

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

Ray Bai and Malay Ghosh

University of South Carolina and University of Florida

Supplementary Material

In Section S1, we provide the tables reporting the results from our simulation studies in Section 6 and our data analysis in Section 7 of the main manuscript. In Section S2, we provide proofs for Theorems 3.1 and 4.1 of the main manuscript. In Section S3, we provide technical details for the Monte Carlo EM and variational EM algorithms described in Section 5 of the main manuscript.

S1 Results for Simulations and Data Analysis

Table 1: Simulation results for Experiments 1 and 2 for the NBP, HS-HC, HS-REML, $\text{SSL-}\mathcal{B}(1, p)$, $\text{SSL-}\mathcal{B}(1, 1)$, MCP, SCAD, and ENet models, averaged across 100 replications when $n = 60, p = 100$.

| Experiment 1: sparse model (10 active predictors) | | | | |
|----------------------------------------------------------|--------------|--------------|--------------|--------------|
| Method | MSE | FDR | FNR | MP |
| NBP | 0.019 | 0.214 | 0.011 | 0.039 |
| HS-HC | 0.020 | 0.128 | 0.014 | 0.029 |
| HS-REML | 0.021 | 0.023 | 0.023 | 0.023 |
| $\text{SSL-}\mathcal{B}(1, p)$ | 0.020 | 0.066 | 0.019 | 0.026 |
| $\text{SSL-}\mathcal{B}(1, 1)$ | 0.025 | 0.151 | 0.017 | 0.036 |
| MCP | 0.020 | 0.238 | 0.014 | 0.046 |
| SCAD | 0.028 | 0 | 0.1 | 0.1 |
| ENet | 0.037 | 0.730 | 0.006 | 0.284 |
| Experiment 2: fairly sparse model (20 active predictors) | | | | |
| Method | MSE | FDR | FNR | MP |
| NBP | 0.077 | 0.202 | 0.050 | 0.083 |
| HS-HC | 0.110 | 0.235 | 0.084 | 0.11 |
| HS-REML | 0.286 | 0.130 | 0.115 | 0.119 |
| $\text{SSL-}\mathcal{B}(1, p)$ | 0.090 | 0.175 | 0.053 | 0.078 |
| $\text{SSL-}\mathcal{B}(1, 1)$ | 0.090 | 0.222 | 0.048 | 0.086 |
| MCP | 0.238 | 0.321 | 0.091 | 0.142 |
| SCAD | 0.226 | 0.791 | 0.199 | 0.252 |
| ENet | 0.096 | 0.610 | 0.031 | 0.310 |

S1. RESULTS FOR SIMULATIONS AND DATA ANALYSIS

Table 2: Simulation results for Experiments 3 and 4 for the NBP, HS-HC, HS-REML, $\text{SSL-}\mathcal{B}(1, p)$, $\text{SSL-}\mathcal{B}(1, 1)$, MCP, SCAD, and ENet models, averaged across 100 replications when $n = 60, p = 100$.

| Experiment 3: fairly dense model (40 active predictors) | | | | |
|---------------------------------------------------------|--------------|--------------|--------------|--------------|
| Method | MSE | FDR | FNR | MP |
| NBP | 0.448 | 0.251 | 0.240 | 0.246 |
| HS-HC | 0.535 | 0.243 | 0.256 | 0.254 |
| HS-REML | 1.10 | 0.233 | 0.338 | 0.325 |
| $\text{SSL-}\mathcal{B}(1, p)$ | 0.728 | 0.300 | 0.270 | 0.279 |
| $\text{SSL-}\mathcal{B}(1, 1)$ | 0.665 | 0.308 | 0.260 | 0.276 |
| MCP | 1.31 | 0.298 | 0.343 | 0.344 |
| SCAD | 1.21 | 0.604 | 0.401 | 0.440 |
| ENet | 0.453 | 0.423 | 0.198 | 0.320 |

| Experiment 4: dense model (60 active predictors) | | | | |
|--------------------------------------------------|--------------|--------------|--------------|--------------|
| Method | MSE | FDR | FNR | MP |
| NBP | 0.760 | 0.173 | 0.467 | 0.344 |
| HS-HC | 1.10 | 0.184 | 0.495 | 0.395 |
| HS-REML | 1.76 | 0.149 | 0.552 | 0.489 |
| $\text{SSL-}\mathcal{B}(1, p)$ | 1.53 | 0.223 | 0.506 | 0.409 |
| $\text{SSL-}\mathcal{B}(1, 1)$ | 1.40 | 0.226 | 0.495 | 0.395 |
| MCP | 1.31 | 0.298 | 0.343 | 0.359 |
| SCAD | 2.18 | 0.430 | 0.603 | 0.589 |
| ENet | 0.892 | 0.260 | 0.426 | 0.336 |

Table 3: Simulation results for Experiments 5 and 6 for NBP, HS-HC, HS-REML, SSL- $\mathcal{B}(1, p)$, SSL- $\mathcal{B}(1, 1)$, MCP, SCAD, and ENet models, averaged across 100 replications.

