Bayesian Bootstrap Spike-and-Slab LASSO

Lizhen Nie* and Veronika Ročková†

October 5, 2020

Abstract

The impracticality of posterior sampling has prevented the widespread adoption of spike-and-slab priors in high-dimensional applications. To alleviate the computational burden, optimization strategies have been proposed that quickly find local posterior modes. Trading off uncertainty quantification for computational speed, these strategies have enabled spike-and-slab deployments at scales that would be previously unfeasible. We build on one recent development in this strand of work: the Spike-and-Slab LASSO procedure of Ročková and George (2018). Instead of optimization, however, we explore multiple avenues for posterior sampling, some traditional and some new. Intrigued by the speed of Spike-and-Slab LASSO mode detection, we explore the possibility of sampling from an approximate posterior by performing MAP optimization on many independently perturbed datasets. To this end, we explore Bayesian bootstrap ideas and introduce a new class of jittered Spike-and-Slab LASSO priors with random shrinkage targets. These priors are a key constituent of the *Bayesian Bootstrap Spike-and-Slab LASSO* (BB-SSL) method proposed here. BB-SSL turns fast optimization into approximate posterior sampling. Beyond its scalability, we show that BB-SSL has a strong theoretical support. Indeed, we find that the induced pseudo-posteriors contract around the truth at a near-optimal rate in sparse normal-means and in high-dimensional regression. We compare our algorithm to the traditional Stochastic Search Variable Selection (under Laplace priors) as well as many state-of-the-art methods for shrinkage priors. We show, both in simulations and on real data, that our method fares superbly in these comparisons, often providing substantial computational gains.

Keywords: *Bayesian Bootstrap, Posterior Contraction, Spike-and-Slab LASSO, Weighted Likelihood Bootstrap.*

*4^{th} year PhD Student at the *Department of Statistics, University of Chicago*

†Associate Professor in Econometrics and Statistics and James S. Kemper Faculty Scholar at the *Booth School of Business, University of Chicago*.

The authors gratefully acknowledge the support from the James S. Kemper Faculty Fund at the Booth School of Business and the National Science Foundation (Grant No. NSF DMS-1944740).

1 Posterior Sampling under Shrinkage Priors

Variable selection is arguably one of the most widely used dimension reduction techniques in modern statistics. The default Bayesian approach to variable selection assigns a probabilistic blanket over models via spike-and-slab priors George and McCulloch (1993); Mitchell and Beauchamp (1988). The major conceptual appeal of the spike-and-slab approach is the availability of *uncertainty quantification* for both model parameters *as well as* models themselves (Madigan and Raftery (1994)). However, practical costs of posterior sampling can be formidable given the immense scope of modern analyses. The main thrust of this work is to extend the reach of existing posterior sampling algorithms in new faster directions.

This paper focuses on the canonical linear regression model, where a vector of responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is stochastically linked to fixed predictors $\boldsymbol{x}_i \in \mathbb{R}^p$ through

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_0 + \epsilon_i \quad \text{with} \quad \epsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{for} \quad 1 \leq i \leq n, \tag{1}$$

where $\sigma^2 > 0$ and where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a possibly sparse vector of regression coefficients. In this work, we assume that σ^2 is known and we refer to Moran et al. (2018) for elaborations with an unknown variance. We assume that the vector $\boldsymbol{Y}$ and the regressors $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p]$ have been centered and have thereby omitted the intercept. In the presence of uncertainty about which subset of $\boldsymbol{\beta}_0$ is in fact nonzero, one can assign a prior distribution over the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ as well as the pattern of nonzeroes $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ where $\gamma_j \in \{0, 1\}$ for whether or not the effect β_j is active. This formalism can be condensed into the usual spike-and-slab prior form

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}) = \prod_{j=1}^{p} [\gamma_j \psi_0(\beta_j) + (1 - \gamma_j)\psi_1(\beta_j)], \quad \mathbb{P}(\gamma_j = 1 \,|\, \theta) = \theta, \quad \theta \sim \text{Beta}(a, b), \tag{2}$$

where $a, b > 0$ are scale parameters and where $\psi_0(\cdot)$ is a highly concentrated prior density around zero (the spike) and $\psi_1(\cdot)$ is a diffuse density (the slab). The dual purpose of the spike-and-slab prior is to (a) shrink small signals towards zero and (b) keep large signals intact. The most popular incarnations of the spike-and-slab prior include: the point-mass spike (Mitchell and Beauchamp (1988)), the non-local slab priors (Johnson and Rossell (2012)), the Gaussian mixture (George and McCulloch (1993)), the Student

mixture (Ishwaran and Rao (2005)). More recently, Ročková (2018a) proposed the Spike-and-Slab LASSO (SSL) prior, a mixture of two Laplace distributions $\psi_0(\beta) = \frac{\lambda_0}{2}e^{-|\beta|\lambda_0}$ and $\psi_1(\beta) = \frac{\lambda_1}{2}e^{-|\beta|\lambda_1}$ where $\lambda_0 \gg \lambda_1$, which forms a continuum between the point-mass mixture prior and the LASSO prior (Park and Casella (2008)).

Posterior sampling under spike-and-slab priors is notoriously difficult. Dating back to at least 1993 (George and McCulloch, 1993), multiple advances have been made to speed up spike-and-slab posterior simulation (George and McCulloch (1997), Bottolo and Richardson (2010), Clyde et al. (2011), Hans (2009), Johndrow et al. (2017), Welling and Teh (2011), Xu et al. (2014)). More recently, several clever computational tricks have been suggested that avoid costly matrix inversions by using linear solvers (Bhattacharya et al., 2016) or by disregarding correlations between active and inactive coefficients (Narisetty et al., 2019). Neuronized priors have been proposed (Shin and Liu (2018)) that offer computational benefits by using a close approximations to spike-and-slab priors without latent binary indicators. Modern applications have nevertheless challenged MCMC algorithms and new computational strategies are desperately needed to keep pace with big data.

Optimization strategies have shown great promise and enabled deployment of spike-and-slab priors at scales that would be previously unfeasible (Ročková and George (2014), Ročková and George (2018), Carbonetto and Stephens (2012)). Fast posterior mode detection is effective in structure discovery and data exploration, a little less so for inference. In this paper, we review and propose new strategies for posterior sampling under the Spike-and-Slab LASSO priors, filling the gap between exploratory data analysis and proper statistical inference.

We capitalize on the latest MAP optimization and MCMC developments to provide several posterior sampling implementations for the Spike-and-Slab LASSO method of Ročková and George (2018). The first one (presented in Algorithm 1) is exact and conventional, following in the footsteps of Stochastic Search Variable Selection (George and McCulloch, 1993)). The second one is approximate and new. The cornerstone of this strategy is the Weighted Likelihood Bootstrap (WLB) of Newton and Raftery (1994) which was recently resurrected in the context of posterior sampling with sparsity priors by Newton et al. (2018)

and Fong et al. (2019). The main idea behind WLB is to perform approximate sampling by independently optimizing randomly perturbed likelihood functions. This optimization-based sampling is (a) unapologetically parallelisable, and (b) it does not require costly matrix inversions, thereby having the potential to meet the demands of large datasets. We extend the WLB framework by incorporating perturbations *both* in the likelihood and in the prior. The main contributions of this work are two-fold. First, we introduce BB-SSL (*Bayesian Bootstrap Spike-and-Slab LASSO*), a novel algorithm for approximate posterior sampling in high-dimensional regression under Spike-and-Slab LASSO priors. Second, we show that suitable "perturbations" lead to approximate posteriors that contract around the truth *at the same speed* (rate) as the actual posterior. These theoretical results have nontrivial practical implications as they offer guidance on the choice of the distribution for perturbing weights. Up until now, theoretical properties of WLB have been confined to consistency statements in low dimensions for iid data (Newton and Raftery (1994)). Our theoretical results, on the other hand, allow the dimensionality to increase with the sample size and go beyond mere consistency by showing that BB-SSL leads to rate-optimal estimation in sparse normal-means and high-dimensional regression under standard assumptions. Last but not least, we make thorough comparisons with the gold standard (i.e. exact MCMC sampling) on multiple simulated and real datasets, concluding that the proposed algorithm is scalable and reliable in practice.

The structure of this paper is as follows. Section 1.1 introduces the notation. Section 2 revisits Spike-and-Slab LASSO and presents a traditional algorithm for posterior sampling. Section 3 investigates performance of weighted Bayesian bootstrap, the building block of this work, in high dimensions. In Section 4, we introduce BB-SSL and present our theoretical study showing rate-optimality as well as its connection with other bootstrap methods. Section 5 shows simulated examples and Section 6 shows performance on real data. We conclude the paper with a discussion in Section 7.

