. 1-19.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, pp. 1348-1360.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33, pp. 1-22.
- George, E. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika* 87, pp. 731-747.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* 28, pp. 500-531.
- Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5, pp. 171-188.
- Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian

## REFERENCES

---

- linear models: A posterior summary perspective. *Journal of the American Statistical Association* 110, pp. 435-448.
- Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2017). Convergence and approximation theory of a scalable meme algorithm for the horseshoe prior. *ArXiv e-prints*, 2017. *arXiv:1705.00841*.
- Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* 107, pp. 649-660.
- Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical bayes estimates of possibly sparse sequences. *The Annals of Statistics* 32, pp. 1594-1649.
- Leday, G. G. R., de Gunst, M. C. M., Kpogbezan, G. B., van der Vaart, A. W., van Wieringen, W. N., and van de Wiel, M. A. (2017). Gene network reconstruction using global-local shrinkage priors. *The Annals of Applied Statistics* 11, pp. 41-68.
- Martin, R., Mess, R., and Walker, S. G. (2017). Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* 23, pp. 1822-1847.
- Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* 42, pp. 789-817.
- Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* 103, pp. 681-686.
- Pérez, M.-E., Pericchi, L. R., and Ramírez, I. C. (2017). The scaled beta2 distribution as a

## REFERENCES

---

- robust prior for scales. *Bayesian Analysis* 12, pp. 615-637.
- Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics 9*, pp. 501-538.
- Polson, N. G. and Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis* 7, pp. 887-902.
- Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls. *IEEE Transactions on Information Theory* 57, pp. 6976-6994.
- Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association* 113, pp. 431-444.
- Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association* 112, pp. 254-265.
- Rousseau, J. and Szabó, B. (2017). Asymptotic behavior of the empirical bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics* 45, pp. 833-865.
- Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Kudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* 103, pp. 14429-14434.

## REFERENCES

---

- Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* 38, pp. 2587-2619.
- Shin, M., Bhattacharya, A., and Johnson, V. E. (2018). Scalable bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statistica Sinica* 28, pp. 1053-1078.
- Song, Q. and Liang, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *ArXiv e-prints, 2017. arXiv:1712.08964*.
- Sparks, D. K., Khare, K., and Ghosh, M. (2015). Necessary and sufficient conditions for high-dimensional posterior consistency under  $g$ -priors. *Bayesian Analysis* 10, pp. 627-664.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, pp. 583-639.
- Tiao, G. C. and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance-components. *Biometrika* 52, pp. 37-54.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58, pp. 267-288.
- van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics* 10, pp. 976-1000.

## REFERENCES

---

van der Pas, S., Szabó, B., and van der Vaart, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics* 11, pp. 3196-3225.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics* 44, pp. 2497-2532.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38, pp. 894-942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101, pp. 1418-1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, pp. 301-320.

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania

E-mail: Ray.Bai@penncmedicine.upenn.edu

Department of Statistics, University of Florida

E-mail: ghoshm@ufl.edu