Experiment 5: $n = 100$, $p = 500$, 8 active predictors set equal to 5.

| Method | MSE | FDR | FNR | MP |
|--------------------------|---------------|----------|----------|----------|
| NBP | 0.0007 | 0 | 0 | 0 |
| HS-HC | 0.0005 | 0 | 0 | 0 |
| HS-REML | 0.0005 | 0 | 0 | 0 |
| SSL- $\mathcal{B}(1, p)$ | 0.0005 | 0.037 | 0 | 0.0007 |
| SSL- $\mathcal{B}(1, 1)$ | 0.0008 | 0.089 | 0 | 0.0017 |
| MCP | 0.078 | 0.124 | 0.0012 | 0.011 |
| SCAD | 0.081 | 0.984 | 0.016 | 0.031 |
| ENet | 0.067 | 0.859 | 0 | 0.104 |

Experiment 6: $n = 200$, $p = 400$, 200 active predictors set equal to 0.6

| Method | MSE | FDR | FNR | MP |
|--------------------------|--------------|--------------|-------|--------------|
| NBP | 0.031 | 0.273 | 0.400 | 0.351 |
| HS-HC | 0.041 | 0.261 | 0.423 | 0.384 |
| HS-REML | 0.049 | 0.204 | 0.469 | 0.444 |
| SSL- $\mathcal{B}(1, p)$ | 0.095 | 0.311 | 0.462 | 0.437 |
| SSL- $\mathcal{B}(1, 1)$ | 0.093 | 0.334 | 0.458 | 0.433 |
| MCP | 0.058 | 0.213 | 0.479 | 0.462 |
| SCAD | 0.051 | 0.488 | 0.499 | 0.498 |
| ENet | 0.038 | 0.346 | 0.362 | 0.355 |

Table 4: Results for data analysis of the Bardet-Biedl syndrome (BBS) data set.

| Method | Number of Genes Selected | MSPE |
|--------------------------|--------------------------|--------------|
| NBP | 31 | 0.466 |
| HS-HC | 6 | 0.797 |
| HS-REML | 4 | 0.616 |
| SSL- $\mathcal{B}(1, p)$ | 3 | 0.594 |
| SSL- $\mathcal{B}(1, 1)$ | 3 | 0.504 |
| MCP | 5 | 0.582 |
| SCAD | 5 | 0.603 |
| ENet | 26 | 0.462 |

S2 Proofs of Main Theorems

Before proving Theorem 3.1, we restate the main results on posterior consistency from Song and Liang (2017). Proposition S2.1 is a restatement of Theorems A.1 and A.2 in Song and Liang (2017).

Proposition S2.1. *Consider the linear regression model (3.1) and suppose that conditions (A1)-(A5) hold. Suppose that the prior for $\pi(\boldsymbol{\beta}, \sigma^2)$ is of the form,*

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{i=1}^p [g(\beta_i/\sigma)/\sigma], \quad \sigma^2 \sim \mathcal{IG}(c, d). \quad (\text{S2.1})$$

Suppose $r_n = M\sqrt{s_n \log p_n/n}$, where $M > 0$ is sufficiently large. If the

density $g(\cdot)$ in (S2.1) satisfies

$$\begin{aligned} 1 - \int_{-k_n}^{k_n} g(x)dx &\leq p_n^{-(1+u)}, \\ -\log \left(\inf_{x \in [-E_n, E_n]} g(x) \right) &= O(\log p_n), \end{aligned} \quad (\text{S2.2})$$

where $u > 0$ is a constant and $k_n \asymp \sqrt{s_n \log p_n / n} / p_n$, then the following results hold:

$$\begin{aligned} \Pr_{\beta_0} \left(\Pi(\beta : \|\beta - \beta_0\|_2 \geq c_1 \sigma_0 r_n | \mathbf{y}_n) \geq e^{-c_2 n r_n^2} \right) &\leq e^{-c_3 n r_n^2}, \\ \Pr_{\beta_0} \left(\Pi(\beta : \|\beta - \beta_0\|_1 \geq c_1 \sigma_0 \sqrt{s_n} r_n | \mathbf{y}_n) \geq e^{-c_2 n r_n^2} \right) &\leq e^{-c_3 n r_n^2}, \\ \Pr_{\beta_0} \left(\Pi(\beta : \|\mathbf{X}_n \beta - \mathbf{X}_n \beta_0\|_2 \geq c_0 \sigma_0 \sqrt{n} r_n | \mathbf{y}_n) < 1 - e^{-c_2 n r_n^2} \right) &\leq e^{-c_3 n r_n^2}, \\ \Pr_{\beta_0} \left(\Pi(\beta : \text{at least } \tilde{q}_n \text{ entries of } |\beta / \sigma| \text{ are larger than } k_n | \mathbf{y}_n) > e^{-c_2 n r_n^2} \right) \\ &\leq e^{-c_3 n r_n^2}, \end{aligned}$$

for some constants $c_0, c_1, c_2, c_3 > 0$, and $\tilde{q}_n \asymp s_n$.

Before proving Theorem 3.1, we also prove the following two lemmas.