1.1 Notation

With $\phi(y;\mu;\sigma^2)$ we denote the Gaussian density with a mean μ and a variance σ^2. We use $\xrightarrow{\text{d}}$ to denote convergence in distribution. We write $a_n = O_p(b_n)$ if for any $M > 0$ and $\epsilon > 0$, there exists $N > 0$ such that $\mathbb{P}(|a_n/b_n| > M) < \epsilon$ for any $n > N$. We write $a_n = o_p(b_n)$ if for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|a_n/b_n| > \epsilon) = 0$. We also write $a_n = O(b_n)$ as $a_n \lesssim b_n$. We use $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$. We use $a_n \gg b_n$ to denote $b_n = o(a_n)$ and $a_n \ll b_n$ to denote $a_n = o(b_n)$. We denote with $\boldsymbol{X}_A$ a sub-matrix consisting of columns of $\boldsymbol{X}$'s indexed by a subset $A \subset \{1, \cdots, p\}$ and with $\boldsymbol{P}_A$ the orthogonal projection to the range of $\boldsymbol{X}_A$ (Zhang and Zhang, 2012), i.e., $\boldsymbol{P}_A = \boldsymbol{X}_A \boldsymbol{X}_A^+$ where $\boldsymbol{X}_A^+$ is the Moore-Penrose inverse of $\boldsymbol{X}_A$. We denote with $\|\boldsymbol{X}\|$ the matrix operator norm of $\boldsymbol{X}$.

2 Spike-and-Slab LASSO Revisited

The Spike-and-Slab LASSO (SSL) procedure of Ročková and George (2018) recently emerged as one of the more successful non-convex penalized likelihood methods. Various SSL incarnations have spawned since its introduction, including a version for group shrinkage (Bai et al. (2020), Tang et al. (2018)), survival analysis (Tang et al. (2017)), varying coefficient models (Bai et al. (2020)) and/or Gaussian graphical models (Deshpande et al. (2019), Li et al. (2019)). The original procedure proposed for Gaussian regression targets a posterior mode

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \prod_{i=1}^{n} \phi(Y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}; \sigma^2) \times \int_\theta \prod_{j=1}^{p} \pi(\beta_j \mid \theta) d\pi(\theta) \right\}, \tag{3}$$

where $\pi(\beta_j \mid \theta) = \theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)$ is obtained from (2) by integrating out the missing indicator γ_j and by deploying $\psi_1(\beta_j) = \lambda_1/2\mathrm{e}^{-|\beta_j|\lambda_1}$ and $\psi_0(\beta_j) = \lambda_0/2\mathrm{e}^{-|\beta_j|\lambda_0}$ with $\lambda_0 \gg \lambda_1$. Ročková and George (2018) develop a coordinate-ascent strategy which targets $\widehat{\boldsymbol{\beta}}$ and which quickly finds (at least a local) mode of the posterior landscape. This strategy (summarized in Theorem 3.1 of Ročková and George (2018)) iteratively updates each $\widehat{\beta}_j$

using an implicit equation[1]

$$\widehat{\beta}_j = \frac{1}{\|\boldsymbol{X}_j\|_2^2}\left(|z_j| - \sigma^2 \lambda^*_{\widehat{\theta}_j}(\widehat{\beta}_j)\right)_+ \text{sign}(z_j) \times \mathbb{I}(|z_j| > \Delta_j) \tag{4}$$

where $\widehat{\theta}_j = \mathbb{E}[\theta \,|\, \widehat{\boldsymbol{\beta}}_{\setminus j}]$, $z_j = \boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}_{\setminus j}\widehat{\boldsymbol{\beta}}_{\setminus j})$ and

$$\Delta_j = \inf_{t>0}\left(\|\boldsymbol{X}_j\|^2 t/2 - \sigma^2 \rho(t \,|\, \widehat{\theta}_j)/t\right) \tag{5}$$

with $\rho(t \,|\, \theta) = -\lambda_1|t| + \log[p^\star_\theta(0)/p^\star_\theta(t)]$, where

$$p^\star(t) = \frac{\theta\psi_1(\beta_j)}{\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)} \quad \text{and} \quad \lambda^*_\theta(t) = \lambda_1 p^\star(t) + \lambda_0(1 - p^\star(t)). \tag{6}$$

Ročková and George (2018) also provide fast updating schemes for Δ_j and $\widehat{\theta}_j$. In this work, we are interested in *sampling from the posterior* as opposed to mode hunting.

One immediate strategy for sampling from the Spike-and-Slab LASSO posterior is the Stochastic Search Variable Selection (SSVS) algorithm of George and McCulloch (1993). One can regard the Laplace distribution (with a penalty $\lambda > 0$) as a scale mixture of Gaussians with an exponential mixing distribution (with a rate $\lambda^2/2$ as in Park and Casella (2008)) and rewrite the SSL prior using the following hierarchical form:

$$\begin{aligned}
\boldsymbol{\beta} \,|\, \boldsymbol{\tau} &\sim \mathcal{N}(\boldsymbol{0}, D_{\boldsymbol{\tau}}) \quad \text{with} \quad D_{\boldsymbol{\tau}} = \text{Diag}(1/\tau_1^2, 1/\tau_2^2, \cdots, 1/\tau_p^2), \\
\boldsymbol{\tau}^{-1} \,|\, \boldsymbol{\gamma} &\sim \prod_{i=1}^p \frac{\lambda_i^2}{2} e^{-\lambda_i^2/2\tau_i^2}, \quad \text{where} \quad \lambda_i = \gamma_i\lambda_1 + (1-\gamma_i)\lambda_0, \\
\gamma_i \,|\, \theta &\sim \text{Bernoulli}(\theta) \quad \text{with} \quad \theta \sim \text{Beta}(a, b),
\end{aligned}$$

where $\boldsymbol{\tau}^{-1} = (1/\tau_1^2, \ldots, 1/\tau_p^2)^T$ is the vector of variances. The conditional conjugacy of the SSL prior enables direct Gibbs sampling for $\boldsymbol{\beta}$ (Table 1). However, as with any other Gibbs sampler for Bayesian shrinkage models (Bhattacharya et al., 2015), this algorithm involves costly matrix inversions and can be quite slow when both n and p are large. In order to improve the MCMC computational efficiency when $p > n$, Bhattacharya et al. (2016) proposed a clever trick. By recasting the sampling step as a solution to a linear system, one

[1]Here we are not necessarily assuming that $\|\boldsymbol{X}_j\|_2^2 = n$ and the above formula is hence slightly different from Theorem 3.1 of Ročková and George (2018).

Algorithm 1 : SSVS

Set: $\lambda_0 \gg \lambda_1$, $a, b > 0$, T (number of MCMC iterations), B (number of samples to discard as burn-in).

Initialize: $\boldsymbol{\beta}^0$ (e.g. LASSO solution after 10-fold cross validation) and $\boldsymbol{\tau}^0$

for $t = 1, 2, \cdots, T$ **do**

(a) Sample $\boldsymbol{\beta}^t \sim \mathcal{N}(\mu_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}})$, where $\Sigma_{\boldsymbol{\gamma}} = (\boldsymbol{X}^T\boldsymbol{X}/\sigma^2 + D^{-1}_{\boldsymbol{\tau}^{t-1}})^{-1}$ and $\mu_{\boldsymbol{\gamma}} = \Sigma_{\boldsymbol{\gamma}}\boldsymbol{X}^T\boldsymbol{Y}/\sigma^2$.

(b) Sample $(\tau_j^t)^2 \sim \text{Inv-Gaus}(\mu_j', (\lambda_j')^2)$ for $j = 1, 2, \ldots, p$, where

$$\mu_j' = \frac{|\lambda_j'|}{|\beta_j^t|} \quad \text{and} \quad (\lambda_j')^2 = \gamma_j^{t-1}\lambda_1^2 + (1-\gamma_j^{t-1})\lambda_0^2$$

(c) Sample $\gamma_j^t \sim \text{Bernoulli}\left(\frac{\pi_1}{\pi_1+\pi_0}\right)$, where

$$\pi_1 = \theta^{t-1}\lambda_1^2 \mathrm{e}^{-\lambda_1^2/2(\tau_j^t)^2}/2 \quad \text{and} \quad \pi_0 = (1-\theta^{t-1})\lambda_0^2 \mathrm{e}^{-\lambda_0^2/2(\tau_j^t)^2}/2.$$

(d) Sample $\theta^t \sim \text{Beta}(\sum_{i=1}^p \gamma_i^t + a, p - \sum_{i=1}^p \gamma_i^t + b)$.

end

Return: $\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t, \theta^t$ where $t = B+1, B+2, \cdots, T$.

can circumvent a Cholesky factorization which would otherwise have a complexity $O(n^2p)$ per iteration. Building on this development, Johndrow et al. (2017) developed a blocked Metropolis-within-Gibbs algorithm to sample from horseshoe posteriors (Carvalho et al., 2010) and designed an approximate algorithm which thresholds small effects based on the sparse structure of the target. The exact method has a per-step complexity $O(n^2p)$ while the approximate one has only $O(np)$. In similar vein, the Skinny Gibbs MCMC method of Narisetty et al. (2019) also bypasses large matrix inversions by independently sampling from active and inactive β_i's. While the method is only approximate, it has a rather favorable computational complexity $O(np)$.

The impressive speed of the Spike-and-Slab LASSO mode detection makes one wonder whether performing many independent optimizations on randomly perturbed datasets will lead to posterior simulation that is more economical. Moreover, one may wonder whether the induced approximate posterior is sufficiently close to the actual posterior $\pi(\boldsymbol{\beta} \mid \boldsymbol{Y})$ and/or whether it can be used for meaningful estimation/uncertainty quantification. We attempt to address these intriguing questions in the next sections.