Lemma S2.1. *Suppose that $a_n \rightarrow 0$ as $n \rightarrow \infty$ and $b \in (1, \infty)$ as $n \rightarrow \infty$.*

Then

$$\frac{\Gamma(a_n + b)}{\Gamma(a_n)\Gamma(b)} \asymp a_n. \quad (\text{S2.3})$$

Proof of Lemma S2.1. Rewrite (S2.3) as

$$\begin{aligned} \frac{\Gamma(a_n + b)}{\Gamma(a_n)\Gamma(b)} &= \frac{a_n \Gamma(a_n + b + 1)}{(a_n + b)\Gamma(a_n + 1)\Gamma(b)} \\ &= \frac{a_n}{a_n + b} \left(\frac{1}{\int_0^1 u^{a_n} (1-u)^{b-1} du} \right). \end{aligned} \quad (\text{S2.4})$$

We have the following inequalities:

$$\int_0^1 u^{a_n}(1-u)^{b-1} du \leq \int_0^1 (1-u)^{b-1} du = b^{-1}, \quad (\text{S2.5})$$

and

$$\begin{aligned} \int_0^1 u^{a_n}(1-u)^{b-1} du &\geq \int_{1/2}^1 u^{a_n}(1-u)^{b-1} du \\ &\geq 2^{-a_n} \int_{1/2}^1 (1-u)^{b-1} du \\ &= 2^{-a_n} 2^{-b} b^{-1}. \end{aligned} \quad (\text{S2.6})$$

Thus, from (S2.4)-(S2.6), we have

$$\frac{a_n b}{a_n + b} \leq \frac{\Gamma(a_n + b)}{\Gamma(a_n)\Gamma(b)} \leq \frac{a_n 2^{a_n + b} b}{a_n + b}. \quad (\text{S2.7})$$

Since $a_n \rightarrow 0$ as $n \rightarrow \infty$, we have $b/(a_n + b) \sim 1$ and $2^{a_n + b} b/(a_n + b) \sim 2^b$, and thus, from (S2.7), we have $\Gamma(a_n + b)/\Gamma(a_n)\Gamma(b) \asymp a_n$ as $n \rightarrow \infty$. \square

Lemma S2.2. *Let $b > 1$. Then for any $a > 0$, $\beta'(a, b)$ is stochastically dominated by $\beta'(a, 1)$.*

Proof of Lemma S2.2. Let $f(x|a, b)$ denote the probability density function (pdf) for the beta prime density, $\beta'(a, b)$. We have

$$\frac{f(x|a, 1)}{f(x|a, b)} \propto \frac{x^{a-1}(1+x)^{-a-1}}{x^{a-1}(1+x)^{-a-b}} = (1+x)^{b-1},$$

which is increasing in x due to our assumption that $b > 1$. Hence, by the monotone likelihood ratio property, $\beta'(a, b)$ is stochastically dominated by $\beta'(a, 1)$ for any $b > 1$. \square

Proof of Theorem 3.1. By Proposition S2.1, it is sufficient to verify that the NBP prior for each coefficient $\pi(\beta_i), i = 1, \dots, p_n$, satisfies the two conditions (S2.2). We first verify the first condition. Let $g(\cdot)$ be the marginal pdf of $\pi(\beta)$ for a single coefficient β . The pdf $g(x)$ under the NBP prior is

$$g(x) = \frac{\Gamma(a_n + b)}{(2\pi)^{1/2}\Gamma(a_n)\Gamma(b)} \int_0^\infty \exp\left(-\frac{x^2}{2\omega^2}\right) (\omega^2)^{a_n-3/2}(1 + \omega^2)^{-a_n-b} d\omega^2. \quad (\text{S2.8})$$

By the symmetry of $g(x)$ and Fubini's Theorem, we have from (S2.8) that

$$\begin{aligned} 1 - \int_{-k_n}^{k_n} g(x) dx &= 2 \int_{k_n}^\infty g(x) dx \\ &= \frac{2\Gamma(a_n + b)}{(2\pi)^{1/2}\Gamma(a_n)\Gamma(b)} \int_{k_n}^\infty \int_0^\infty \exp\left(-\frac{x^2}{2\omega^2}\right) (\omega^2)^{a_n-3/2}(1 + \omega^2)^{-a_n-b} d\omega^2 dx \\ &= \frac{\Gamma(a_n + b)}{\Gamma(a_n)\Gamma(b)} \int_0^\infty (\omega^2)^{a_n-1}(1 + \omega^2)^{-a_n-b} \left[2 \int_{k_n}^\infty (2\pi\omega^2)^{-1/2} \exp\left(-\frac{x^2}{2\omega^2}\right) dx \right] d\omega^2 \end{aligned} \quad (\text{S2.9})$$

Letting $X \sim \mathcal{N}(0, \omega^2)$, we see the inner integral in (S2.9) is $\Pr(|X| \geq k_n)$.

We use the tail bound, $\Pr(|X| \geq k_n) \leq 2e^{-k_n^2/2\omega^2}$, to further bound (S2.9)

from above as

$$\begin{aligned}
2 \int_{k_n}^{\infty} g(x) dx &\leq \frac{2\Gamma(a_n + b)}{\Gamma(a_n)\Gamma(b)} \int_0^{\infty} (\omega^2)^{a_n-1} (1 + \omega^2)^{-a_n-b} e^{-k_n^2/2\omega^2} d\omega^2 \\
&\leq 2a_n \int_0^{\infty} (\omega^2)^{a_n-1} (1 + \omega^2)^{-a_n-1} e^{-k_n^2/2\omega^2} d\omega^2 \\
&= 2a_n \int_0^{\infty} (1 + u)^{-a_n-1} e^{-u(k_n^2/2)} du \\
&\leq 2a_n \int_0^{\infty} e^{-u(k_n^2/2)} du \\
&= \frac{4a_n}{k_n^2} \\
&\lesssim p_n^{-(1+u)}, \tag{S2.10}
\end{aligned}$$

where we used the fact that $b \in (1, \infty)$ and Lemma S2.2 in the second inequality, a transformation of variables $u = 1/\omega^2$ in the first equality, and the fact that $a_n \lesssim k_n^2 p_n^{-(1+u)}$ for the final inequality of the above display. Thus, combining (S2.9)-(S2.10) shows that the first condition in (S2.2) holds.