3 Likelihood Reweighting and Bayesian Bootstrap

The jumping-off point of our methodology is the weighted likelihood bootstrap (WLB) method introduced by Newton and Raftery (1994). The premise of WLB is to draw approximate samples from the posterior by independently maximizing randomly reweighted likelihood functions. Such a sampling strategy is computationally beneficial when, for instance, maximization is easier than Gibbs sampling from conditionals.

In the context of linear regression (1), the WLB method of Newton and Raftery (1994) will produce a series of draws $\widetilde{\boldsymbol{\beta}}_t$ by first sampling random weights $\boldsymbol{w}_t = (w_1^t, w_2^t, \cdots, w_n^t)^T$ from some weight distribution $\pi(\boldsymbol{w})$ and then maximizing a reweighted likelihood

$$\widetilde{\boldsymbol{\beta}}_t = \arg\max_{\boldsymbol{\beta}} \widetilde{L}^{\boldsymbol{w}_t}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)}) \tag{7}$$

where

$$\widetilde{L}^{\boldsymbol{w}_t}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)}) = \prod_{i=1}^{n} \phi(Y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}; \sigma^2)^{w_i^t}.$$

Newton and Raftery (1994) argue that for certain weight distributions $\pi(\boldsymbol{w})$, the conditional distribution of $\widetilde{\boldsymbol{\beta}}_t$'s given the data can provide a good approximation to the posterior distribution of $\boldsymbol{\beta}$. Moreover, WLB was shown to have nice theoretical guarantees when the number of parameters does not grow. Namely, under uniform Dirichlet weights (more below) and iid data samples, WLB is consistent (i.e. concentrating on any arbitrarily small neighborhood around an MLE estimator) and asymptotically first-order correct (normal with the same centering) for almost every realization of the data. The WLB method, however, is only approximate and it does not naturally accommodate a prior. Uniform Dirichlet weights provide a higher-order asymptotic equivalence when one chooses the squared Jeffrey's prior. However, for more general prior distributions (such as shrinkage priors considered here), the correspondence between the prior $\pi(\boldsymbol{\beta})$ and $\pi(\boldsymbol{w})$ is unknown. Newton and Raftery (1994) suggest post-processing the posterior samples with importance sampling to leverage prior information. This pertains to Efron (2012), who proposes a posterior sampling method for exponential family models with importance sampling on parametric bootstrap distributions.

Alternatively, Newton et al. (2018) suggested blending the prior directly into WLB by including a weighted prior term, i.e. replacing (7) with

$$\widetilde{\boldsymbol{\beta}}_t = \arg\max_{\boldsymbol{\beta}} \widetilde{L}^{\boldsymbol{w}_t}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)})\pi(\boldsymbol{\beta})^{\widetilde{w}^t},$$

where $w_i^t \overset{\text{i.i.d}}{\sim} Exp(1)$[2]. This so called Weighted Bayesian Bootstrap (WBB) method treats the prior weight $\widetilde{w}^t$ as either fixed (and equal to one) or as one of the random data weights arising from the exponential distribution. We explore these two strategies in the next section within the context of the Spike-and-Slab LASSO where $\pi(\boldsymbol{\beta})$ is the SSL shrinkage prior implied by (2).

3.1 WBB meets Spike-and-Slab LASSO

Since SSL is a thresholding procedure (see (4)), WBB will ultimately create samples from pseudo-posteriors that have a point mass at zero. This is misleading since the posterior under the Gaussian likelihood and a single Laplace prior is half-normal (Hans (2009), Park and Casella (2008)). Deploying the WBB method thus does not guarantee that uncertainty be propertly captured for the zero (negligible) effects since their posterior samples may very often be exactly zero. We formalize this intuition below. We want to understand the extent to which the WBB (or WLB) pseudo-posteriors correspond to the actual posteriors. To this end, we focus on the canonical Gaussian sequence model

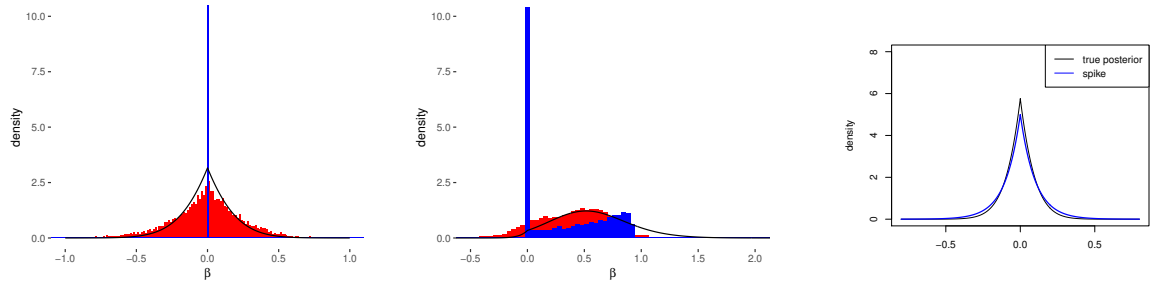
$$y_i = \beta_i^0 + \epsilon_i/\sqrt{n} \quad \text{for} \quad i = 1, 2, \cdots n. \tag{8}$$

Under the separable SSL prior (i.e. θ fixed), the true posterior is a mixture

$$\pi\left(\beta_i \,|\, y_i\right) = w_1 \pi\left(\beta_i \,|\, y_i, \gamma_i = 1\right) + w_0 \pi\left(\beta_i \,|\, y_i, \gamma_i = 0\right) \tag{9}$$

where $w_1 = \pi\left(\gamma_i = 1 \,|\, y_i\right)$ and $w_0 = \pi\left(\gamma_i = 0 \,|\, y_i\right)$. From Hans (2009), we know that $\pi\left(\beta_i \,|\, y_i, \gamma_i = 1\right)$ and $\pi\left(\beta_i \,|\, y_i, \gamma_i = 0\right)$ are orthant truncated Gaussians and thus $\pi\left(\beta_i \,|\, y_i\right)$ is a mixture of orthant truncated Gaussians.

[2]Note that if $w_1, w_2, \cdots, w_n \overset{\text{i.i.d}}{\sim} Exp(1)$, then $\frac{w_1}{\sum w_i}, \frac{w_2}{\sum w_i}, \cdots, \frac{w_n}{\sum w_i} \sim Dir(1, 1, \cdots, 1)$ which brings us back to the uniform Dirichlet distribution.



(a) $y_i = 0.1$ (b) $y_i = 1$ (c) True posterior versus spike.

Figure 1: True and WBB approximated posterior distribution $\pi(\beta_i \,|\, y_i)$ under the separable SSL prior with $\lambda_0 = 5, \lambda_1 = 0.1$ and $\theta = 0.2$ and a Gaussian sequence model $y_i = x_i + \epsilon_i/\sqrt{n}$ with $n = 20$. Red bins represent BB-SSL pseudo-posterior, blue bins represent WBB pseudo-posterior (with a random prior weight), black line is the true posterior and $\alpha = 2.5$. The plot (c) is for the same setting except for $\lambda_0 = 10$.

We start by examining the posterior distribution of *active coordinates* such that $|y_i| > |\beta_i^0|/2 > 0$ (this event happens with high probability when n is sufficiently large). For the true posterior, we show in the Appendix (Proposition 1) that $w_0 \to 0$ and $w_1 \to 1$. The true posterior $\pi\left(\beta_i \,|\, y_i\right)$ is hence dominated by the component $\pi\left(\beta_i \,|\, y_i, \gamma_i = 1\right)$, which takes the following form

$$\pi\left(\beta_i \,|\, y_i, \gamma_i = 1\right) = \frac{\mathbb{I}(\beta_i \geq 0) c_1^{(-)} \phi_1^{(-)}(\beta_i) + \mathbb{I}(\beta_i < 0) c_1^{(+)} \phi_1^{(+)}(\beta_i)}{\int_0^\infty c_1^{(-)} \phi_1^{(-)}(\beta_i) d\beta_i + \int_{-\infty}^0 c_1^{(+)} \phi_1^{(+)}(\beta_i) d\beta_i}$$

where

$$c_1^{(-)} = \theta \lambda_1 e^{-y_i \lambda_1 + \lambda_1^2/2n} \quad \text{and} \quad c_1^{(+)} = \theta \lambda_1 e^{y_i \lambda_1 + \lambda_1^2/2n}, \tag{10}$$

$$\phi_1^{(-)}(x) = \phi\left(x; y_i - \frac{\lambda_1}{n}, \frac{1}{n}\right) \quad \text{and} \quad \phi_1^{(+)}(x) = \phi\left(x; y_i + \frac{\lambda_1}{n}, \frac{1}{n}\right). \tag{11}$$

Intuitively, λ_1/n vanishes when n is large, so both $\phi_1^{(-)}(\beta_i)$ and $\phi_1^{(+)}(\beta_i)$ will be close to $\phi(\beta_i; y_i, \frac{1}{n})$. This intuition is proved rigorously in the Appendix (Section A.12.2), where we show that the density of the transformed variable $\sqrt{n}(\beta_i - y_i)$ converges pointwise to the standard normal density and thereby the posterior $\pi\left(\sqrt{n}(\beta_i - y_i) \,|\, y_i, \gamma_i = 1\right)$ converges to $N(0, 1)$ in total variation (Scheffé (1947)).