We now show that the second condition of (S2.2) also holds under our assumptions on (a_n, b) and our assumption on the rate of growth on E_n in (A5). With a change of variables, $z = x^2/2\omega^2$, in (S2.8), we can rewrite the marginal pdf of the NBP prior, $g(x)$, as

$$g(x) = \frac{\Gamma(a_n + b)}{2^{1-b} \pi^{1/2} \Gamma(a_n) \Gamma(b)} (x^2)^{a_n-1/2} \int_0^{\infty} e^{-z} z^{b-1/2} (x^2 + 2z)^{-a_n-b} dz. \tag{S2.11}$$

By the symmetry of $g(x)$, the infimum of $g(x)$ on the interval $[-E_n, E_n]$

occurs at either $-E_n$ or E_n . From (S2.3) in Lemma S2.1, (S2.11), and the assumptions that E_n is nondecreasing and $b \in (1, \infty)$, we have

$$\begin{aligned}
 \inf_{x \in [-E_n, E_n]} g(x) &\gtrsim a_n (E_n^2)^{a_n-1/2} \int_0^\infty e^{-z} z^{b-1/2} (E_n^2 + 2z)^{-a_n-b} dz \\
 &= a_n (E_n^2)^{a_n-1/2} \int_0^\infty e^{-z} \left(\frac{z}{E_n^2 + 2z} \right)^{b-1/2} \left(\frac{1}{E_n^2 + 2z} \right)^{a_n+1/2} dz \\
 &\geq a_n (E_n^2)^{a_n-1/2} \int_1^2 e^{-z} \left(\frac{z}{E_n^2 + 2z} \right)^{b-1/2} \left(\frac{1}{E_n^2 + 2z} \right)^{a_n+1/2} dz \\
 &\gtrsim a_n (E_n^2)^{a_n-1/2} (E_n^2 + 2)^{-b+1/2} (E_n^2 + 4)^{-a_n-1/2} \\
 &\asymp a_n (E_n^2)^{-b-1/2}. \tag{S2.12}
 \end{aligned}$$

By assumption, $a_n \lesssim k_n^2 p_n^{-(1+u)}$ for some $u > 0$, and $\log(E_n) = O(\log p_n)$.

Therefore, it follows from (S2.12) that

$$\begin{aligned}
 -\log \left(\inf_{x \in [-E_n, E_n]} g(x) \right) &\lesssim -\log(k_n^2 p_n^{-(1+u)}) + (b + 1/2) \log p_n \\
 &\lesssim -\log(p_n^{-(3+u)}) + (b + 1/2) \log p_n \\
 &\lesssim \log p_n, \tag{S2.13}
 \end{aligned}$$

where we used the fact that $k_n \asymp \sqrt{s_n \log p_n / n} / p_n$ and Assumption (A4) that $s_n = o(n / \log p_n)$, and so $k_n \lesssim p_n^{-1}$. Thus, the second condition in (S2.2) also holds.

We have shown that as long as $a_n \lesssim k_n^2 p_n^{-(1+u)}$, $u > 0$, $b \in (1, \infty)$, and $\log(E_n) = O(\log p_n)$ in Assumption (A5), the two conditions (S2.2) in Proposition S2.1 are satisfied. Hence, Theorem 3.1 has been proven. \square

Proof of Theorem 4.1. At the k th iteration of the EM algorithm, the (a, b) that solves (4.4) is

$$\begin{aligned}\psi(a) &= \frac{1}{p} \sum_{i=1}^p U_i(\lambda_i^2), & a \geq 0, \\ \psi(b) &= -\frac{1}{p} \sum_{i=1}^p V_i(\xi_i^2), & b \geq 0,\end{aligned}\tag{S2.14}$$

where $U_i(\lambda_i^2)$ is an estimate of $\mathbb{E}_{a^{(k-1)}} [\log(\lambda_i^2) | \mathbf{y}]$ and $V_i(\xi_i^2)$ is an estimate of $\mathbb{E}_{b^{(k-1)}} [\log(\xi_i^2) | \mathbf{y}]$ taken from either the Gibbs sampler or the MFVB coordinate ascent algorithm. Since the λ_i 's and ξ_i 's, $i = 1, \dots, p$, are strictly greater than zero and are drawn from \mathcal{GIG} and \mathcal{IG} densities in the Gibbs sampling algorithm or the MFVB algorithm (and thus, expectations of $\log(\lambda_i^2)$ and $\log(\xi_i^2)$, $i = 1, \dots, p$, are well-defined and finite), U_i and V_i , $i = 1, \dots, p$, exist and are finite.

The digamma function $\psi(x)$ is continuous and monotonically increasing for all $x \in (0, \infty)$, with a range of $(-\infty, \infty)$ on the domain of positive reals. Therefore, for any $y \in \mathbb{R}$, there exists a unique $x \in (0, \infty)$ so that $\psi(x) = y$. Since we impose the constraint that $a \geq 0$, there must be a unique $\widehat{a}^{(k)} > 0$ that solves the first equation in (S2.14). Similarly, there exists a unique $\widehat{b}^{(k)} > 0$ that solves the second equation in (S2.14). \square

S3 Details for the Monte Carlo EM and Variational EM Algorithms for the Self-Adaptive NBP Model

S3.1 Monte Carlo EM Algorithm

After initializing $(\boldsymbol{\beta}, \lambda_1, \dots, \lambda_p, \xi_1, \dots, \xi_p, \sigma^2)$, we iteratively cycle through sampling from the full conditional densities in (5.1). To speed up computation, the λ_i 's and ξ_i 's, $i = 1, \dots, p$, are block-updated in parallel, and we utilize the fast sampling algorithm of Bhattacharya et al. (2016) to sample from the full conditional for $\boldsymbol{\beta}$ in $O(n^2p)$ time.

As described in Section 5.1, we incorporate the EM algorithm for obtaining the MML estimates of (a, b) by solving for (a, b) in (4.4) every $M = 100$ iterations of the Gibbs sampler. To assess convergence, we compute the square of the Euclidean distance between $(\widehat{a}^{(k-1)}, \widehat{b}^{(k-1)})$ and $(\widehat{a}^{(k)}, \widehat{b}^{(k)})$ at the k th iteration of the EM Monte Carlo algorithm, and if it falls below a small $\delta > 0$, then we set our MML estimates as $(\widehat{a}, \widehat{b}) = (\widehat{a}^{(k)}, \widehat{b}^{(k)})$ and draw a final sample from the Gibbs sampler.

We recommend setting $\delta = 10^{-6}$. If the square of the ℓ_2 distance has not fallen below δ after 100 iterations (so 10,000 total iterations of the Gibbs sampler have been sampled at this point), then we terminate the EM algorithm and use the estimate from the 100th iteration as $(\widehat{a}, \widehat{b})$. In

our experience, even if the square of the ℓ_2 distance between $(\widehat{a}^{(k-1)}, \widehat{b}^{(k-1)})$ and $(\widehat{a}^{(k)}, \widehat{b}^{(k)})$ does not quite fall underneath the small $\delta > 0$ after $k = 100$ updates, the successive iterates are still very close to one another at this point. Thus, all these later estimates of (a, b) would have a similar effect on posterior inference. Algorithm 1 at the end of Section S2 gives the complete steps for implementing the EM/Gibbs algorithm for our model.

S3.2 Variational EM Algorithm

Let $\boldsymbol{\lambda} = (\lambda_1^2, \dots, \lambda_p^2)$ and $\boldsymbol{\xi} = (\xi_1^2, \dots, \xi_p^2)$ from (5.1). The mean field variational Bayes (MFVB) approach stems from the following lower bound:

$$\begin{aligned} \log \pi(\mathbf{y}) &\geq \int_{(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2)} q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2) \log \left(\frac{\pi(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2, \gamma)}{q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2)} \right) d(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2) \\ &\equiv \mathcal{L}[q(\cdot)], \end{aligned} \tag{S3.1}$$

where $\mathcal{L}[q(\cdot)]$ is known as the evidence lower bound (ELBO). We constrain $q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2) = q_1^*(\boldsymbol{\beta})q_2^*(\boldsymbol{\lambda})q_3^*(\boldsymbol{\xi})q_4^*(\sigma^2)$ and the q_i 's, $i = 1, \dots, 4$, to be families that ensure that (S3.1) is tractable. This is also known as mean field variational Bayes (MFVB). The parameters in q_1^* , q_2^* , q_3^* , and q_4^* are found by maximizing (S3.1), which is equivalent to minimizing the Kullback-Leibler (KL) distance between $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2 | \mathbf{y})$ and $q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2)$. $\pi(\boldsymbol{\beta} | \mathbf{y})$ can be approximated by $q_1^*(\boldsymbol{\beta})$ and posterior inference can be carried out through $q_1^*(\boldsymbol{\beta})$. For a detailed review of variational inference, see Blei et al. (2017).

Based on the conditional densities in (5.1), we use the approximation,

$$q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2 | \mathbf{y}) \approx q_1^*(\boldsymbol{\beta})q_2^*(\boldsymbol{\lambda}), q_3^*(\boldsymbol{\xi})q_4^*(\sigma^2),$$