We now investigate the limiting shape of the pseudo-distribution obtained from WBB. For a given weight $w_i > 0$, the WBB estimator $\widehat{\beta}_i$ equals

$$\widehat{\beta}_i = \begin{cases} 0, & \text{if } |\sqrt{w_i} y_i| \leq \Delta/n. \\ \left[|y_i| - \frac{1}{w_i n} \lambda^*(\widehat{\beta}_i) \right]_+ \text{sign}(\sqrt{w_i n} y_i), & \text{otherwise.} \end{cases} \tag{12}$$

When $\widehat{\beta}_i \neq 0$, we show in the Appendix (Section A.12.2) that $n\,(\widehat{\beta}_i - y_i) \to -\frac{1}{w_i}\lambda_1$. Under the condition $|y_i| > \frac{|\beta_i^0|}{2} > 0$, it can be shown (Section A.12.2) that $\mathbb{P}_{w_i}\,(\widehat{\beta}_i = 0 \,|\, y_i) \to 0$. For active coordinates, the distribution of the WBB samples $\widehat{\beta}_i$ is thus purely determined by that of $-\frac{1}{w_i}$. The shape of this posterior can be very different from the standard normal one, as can be seen from Figure 1. In particular, Figure 1b shows how WBB (a) assigns a non-negligible prior mass to zero (in spite of evidence of signal) (b) incurs bias in estimation and (c) underestimates variance with a skewed misrepresentation of the posterior distribution. This last aspect is particularly pronounced when the signal is even stronger.[3]

The approximability of WBB does not get any better for *inactive coordinates* such that $\beta_i^0 = 0$ and thereby $|y_i| = O_p(\frac{1}{\sqrt{n}})$ from (8). The following arguments will be under the assumption $|y_i| \asymp \frac{1}{\sqrt{n}}$. One can show (Section A.12.3 in the Appendix) that the true posterior $\pi\,(\beta_i \,|\, y_i)$ is dominated by the component $\pi\,(\beta_i \,|\, y_i, \gamma_i = 0)$ since $w_0 \to 1$ and $w_1 \to 0$. When n is sufficiently large, one can then approximate this distribution with the Laplace spike, indeed $\pi\,(\lambda_0 \beta_i \,|\, y_i, \gamma_i = 0)$ converges to $\frac{1}{2} e^{-|\lambda_0 \beta_i|}$ in total variation (Section A.12.3 in the Appendix). When the signal is weak, the posterior thus closely resembles the spike Laplace distribution, as can be seen from Figure 1c. For the fixed (and also random) WBB pseudo-posteriors, we show (in Section A.12.3 in the Appendix) that whenever $\mathbb{E} e^{t w_i}$ (or $\mathbb{E} e^{t w_i / w_p}$ for random WBB) is bounded for some $t > 0$, the posterior converges to a point mass at 0, i.e. $\mathbb{P}_{w_i}\,(\widehat{\beta}_i = 0 \,|\, y_i) \to 1$. This is a misleading approximation of the actual posterior (Figure 1a). To conclude, since SSL is always shrinking the estimates towards 0, WBB samples will often be zero. The true posterior, however, follows roughly a spike

[3] In fact, under the uniform Dirichlet distribution, the marginal distribution becomes $w_i \sim n \times \text{Beta}(1, n-1)$. Since $n \times \text{Beta}(1, n-1) \xrightarrow{\text{d}} \text{Gamma}(1,1)$, the distribution of $-\frac{1}{w_i}$ converges to Inverse-Gamma(1,1) which exhibits a skewed shape, which is in sharp contrast to the symmetric Gaussian distribution of the true posterior.

Algorithm 2 : BB-SSL Sampling

Set: $\lambda_0 \gg \lambda_1$, $a, b > 0$ and T (number of iterations).
for $t = 1, 2, \cdots, T$ **do**
  (a) Sample $\boldsymbol{w}_t \sim \pi(\boldsymbol{w})$.
  (b) Sample $\boldsymbol{\mu}_t$ from $\mu_j^t \overset{iid}{\sim} \psi_0(\mu)$.
  (c) Calculate $\widetilde{\boldsymbol{\beta}}_t$ from (14).
end

Laplace distribution when the signal is weak. Motivated by Papandreou and Yuille (2010), one possible solution is to introduce randomness in the shrinkage target of the prior.

4 Introducing BB-SSL

Similarly as Newton et al. (2018), we argue that the random perturbation should affect both the prior and the data. Instead of inflating the prior contribution by fixed or random weight, we *perturb the prior mean for each coordinate*. This creates a random shift in the centering of the prior so that the posterior can shrink to a random location as opposed to zero. Instead of the prior $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma})$ in (2) which is centered around zero, we consider a variant that uses hierarchical jittered Laplace distributions.

Definition 4.1. *For $\lambda_0 \gg \lambda_1$, a location shift vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)^T \in \mathbb{R}^p$ and a prior inclusion weight $\theta \in (0, 1)$, the* jittered Spike-and-Slab LASSO prior *is defined as*

$$\widetilde{\pi}(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}, \theta) = \prod_{i=1}^{p} [\theta \psi_1(\beta_i; \mu_i) + (1 - \theta)\psi_0(\beta_i; \mu_i)], \quad \textit{where} \quad \mu_i \overset{iid}{\sim} \psi_0(\mu) \tag{13}$$

and where $\psi_1(\beta; \mu) = \lambda_1/2e^{-|\beta-\mu|\lambda_1}$, $\psi_0(\beta; \mu) = \lambda_0/2e^{-|\beta-\mu|\lambda_0}$ and $\psi_0(\beta) = \psi_0(\beta; 0)$.

The Bayesian Bootstrap Spike-and-Slab LASSO (which we abbreviate as BB-SSL) is obtained by maximizing a pseudo-posterior obtained by reweighting the likelihood and recentering the prior. Namely, one first samples weights w_i^t, one for each observation, from $\boldsymbol{w}_t = (w_1^t, \ldots, w_n^t)^T \sim n\,\text{Dirichlet}(\alpha, \cdots, \alpha)$ for some $\alpha > 0$ (we discuss choices in the next section). Second, one samples the location shifts $\boldsymbol{\mu}_t = (\mu_1^t, \ldots, \mu_p^t)^T$ from the spike distribution as in (13). Lastly, a draw $\widetilde{\boldsymbol{\beta}}_t$ from the BB-SSL posterior is obtained as a

pseudo-MAP estimator

$$\widetilde{\boldsymbol{\beta}}_t = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \widetilde{L}^{\boldsymbol{w}_t}(\boldsymbol{\beta}, \sigma^2; \boldsymbol{X}^{(n)}, \boldsymbol{Y}^{(n)}) \times \int_\theta \widetilde{\pi}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_t, \theta) \mathrm{d}\, \pi(\theta) \right\}. \tag{14}$$

BB-SSL can be implemented by directly applying the SSL algorithm we described in the previous section on randomly perturbed data. In particular, denote with $\boldsymbol{\beta}^* = \boldsymbol{\beta} - \boldsymbol{\mu}_t$, $Y_i^* = \sqrt{w_i^t}\left(Y_i - \boldsymbol{x}_i^T \boldsymbol{\mu}_t\right)$, $\boldsymbol{x}_i^* = \sqrt{w_i^t}\boldsymbol{x}_i$. We can first calculate

$$\widehat{\boldsymbol{\beta}_t^*} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \widetilde{L}^{\boldsymbol{w}_t=(1,1,\cdots,1)}(\boldsymbol{\beta}^*, \sigma^2; \boldsymbol{X}^*, \boldsymbol{Y}^*) \times \int_\theta \widetilde{\pi}(\boldsymbol{\beta} \mid \boldsymbol{0}, \theta) \mathrm{d}\, \pi(\theta) \right\},$$

and then get $\widetilde{\boldsymbol{\beta}}_t$ through a post-processing step $\widetilde{\boldsymbol{\beta}}_t = \widehat{\boldsymbol{\beta}_t^*} + \boldsymbol{\mu}_t$.

4.1 Theory for BB-SSL

For the uniform Dirichlet weights, Newton and Raftery (1994) (Theorem 1 and 2) show first-order correctness, i.e. consistency and asymptotic normality, of WLB in low dimensional settings (a fixed number of parameters) and iid observations. Their result can be generalized to WBB (Newton et al., 2018) as well as BB-SSL. While the uniform Dirichlet weight distribution is a natural choice, Newton and Raftery (1994) point out that it is doubtful that such weights would yield good higher-order approximation properties. The authors leave open the question of relating the weighting distribution to the model itself and to a more general prior. Here, we address this question by looking into asymptotics for guidance about the weight distribution. We focus on high-dimensional scenarios where the number of parameters ultimately increases with the sample size.