where

$$\begin{aligned} q_1^*(\boldsymbol{\beta}) &\sim \mathcal{N}_p(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*), \\ q_2^*(\boldsymbol{\lambda}) &\sim \prod_{i=1}^p \mathcal{GIG}(k_i^*, l^*, m^*), \\ q_3^*(\boldsymbol{\xi}) &\sim \prod_{i=1}^p \mathcal{IG}(u^*, v_i^*), \\ q_4^*(\sigma^2) &\sim \mathcal{IG}(c^*, d^*), \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\beta}^* &= (\mathbf{X}^\top \mathbf{X} + \mathbf{D}^*)^{-1} \mathbf{X}^\top \mathbf{y}, \quad \boldsymbol{\Sigma}^* = \mathbb{E}_{q_4^*}(\sigma^2) (\mathbf{X}^\top \mathbf{X} + \mathbf{D}^*)^{-1}, \\ \mathbf{D}^* &= \text{diag}(\mathbb{E}_{q_2^*}(\lambda_1^{-2})\mathbb{E}_{q_3^*}(\xi_1^{-2}), \dots, \mathbb{E}_{q_2^*}(\lambda_p^{-2})\mathbb{E}_{q_3^*}(\xi_p^{-2})), \\ k_i &= \mathbb{E}_{q_1^*}(\beta_i^2)\mathbb{E}_{q_4^*}(\sigma^{-2})\mathbb{E}_{q_3^*}(\xi_i^{-2}), \quad i = 1, \dots, p, \quad l^* = 2, \quad m^* = a - \frac{1}{2}, \\ u^* &= b + \frac{1}{2}, \quad v_i^* = \frac{1}{2}\mathbb{E}_{q_1^*}(\beta_i^2)\mathbb{E}_{q_4^*}(\sigma^{-2})\mathbb{E}_{q_2^*}(\lambda_i^{-2}) + 1, \quad i = 1, \dots, p, \\ c^* &= \frac{n+p+2c}{2}, \quad d^* = \frac{\mathbb{E}_{q_1^*}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2) + \mathbb{E}_q(\boldsymbol{\beta}^\top \mathbf{D}^* \boldsymbol{\beta}) + 2d}{2}. \end{aligned} \tag{S3.2}$$

From (5.3) and (S3.2), we can easily construct our coordinate ascent updates. The expectations, $\mathbb{E}_{q_2^*}(\lambda_i^{-2})$, $\mathbb{E}_{q_3^*}(\xi_i^{-2})$, $\mathbb{E}_{q_4^*}(\sigma^2)$, and $\mathbb{E}_{q_4^*}(\sigma^{-2})$ can be computed using properties of the \mathcal{GIG} and \mathcal{IG} densities. We also have

$$\begin{aligned} \mathbb{E}_{q_1^*}(\beta_i^2) &= (\beta_i^*)^2 + \boldsymbol{\Sigma}_{ii}^*, \\ \mathbb{E}_{q_1^*}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2) &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 + \text{tr}(\mathbf{X}^\top \mathbf{X} \boldsymbol{\Sigma}^*), \\ \mathbb{E}_q(\boldsymbol{\beta}^\top \mathbf{D}^* \boldsymbol{\beta}) &= \sum_{i=1}^p (\beta_i^*)^2 \mathbb{E}_{q_2^*}(\lambda_i^{-2}) \mathbb{E}_{q_3^*}(\xi_i^{-2}) + \text{tr}(\mathbf{D}^* \boldsymbol{\Sigma}^*). \end{aligned}$$

At each iteration, we compute the evidence lower bound (ELBO),

$$\mathcal{L} = \mathbb{E}_q \log f(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2) - \mathbb{E}_q \log q(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2), \quad (\text{S3.3})$$

where f is the joint density over \mathbf{y} and all parameters. In particular, (S3.3)

can be derived as

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_q \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) + \mathbb{E}_q \log(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{\xi}, \sigma^2) + \mathbb{E}_q \log \pi(\boldsymbol{\xi}) + \mathbb{E}_q \log \pi(\boldsymbol{\xi}) \\ &\quad + \mathbb{E}_q \log \pi(\sigma^2) - \mathbb{E}_q \log q_1^*(\boldsymbol{\beta}) - \mathbb{E}_q \log q_2^*(\boldsymbol{\lambda}) - \mathbb{E}_q \log q_3^*(\boldsymbol{\xi}) - \mathbb{E}_q \log q_4^*(\sigma^2) \\ &= -\frac{n}{2} \log(2\pi) + \frac{n}{2} + p \log 2 + p \log \Gamma(u^*) - p \log \Gamma(a) - p \log \Gamma(b) \\ &\quad + c \log d - c^* \log d^* + \log \Gamma(c^*) - \log \Gamma(c) + \frac{1}{2} \log |\boldsymbol{\Sigma}^*| \\ &\quad - \sum_{i=1}^p \log \left[\frac{(k_i^*/l^*)^{m^*/2}}{K_{m^*}(\sqrt{k_i^* l^*})} \right] - u^* \sum_{i=1}^p \log v_i^* + \sum_{i=1}^p \left(\frac{k_i^*}{2} - 1 \right) \mathbb{E}_{q_2^*}(\lambda_i^2) \\ &\quad + \sum_{i=1}^p (v_i^* - 1) \mathbb{E}_{q_3^*}(\xi_i^{-2}) + \frac{l^*}{2} \sum_{i=1}^p \mathbb{E}_{q_2^*}(\lambda_i^{-2}), \end{aligned} \quad (\text{S3.4})$$

where $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind.

In each step of our algorithm, we compute the ELBO (S3.3). Convergence is assessed by computing the absolute difference, $\text{dif} = |\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}|$, at each iteration, and terminating the algorithm if $\text{dif} < \delta$, for some small tolerance $\delta > 0$. We run the MFVB algorithm until convergence or until a maximum of 1000 iterations have been reached.

To incorporate the EM algorithm for computing hyperparameters (a, b) into the MFVB scheme, we solve for (a, b) in (4.4) in every iteration of coordinate ascent algorithm, using $\mathbb{E}_{q_2^{*(t-1)}, a^{(t-1)}} [\log(\lambda_i^2)]$ and $\mathbb{E}_{q_3^{*(t-1)}, b^{(t-1)}} [\log(\xi_i^2)]$

in place of the summands in (4.4) at the t th iteration:

$$\begin{aligned}\mathbb{E}_{q_2^{*(t-1)}, a^{(t-1)}} [\log(\lambda_i^2)] &= \log \left(\frac{\sqrt{k_i^{*(t-1)}}}{\sqrt{l^*}} \right) + \frac{\partial}{\partial m^{*(t-1)}} \log \left[K_{m^{*(t-1)}} \left(\sqrt{k_i^{*(t-1)}} l^* \right) \right], \\ \mathbb{E}_{q_3^{*(t-1)}, b^{(t-1)}} [\log(\xi_i^2)] &= \log \left(v_i^{*(t-1)} \right) - \psi \left(u^{*(t-1)} \right),\end{aligned}\tag{S3.5}$$

where $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind, and $a^{*(t-1)}$, $b_i^{*(t-1)}$, $k_i^{*(t-1)}$, l^* , and $m^{*(t-1)}$ are taken from the $(t-1)$ st iteration and defined in (S3.2). Numerical differentiation is used to evaluate the derivative in the first equation of (S3.5). Algorithm 2 at the end of Appendix S2 provides the complete steps for implementing the variational EM algorithm for the self-adaptive NBP model. Note that Step 9 in Algorithm 2 involves computing the inverse of a $p \times p$ matrix, $\Phi^{*(t)} = (\mathbf{X}^\top \mathbf{X} + \mathbf{D}^{*(t)})^{-1}$. Since $\mathbf{D}^{*(t)}$ is a diagonal matrix, the computational cost can be substantially reduced when $p \gg n$ by invoking the Sherman-Morrison-Woodbury formula, i.e.

$$\Phi^{*(t)} \leftarrow (\mathbf{D}^{*(t)})^{-1} - (\mathbf{D}^{*(t)})^{-1} \mathbf{X}^\top (\mathbf{I}_n + \mathbf{X} (\mathbf{D}^{*(t)})^{-1} \mathbf{X}^\top)^{-1} \mathbf{X} (\mathbf{D}^{*(t)})^{-1},$$

which only involves inverting an $n \times n$ matrix, rather than $p \times p$ one. In steps 12-14 of Algorithm 2, we can also update $(k_i^{*(t)}, v_i^{*(t)})$, $i = 1, \dots, p$, simultaneously in parallel to save on computing time.

S3. DETAILS FOR THE MONTE CARLO EM AND VARIATIONAL EM
ALGORITHMS FOR THE SELF-ADAPTIVE NBP MODEL

Algorithm 1 Monte Carlo EM algorithm for the self-adaptive NBP

- 1: **Initialize:**
 - 2: $a^{(0)} = b^{(0)} = 0.01, c = d = 10^{-5}, \max = 100, M = 100, J = 20000,$
 $\delta = 10^{-6}, \text{dif} = 1, \text{and } k = 0.$
 - 3: Initialize $\beta^{(0)}, \sigma^{2(0)}, \lambda_i^{2(0)}, \xi_i^{2(0)}, i = 1, \dots, p.$
 - 4: **for** $t = 1$ to J **do**
 - 5: $\mathbf{D}^{(t)} \leftarrow \text{diag} \left(\lambda_1^{2(t-1)} \xi_1^{2(t-1)}, \dots, \lambda_p^{2(t-1)} \xi_p^{2(t-1)} \right)$
 - 6: Draw $\beta^{(t)} \sim \mathcal{N}_p \left((\mathbf{X}^\top \mathbf{X} + (\mathbf{D}^{(t)})^{-1})^{-1} \mathbf{X}^\top \mathbf{y}, \sigma^{2(t-1)} (\mathbf{X}^\top \mathbf{X} + (\mathbf{D}^{(t)})^{-1})^{-1} \right)$
 - 7: **for** $i = 1$ to p **do**
 - 8: Draw $\lambda_i^{2(t)} \sim \mathcal{IG} \left(a^{(k)} + \frac{1}{2}, \frac{(\beta_i^{(t)})^2}{2\sigma^{2(t-1)} \xi_i^{2(t-1)}} + 1 \right)$
 - 9: Draw $\xi_i^{2(t)} \sim \mathcal{GIG} \left(\frac{(\beta_i^{(t)})^2}{\sigma^{2(t-1)} \lambda_i^{2(t)}}, 2, b^{(k)} - \frac{1}{2} \right)$
 - 10: **end for**
 - 11: Draw $\sigma^{2(t)} \sim \mathcal{IG} \left(\frac{n+p+2c}{2}, \frac{\|\mathbf{y} - \mathbf{X}\beta^{(t)}\|_2^2 + (\beta^{(t)})^\top (\mathbf{D}^{(t)})^{-1} \beta^{(t)} + 2d}{2} \right)$
 - 12: **EM: Update hyperparameters** $(a, b).$
 - 13: **if** $t \bmod M = 0$ **and** $k \leq \max$ **and** $\text{dif} \geq \delta$ **then**
 - 14: $k \leftarrow k + 1$
 - 15: $\text{low} \leftarrow t - M + 1$
 - 16: $\text{high} \leftarrow t$
 - 17: **for** $j = 1$ to p **do**
 - 18: $U_j \leftarrow \frac{1}{M} \left[\ln \left(\lambda_j^{2(\text{low})} \right) + \dots + \ln \left(\lambda_j^{2(\text{high})} \right) \right]$
 - 19: $V_j \leftarrow \frac{1}{M} \left[\ln \left(\xi_j^{2(\text{low})} \right) + \dots + \ln \left(\xi_j^{2(\text{high})} \right) \right]$
 - 20: **end for**
 - 21: Solve for a in $-p\psi(a) - \sum_{j=1}^p U_j = 0$
 - 22: $a^{(k)} \leftarrow a$
 - 23: Solve for b in $-p\psi(b) + \sum_{j=1}^p V_j = 0$
 - 24: $b^{(k)} \leftarrow b$
 - 25: $\text{dif} \leftarrow (a^{(k)} - a^{(k-1)})^2 + (b^{(k)} - b^{(k-1)})^2$
 - 26: **end if**
 - 27: **end for**
-