In particular, we provide sufficient conditions for the weight distribution $\pi(\boldsymbol{w})$ so that the pseudo-posterior concentrates at the same rate as the actual posterior under the same prior settings. After stating the result for general weight distributions, we particularize our considerations to Dirichlet and gamma distributions and provide specific guidance for implementation. Our first result is obtained for the canonical high-dimensional normal-means problem, where $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)^T$ is observed as a noisy version of a sparse mean vector $\boldsymbol{\beta}_0 = (\beta_1^0, \ldots, \beta_n^0)^T$, i.e.

$$Y_i = \beta_i^0 + \epsilon_i, \quad \text{where} \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0,1) \quad \text{for } 1 \le i \le n. \tag{15}$$

Theorem 4.1 (Normal Means). *Consider the normal means model (15) with* $q = \|\boldsymbol{\beta}_0\|_0$ *such that* $q = o(n)$ *as* $n \to \infty$. *Assume the SSL prior with* $0 < \lambda_1 < \frac{1}{e^2}$ *and* $\theta \asymp (\frac{q}{n})^\eta, \lambda_0 \asymp (\frac{n}{q})^\gamma$ *with* $\eta, \gamma > 0$ *such that* $\eta + \gamma \geq 1$. *Assume that* $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ *are non-negative and arise from* $\pi(\boldsymbol{w})$ *such that*

(1) $\mathbb{E}\, w_i = 1$ *for each* $1 \leq i \leq n$,

(2) $\exists\, C_1, C_2 > 0$ *such that* $\mathbb{E}\left(\frac{1}{w_i}\right) \leq C_1$ *and* $\mathbb{E}\left(\frac{1}{w_i^2}\right) \leq C_2$ *for each* $1 \leq i \leq n$,

(3) $\exists\, C_3 > 0$ *such that for each* $1 \leq i \leq n$

$$\mathbb{P}(w_i > \eta + \gamma) \leq C_3 \frac{q}{n} \sqrt{\log\left(\frac{n}{q}\right)}.$$

Then, for any $M_n \to \infty$, *the BB-SSL posterior concentrates at the minimax rate, i.e.*

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0} \mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu}} \left[\|\widetilde{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{w}} - \boldsymbol{\beta}_0\|_2^2 > M_n\, q \log\left(\frac{n}{q}\right) \mid \boldsymbol{Y}^{(n)} \right] = 0. \tag{16}$$

Proof. See Section A.1 in the Appendix.

In Theorem 4.1, $\widetilde{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{w}}$ denotes the BB-SSL sample whose distribution, for each given $\boldsymbol{Y}^{(n)}$, is induced by random weights $\boldsymbol{w}$ arising from $\pi(\boldsymbol{w})$ and random recentering $\boldsymbol{\mu}$ arising from $\psi_0(\cdot)$. Despite the approximate nature of BB-SSL, the concentration rate (16) is *minimax optimal* and it is the same rate achieved by the *actual* posterior distribution under the same prior assumptions (Ročková (2018a)). Condition (1) in Theorem 4.1 is not surprising and aligns with considerations in Newton et al. (2018). Conditions (2) and (3) can be viewed as regularizing the tail behavior of w_i's (left and right, respectively). While Newton et al. (2018) only showed consistency for iid models in finite-dimensional settings, Theorem 4.1 is far stronger as it shows optimal convergence rate in a high-dimensional scenario. The following Corollary discusses specific choices of $\pi(\boldsymbol{w})$.

Corollary 4.1. *Assume the same model and prior as in Theorem 4.1. Next, when* $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)^T \sim n \times \mathrm{Dir}(\alpha, \alpha, \cdots, \alpha)$ *with* $\alpha \gtrsim \sigma^2 \log[\frac{(1-\theta)\lambda_0}{\theta\lambda_1}]$ *or* $w_i \overset{i.i.d}{\sim} \frac{1}{\alpha} Gamma(\alpha, 1)$ *with* $\alpha \gtrsim \sigma^2 \log\left[\frac{(1-\theta)\lambda_0}{\theta\lambda_1}\right]$, *the BB-SSL posterior satisfies (16).*

Proof. See the Appendix (Section A.2).

Theorem 4.1 and Corollary 4.1 give insights into what weight distributions are appropriate for sparse normal means. In parametric models, the uniform Dirichlet distribution would be enough to achieve consistency (Newton and Raftery, 1994). It is interesting to note, however, that in the non-parametric normal means model, the assumption $\boldsymbol{w} \sim n \times \text{Dir}(\alpha, \cdots, \alpha)$ for $\alpha < 2$ yields risk (for active coordinates) that can be arbitrarily large (as we show in Section A.3 in the Appendix). The requirement $\alpha \geq 2$ is thus *necessary* for controlling the risk of active coordinates and the plain uniform Dirichlet prior (with $\alpha = 1$) would not be appropriate.

In the following theorems, we study the high-dimensional regression model (1).

Theorem 4.2 (Regression Model Size)**.** *Consider the regression model (1) with $q = ||\boldsymbol{\beta}||_0$ (unknown). Assume the SSL prior with $(1-\theta)/\theta \asymp p^{\eta}$ and $\lambda_0 \asymp p^{\gamma}$ where $\eta, \gamma \geq 1$. Assume that $\boldsymbol{w} = (w_1, \cdots, w_n)^T$ are non-negative and arise from $\pi(\boldsymbol{w})$ such that*

(1) $\mathbb{E}\, w_i = 1$ *for each* $1 \leq i \leq n$,

(2) $\exists\, m \in (0,1)$ *s.t.* $\lim_{n\to\infty} \mathbb{P}(\min_i w_i > m) = 1$,

(3) $\exists\, M > 1$ *s.t.* $\lim_{n\to\infty} \mathbb{P}(\max_i w_i < M) = 1$,

(4) $Var(w_i) \lesssim \frac{1}{\log n}$, $Cov(w_i, w_j) = C_0 \lesssim \frac{1}{n \log n}$ *for any* $1 \leq i, j \leq n$,

(5) Set $(1+\xi_0)\frac{\sigma}{\widetilde{\eta}}\sqrt{n}(1+\sqrt{\frac{5}{2}\log p}) \leq \lambda_1 \leq 4\sqrt{n \log p}$ *where* $\widetilde{\eta} \in (0, m)$ *and* $\xi_0 > 0$ *satisfies*

$$\max\{\lambda_{max}^{1/2}\left(\boldsymbol{X}_B^T \boldsymbol{P}_A \boldsymbol{X}_B/n\right) : B \cup A = \emptyset, |A| = rank(P_A) = |B| = k,$$
$$k(1+\xi_0)^2\left(1+\sqrt{2.5\log p}\right)^2 \leq 2n\} \leq \xi_0$$

(6) $D = \frac{\left(M\eta^*\sqrt{\eta+\gamma}+2\sqrt{2}\right)^2}{mc^2(\eta+\gamma-1)} < 1-\delta$ *for some* $\delta > 0$, *where* $\eta^* = \max\left\{\widetilde{\eta} + C_n\frac{\|\boldsymbol{X}\|}{\lambda_1}, \frac{\widetilde{\eta}}{m}\right\} \in (0,1)$, C_n *is a sequence s.t.* $C_n \to \infty$ *and* $c = c(\eta^*; \boldsymbol{\beta})$.

Then the BB-SSL posterior satisfies

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0}\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{w}}\left(\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{w}}^{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_0 \leq q(1+K) \mid \boldsymbol{Y}^{(n)}\right) = 1$$

where $K = 2\frac{D}{1-D}$. *The definition of* $c(\eta^*; \boldsymbol{\beta})$ *is in the Appendix, Section A.4.*

Proof. Section A.9 in the Appendix.

Theorem 4.3 (Regression model)**.** *Under the same conditions as in Theorem 4.2, the BB-SSL posterior concentrates at the near-minimax rate, i.e.,*

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\beta}_0}\mathbb{P}_{\boldsymbol{w},\boldsymbol{\mu}}\left(\|\widetilde{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{w}} - \boldsymbol{\beta}_0\|_2^2 > \frac{C_5(\eta^*)^2 M^2}{m\,\phi^2\,c^2} q(1+K)\frac{\log p}{n} \mid \boldsymbol{Y}^{(n)}\right) = 0 \tag{17}$$

where $c = c(\eta^*;\boldsymbol{\beta}), \phi = \phi\left(C(\eta^*;\boldsymbol{\beta}_0)\right)$*, whose definition are in the Appendix A.4.*

Proof. Section A.10 in the Appendix.

It follows from Ročková and George (2018) and from (17) that the BB-SSL posterior achieves the same rate of posterior concentration as the *actual* posterior. In Theorem 4.2, Conditions (1)-(4) regulate the distribution $\pi(\boldsymbol{w})$ while Conditions (5) and (6) impose requirements on $\boldsymbol{X}$, λ_0 and λ_1. Conditions (2) and (3) are counterparts of (2) and (3) in Theorem 4.1 and control the left and right tail of w_i's, respectively. The larger M (or the smaller m) is, the larger D and K will become and the larger the bound on $\|\widetilde{\boldsymbol{\beta}}^{\boldsymbol{\mu}}_{\boldsymbol{w}} - \boldsymbol{\beta}_0\|_2^2$ will be. Compared with the normal means model, we have one additional Condition (4) which requires that each w_i becomes more and more concentrated around its mean and that w_i's are asymptotically uncorrelated. It is interesting to note that distributions $\pi(\boldsymbol{w})$ in Corollary 4.1 both satisfy Condition (4). Moreover, the Dirichlet distribution in Corollary 4.1 achieves both upper bounds tightly. Finally, Condition (5) ensures that identifiability holds with high probability (Zhang and Zhang (2012)) and Condition (6) ensures that our bound is meaningful.[4] In practice, many distributions will satisfy Conditions (1)-(4), e.g. bounded distributions with a proper covariance structure or distributions from Corollary 4.1.