Algorithm 2 Variational EM algorithm for the self-adaptive NBP model

- 1: **Initialize:**
 - 2: $l^* = 2$, $c^* = \frac{n+p+2c}{2}$, $a^{(0)} = b^{(0)} = 0.01$, $\delta = 10^{-3}$, $J = 1000$, and $t = 1$.
 - 3: Initialize $d^{*(0)}$, $k_i^{*(0)}$, $v_i^{*(0)}$, $i = 1, \dots, p$.
 - 4: **while** $|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}| \geq \delta$ and $1 \leq t \leq J$ **do**
 - 5: **E-step:** Update variational parameters in (S3.2).
 - 6: Update $m^{*(t)} \leftarrow a^{(t-1)} - \frac{1}{2}$
 - 7: Update $u^{*(t)} \leftarrow b^{(t-1)} + \frac{1}{2}$
 - 8: Update $\mathbf{D}^{*(t)} \leftarrow \text{diag} \left(\mathbb{E}_{q_2^{*(t-1)}}(\lambda_1^{-2}) \mathbb{E}_{q_3^{*(t-1)}}(\xi_1^{-2}), \dots, \mathbb{E}_{q_2^{*(t-1)}}(\lambda_p^{-2}) \mathbb{E}_{q_3^{*(t-1)}}(\xi_p^{-2}) \right)$
 - 9: Update $\Phi^{*(t)} \leftarrow (\mathbf{X}^\top \mathbf{X} + \mathbf{D}^{*(t)})^{-1}$
 - 10: Update $\Sigma^{*(t)} \leftarrow \mathbb{E}_{q_4^{*(t-1)}}(\sigma^2) \Phi^{*(t)}$
 - 11: Update $\beta^{*(t)} \leftarrow \Phi^{*(t)} \mathbf{X}^\top \mathbf{y}$
 - 12: **for** $i = 1$ to p **do**
 - 13: Update $k_i^{*(t)} \leftarrow \mathbb{E}_{q_1^{*(t-1)}}(\beta_i^2) \mathbb{E}_{q_4^{*(t-1)}}(\sigma^{-2}) \mathbb{E}_{q_3^{*(t-1)}}(\xi_i^{-2})$
 - 14: Update $v_i^{*(t)} \leftarrow \frac{1}{2} \mathbb{E}_{q_1^{*(t-1)}}(\beta_i^2) \mathbb{E}_{q_4^{*(t-1)}}(\sigma^{-2}) \mathbb{E}_{q_2^{*(t-1)}}(\lambda_i^{-2}) + 1$
 - 15: **end for**
 - 16: Update $d^{*(t)} \leftarrow \frac{\mathbb{E}_{q_1^{*(t-1)}}(\|\mathbf{y} - \mathbf{X}\beta^2\|_2^2) + \mathbb{E}_{q_4^{*(t-1)}}(\beta^\top \mathbf{D}^{*(t)} \beta) + 2d}{2}$
 - 17: **M-step:** Update hyperparameters (a, b) .
 - 18: Solve for a in $-p\psi(a) + \sum_{i=1}^p \mathbb{E}_{q_2^{*(t-1)}}[\log(\lambda_i^2)] = 0$
 - 19: $a^{(t)} \leftarrow a$
 - 20: Solve for b in $-p\psi(b) - \sum_{i=1}^p \mathbb{E}_{q_3^{*(t-1)}}[\log(\xi_i^2)] = 0$
 - 21: $b^{(t)} \leftarrow b$
 - 22: Update $\mathcal{L}^{(t)}$, as in (S3.4).
 - 23: $t \leftarrow t + 1$
 - 24: **end while**
-

Bibliography

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* 112, pp. 859-877.

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.

Song, Q. and Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv e-prints, 2017. arXiv:1712.08964*.

Department of Statistics, University of South Carolina

E-mail: RBAI@mailbox.sc.edu

Department of Statistics, University of Florida

E-mail: ghoshm@ufl.edu