[4]In order for η^* to be a bounded real number smaller than 1, we would need $\lambda_1/\|\boldsymbol{X}\| \to \infty$. For example when $p \asymp n$, for random matrix where each element is generated independently by Gaussian distribution, we have $\|\boldsymbol{X}\| = O_p(\sqrt{n}+\sqrt{p})$ (Vivo et al. (2007)). So in order for such a sequence C_n (s.t. $C_n \to \infty$) to exist, we need $\lambda_1/(\sqrt{n}+\sqrt{p}) \to \infty$. We can choose $\lambda_1 = (\sqrt{n}+\sqrt{p})\sqrt{\log p}$ and $C_1 = \log\frac{\lambda_1}{\|\boldsymbol{X}\|}$ under such settings.

Algorithm 3 : Posterior Bootstrap Sampling

Data: Data $(Y_i, \boldsymbol{x}_i)$ for $1 \leq i \leq n, x_i \in \mathbb{R}^p$, truncation limit m

Result: $\tilde{\boldsymbol{\beta}}^t, t = 1, 2, \cdots, T$

for $t = 1, 2, \cdots, T$ **do**

(a) Draw prior pseudo-samples $\tilde{\boldsymbol{x}}_{1:m}, \tilde{y}_{1:m} \sim F_\pi$.

(b) Draw $(w_{1:n}, \tilde{w}_{1:m}) \sim \text{Dir}(1, 1, \cdots, 1, c/m, c/m, \cdots, c/m)$.

(c) Calculate $\widetilde{\boldsymbol{\beta}}^t = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{j=1}^n w_j l(\boldsymbol{x}_j, y_j, \boldsymbol{\beta}) + \sum_{k=1}^m \tilde{w}_k l(\tilde{\boldsymbol{x}}_k, \tilde{y}_k, \boldsymbol{\beta}) \right\}$.

end

Remark 4.1. *In the regression model, when* $\boldsymbol{w}$ *arises from the same distribution as in Corollary 4.1, Conditions (1)-(4) in Theorem 4.3 are satisfied by setting* $m = \frac{1}{e}$ *and* $M = \frac{2}{3}(\eta + \gamma)$. *The detailed proof is in Section A.11 in the Appendix.*

4.2 Connections to Other Bootstrap Approaches

Our approach bears a resemblance to Bayesian non-parametric learning (NPL) introduced by Lyddon et al. (2018) and Fong et al. (2019) which generates exact posterior samples under a Bayesian non-parametric model that assumes less about the underlying model structure. Under a prior on the sampling distribution function F_π, one can use WBB (and also WLB) to draw samples from a posterior of F_π by optimizing a randomly weighted loss function $l(\cdot)$ based on an enlarged sample (observed plus pseudo-samples) with weights following a Dirichlet distribution (see Algorithm 3 which follows from Fong et al. (2019)). Despite the fact that these two procedures have different objectives, there are many interesting connections. In particular, the idea of randomly perturbing the prior has an effect similar to adding pseudo-samples $\widetilde{\boldsymbol{x}}_{1:m}, \widetilde{y}_{1:m}$ from the prior $F_\pi(\boldsymbol{x}, y)$ defined through

$$\widetilde{\boldsymbol{x}}_k \sim F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n \delta(\boldsymbol{x}_i), \quad \widetilde{y}_k \,|\, \widetilde{\boldsymbol{x}}_k = \widehat{y}_k + \widetilde{\boldsymbol{x}}_k^T \boldsymbol{\mu}$$

where $\delta(\cdot)$ is the Dirac measure, $\boldsymbol{\mu}$ is the Spike, and $\widehat{y}_k = y_i$ where i satisfies $\widetilde{\boldsymbol{x}}_k = \boldsymbol{x}_i$. A motivation for this prior is derived in the Appendix (Section B). Under this prior, the NPL posterior samples $\widetilde{\boldsymbol{\beta}}^t$ generated by Algorithm 3 approximately follow the distribution (see

Section B in the Appendix)

$$\widetilde{\boldsymbol{\beta}}^t \overset{\mathrm{d}}{\approx} \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p} \left\{ -\frac{1}{2}\sum_{i=1}^{n} w_i^*(Y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})^2 + \log\left[\int \prod_{j=1}^{p} \pi\left(\tilde{\beta}_j - \frac{c}{c+n}\mu_j^* \mid \theta\right) d\pi(\theta)\right]\right\} - \frac{c}{c+n}\boldsymbol{\mu}^* \tag{18}$$

where $(w_1^*, w_2^*, \cdots, w_n^*)^T \sim n \times \mathrm{Dir}(1+c/n, \cdots, 1+c/n)$ and where c represents the strength of our belief in F_π and can be interpreted as the effective sample size from F_π. In comparison with the BB-SSL estimate

$$\widetilde{\boldsymbol{\beta}}^t = \arg\max_{\tilde{\boldsymbol{\beta}}\in\mathbb{R}^p} \left\{ -\frac{1}{2}\sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}})^2 + \log\left[\int_\theta \prod_{j=1}^{p} \pi\left(\tilde{\beta}_j - \mu_j \mid \theta\right) d\pi(\theta)\right]\right\} \tag{19}$$

where $(w_1, w_2, \cdots, w_n)^T \sim n \times \mathrm{Dir}(\alpha, \cdots, \alpha)$, both (18) and (19) are shrinking towards a random location and both are using Dirichlet weights. The main difference is in the choice of the concentration parameter c. When $c = 0$, (18) reduces to WBB (with a fixed weight on the prior) which reflects less confidence in the prior F_π and thus less prior perturbation (location shift). When c is large, (18) becomes more similar to (19) where the prior F_π is stronger and thereby more prior perturbation is induced. Another difference is that (18), although shrinking towards a random location $\frac{c}{c+n}\mu_j^*$, adds the location back which results in less variance (see Figures in Section B in the Appendix).

5 Simulations

We compare the empirical performance of our method with several existing posterior sampling methods including WBB (Newton et al. (2018)), SSVS (George and McCulloch (1993)), and Skinny Gibbs (Narisetty et al. (2019)). We implement two versions of WBB: WBB1 (with a fixed prior weight) and WBB2 (with a random prior weight). We also implement the original SSVS algorithm (Algorithm 1 further referred to as SSVS1) and compare its complexity and running times with its faster version (further referred to as SSVS2) which uses the trick from Bhattacharya et al. (2016). Comparisons are based on the marginal posterior distributions for β_i's, marginal inclusion probabilities (MIP) $\mathbb{P}(\gamma_i = 1 \mid \boldsymbol{Y}^{(n)})$ as well as the joint posterior distribution $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}^{(n)})$. As the benchmark gold standard for comparisons, we run SSVS initialized at the truth for a sufficiently large

Algorithm	Complexity
SSVS1	$O(p^2 \max(p, n))$
SSVS2	$O(n^2 p)$
Skinny Gibbs	$O(np)$
WLB	$O(np^2)$ when $n \leq p$, not applicable when $p > n$
BB-SSL	$O\left(\min\left(\text{maxiter} \times p(n + \frac{p}{c_1}), (n + \text{maxiter}) \times p^2\right)\right)$ for a single value λ_0

Table 1: A computational complexity analysis (per sample) of each algorithm. `Maxiter` is the user-specified maximum number of iterations with a default value 500. For BB-SSL, c_1 is the pre-specified number of iterations after which the value of theta is updated with a default value 10. By setting $c_1 \propto p$ we have BB-SSL complexity $O(np)$.

number of iterations T and discard the first B samples as a burn-in. We use the same T and B for Skinny Gibbs except that we initialize $\boldsymbol{\beta}$ at the origin. For BB-SSL, we draw weights $\boldsymbol{w} \sim n \times \text{Dir}(\alpha, \cdots, \alpha)$ where α depends on $(n, p, q, \sigma^2)^T$. We run WWB1, WBB2 and BB-SSL for T iterations. Throughout the simulations we set $\sigma^2 = 1$ and assume the prior $\theta \sim B(1, p)$. The computational complexity of each algorithm is summarized in Table 1. The running times of these different algorithms for varying p and n are in Table 1 in Section C of the Appendix.

5.1 The Low-dimensional Case

Similarly to the experimental setting in Ročková (2018b), we generate $n = 50$ observations on $p = 12$ predictors with $\boldsymbol{\beta}_0 = (1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0)^T$, where the predictors have been grouped into 4 blocks. Within each block, predictors have an equal correlation and there is only one active predictor. All the other correlations are set to 0. We choose a single value[5] for $\lambda_0 \propto p$ and generate Dirichlet weights assuming $\alpha = 1$ (for WBB1, WBB2 and BB-SSL).

Uncorrelated Designs Assuming $\rho = 0, \lambda_0 = 12$ and $\lambda_1 = 0.05$ we run SSVS1 and Skinny Gibbs for $T = 10\,000$ iterations with a burn-in $B = 5\,000$. For WBB1, WBB2 and BB-SSL we use $T = 5\,000$ iterations. All methods perform very well under various metrics

[5]Note that SSL deploys an entire sequence of λ_0 values with warm starts

in this setting. We refer the reader to Section D.1 in the Appendix for details.

Correlated Designs Correlated designs are far more interesting for comparisons. We choose $\rho = 0.9, \lambda_0 = 7$ and $\lambda_1 = 0.15$ to deliberately encourage multimodality in the model posterior. For SSVS1 and Skinny Gibbs we set $T = 100\,000$ and $B = 5\,000$. For WBB1, WBB2 and BB-SSL we set $T = 95\,000$.

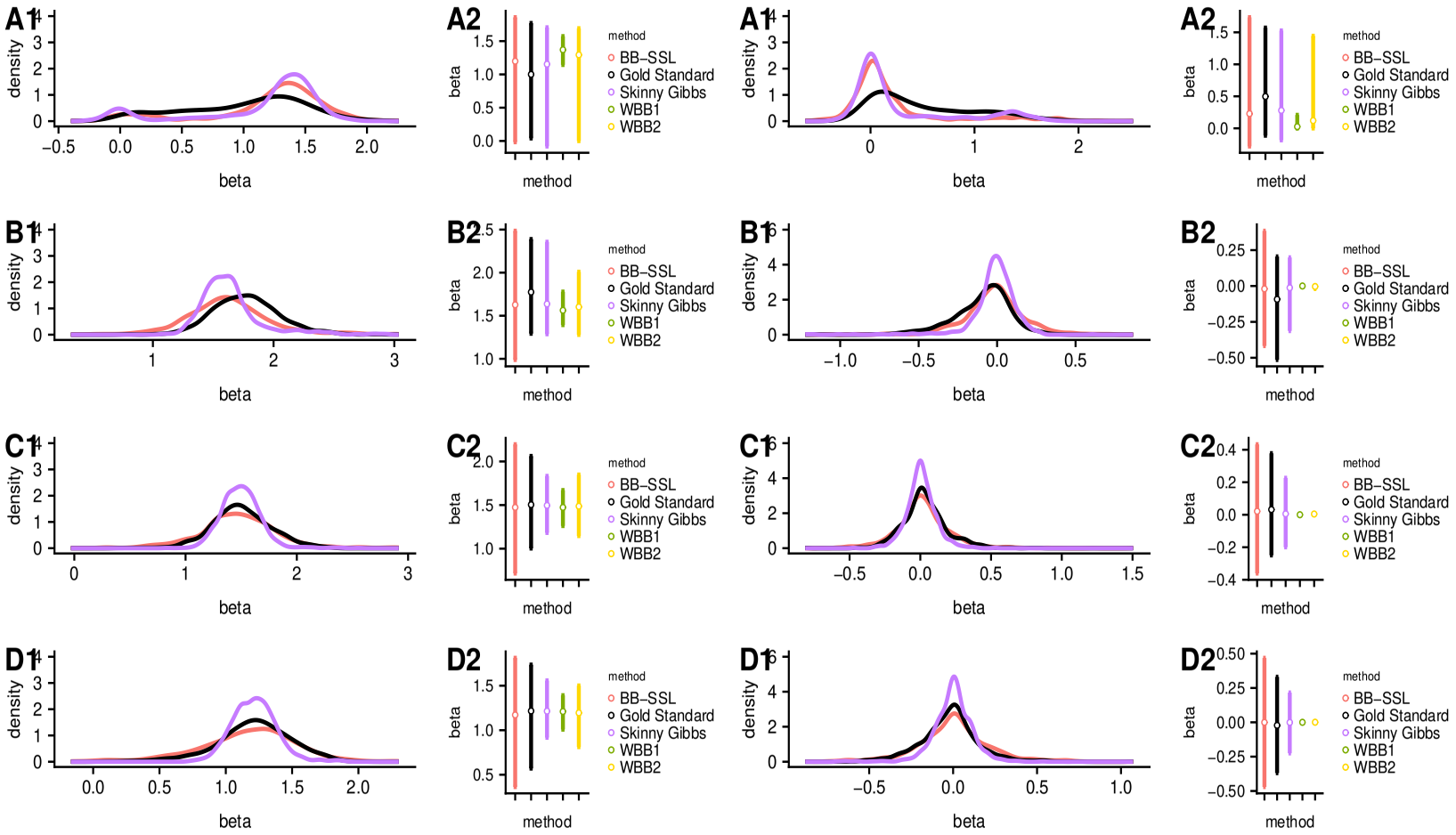
In terms of the marginal densities of β_i's, Figure 2 shows that BB-SSL tracks SSVS1 very closely. All methods, can cope with multi-collinearity where BB-SSL tends to have slightly longer credible intervals and where the opposite is true for Skinny Gibbs, WBB1 and WWB2. In terms of the marginal means of γ_i (Figure 3a) all methods perform well, where the median probability model rule (truncating the marginal means at 0.5) yields the true model. In terms of the overall posterior $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y}^{(n)})$, we identify over 60 unique models using SSVS1 where the true model accounts for most of the posterior mass. In Figure 3b, we show the visited (blue triangle) and not visited (red dots) among these models, where y-axis represents the estimated posterior probability for each model (calculated from SSVS1). All methods, can detect the dominating model. BB-SSL tracked down 99% of the posterior probability, followed by WBB1 (92%), WBB2 (91%) and Skinny Gibbs (73%).

5.2 The High-dimensional Case

We set $n = 100$ and $p = 1\,000$, where the active predictors have regression coefficients $(1, 2, -2, 3)^T$. Predictors are grouped into blocks of size 10, where each group has exactly one active coordinate and where predictors have a within-group correlation ρ. We set $\lambda_0 = 50, \lambda_1 = 0.05$ and $\alpha = 2$ for BB-SSL. For SSVS1 and Skinny Gibbs we set $T = 15\,000$ and $B = 5\,000$ while for WBB1, WBB2 and BB-SSL we set $T = 1\,000$.

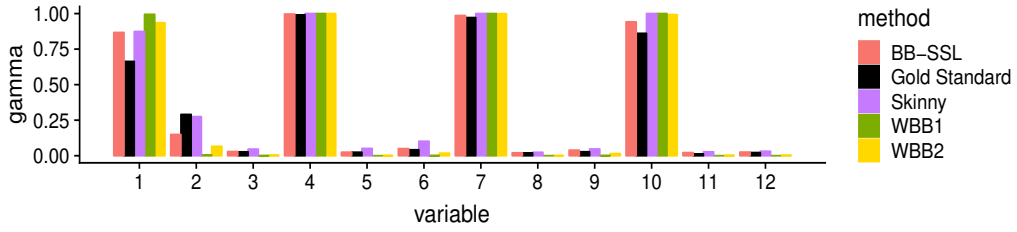
We consider 4 correlation settings, $\rho \in \{0, 0.6, 0.9\}$ and an extreme case $\rho = 0.99$ included in Section D.3 in the Appendix. In the independent ($\rho = 0$) setting, we fit SSL using a single value λ_0. In the correlated settings ($\rho = 0.6, 0.9, 0.99$), we fit SSL with λ_0 in a sequence of length 50 starting at 0.05 and ending at 50.

Since the results are similar for the three cases with $\rho \in \{0, 0.6, 0.9\}$, for brevity we

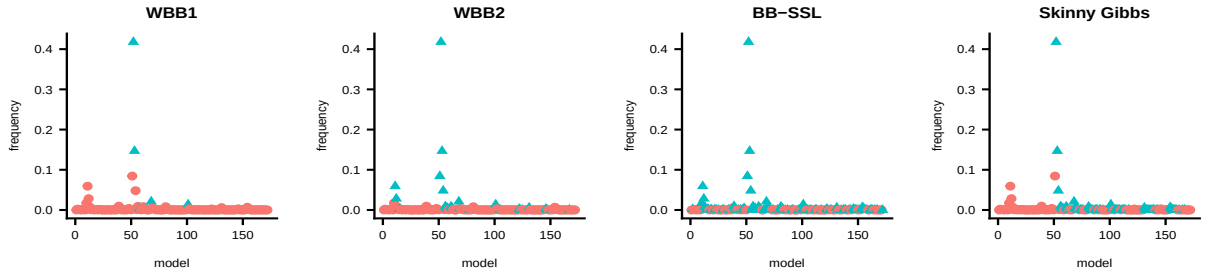


(a) Active predictors, from top to bottom: $\beta_1, \beta_4, \beta_7, \beta_{10}$ (b) Inactive predictors, from top to bottom: $\beta_2, \beta_5, \beta_8, \beta_{11}$

Figure 2: Estimated posterior density (left panel) and credible intervals (right panel) of β_i's in the low-dimensional correlated case. We have $n = 50, p = 12, \beta_{active} = (1.3, 1.3, 1.3, 1.3)', \lambda_0 = 7, \lambda_1 = 0.15, \rho = 0.9$. Each method has 5 000 sample points (after thinning for SSVS and Skinny Gibbs). BB-SSL is fitted using a single value $\lambda_0 = 7$. Since WBB1 and WBB2 produce a point mass at zero, we exclude them from density comparisons.

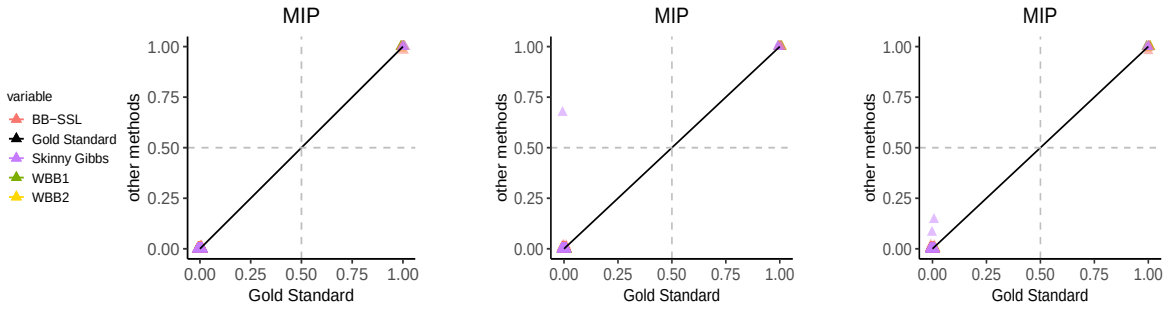


(a) Marginal inclusion probability.



(b) Posterior exploration plot. Blue (red) triangles are visited (unvisited) models.

Figure 3: The low-dimensional correlated case with $n = 50, p = 12, \beta_{active} = (1.3, 1.3, 1.3, 1.3)'$ where predictors are grouped into 4 correlated blocks with $\rho = 0.9$. We choose $\lambda_0 = 7, \lambda_1 = 0.15$.



(a) $\rho = 0$ (b) $\rho = 0.6$ (c) $\rho = 0.9$

Figure 4: Mean of γ_i's (marginal inclusion probability) in high-dimensional settings. $n = 100, p = 1000, \boldsymbol{\beta}_{active} = (1, 2, -2, 3)$ and predictors are grouped into 100 blocks. Each block has one active predictor. We set $\lambda_0 = 50, \lambda_1 = 0.05$. For (a), SSL is fitted using a single value $\lambda_0 = 50$. For (b) and (c), SSL is fitted using a sequence of λ_0's which starts at 0.05, ends at 50 and has length 50.

Metric			Skinny Gibbs	WBB1	WBB2	BB-SSL
β_i's	KL divergence	active	0.08	0.04	0.06	**0.004**
		inactive	0.02	3.09	3.09	**0.008**
	MSE	active	$\mathbf{1.77 \times 10^{-4}}$	2.00×10^{-4}	4.00×10^{-4}	6.50×10^{-4}
		inactive	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
γ_i's	MSE	active	**0**	**0**	$< 10^{-4}$	$< 10^{-4}$
		inactive	**0**	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Selected Model	Hamming distance	all	**0**	**0**	**0**	**0**

(a) $n = 100, p = 1000, \rho = 0$.

Metric			Skinny Gibbs	WBB1	WBB2	BB-SSL
β_i's	KL divergence	active	0.20	0.12	0.11	**0.03**
		inactive	0.02	3.09	3.09	**0.007**
	MSE	active	4.86×10^{-3}	1.29×10^{-3}	$\mathbf{1.11 \times 10^{-3}}$	1.36×10^{-3}
		inactive	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
γ_i's	MSE	active	**0**	**0**	$< 10^{-4}$	$< 10^{-4}$
		inactive	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
Selected Model	Hamming distance	all	1	**0**	**0**	**0**

(b) $n = 100, p = 1000, \rho = 0.6$.

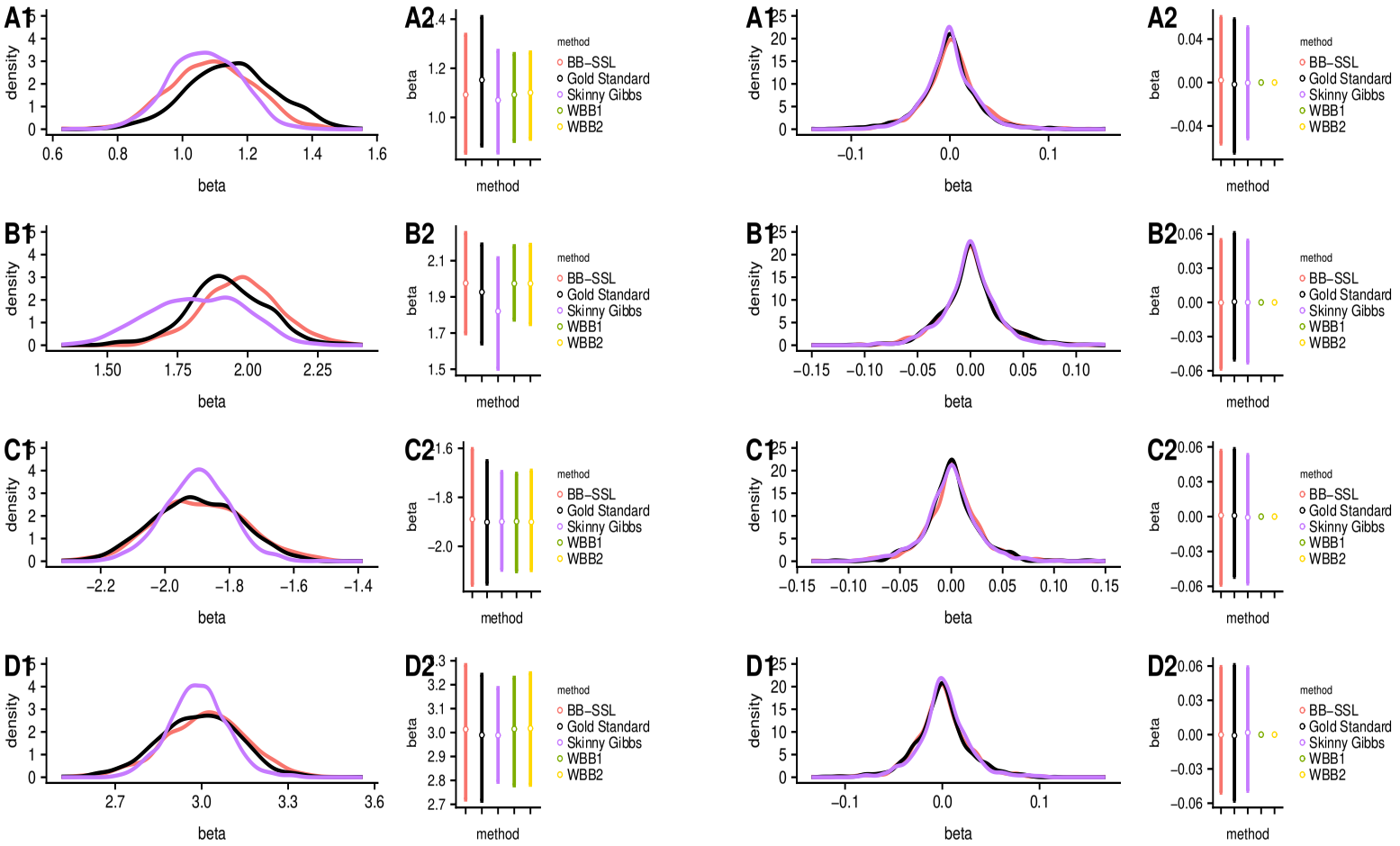
Metric			Skinny Gibbs	WBB1	WBB2	BB-SSL
β_i's	KL divergence	active	0.19	0.26	0.26	**0.02**
		inactive	0.10	3.09	3.09	**0.002**
	MSE	active	1.65×10^{-3}	$\mathbf{1.09 \times 10^{-3}}$	1.36×10^{-3}	1.50×10^{-3}
		inactive	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
γ_i's	MSE	active	**0**	**0**	$< 10^{-4}$	$< 10^{-4}$
		inactive	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$	$\mathbf{< 10^{-4}}$
Selected Model	Hamming distance	all	**0**	**0**	**0**	**0**

(c) $n = 100, p = 1000, \rho = 0.9$.

Table 2: Evaluation of the quality of posterior approximation in the high-dimensional setting. The best performance is marked in bold font.

only show the results for $\rho = 0.6$ with the rest postponed until the Appendix. In terms of the marginal density of β_i's (shown in Figure 5), Skinny Gibbs tends to underestimate the variance for active coordinates and WBB1 and WBB2 produce a point mass at 0 for inactive coordinates. BB-SSL, on the other hand, fares very well. Figure 4 shows that BB-SSL, WBB1 and WBB2 accurately reproduce the MIPs, while Skinny Gibbs tends to slightly overestimate the MIP as ρ increases.

To better quantify the performance of each method, we gauge the quality of posterior approximation under various metrics in Table 2. The KL divergence is calculated using an



(a) Active predictors, from top to bottom: $\beta_1, \beta_{11}, \beta_{21}, \beta_{31}$ (b) Inactive predictors, from top to bottom: $\beta_2, \beta_{12}, \beta_{22}, \beta_{32}$

Figure 5: Estimated posterior density (left panel) and credible intervals (right panel) of β_i's in the high-dimensional correlated case ($\rho = 0.6$). We have $n = 100, p = 1000, \beta_{active} = (1, 2, -2, 3)', \lambda_0 = 50, \lambda_1 = 0.05$. Each method has 5 000 sample points (after thinning for SSVS and Skinny Gibbs). BB-SSL is fitted using a an equal difference sequence of λ_0's, which is of length 50, starts at 0.05 and ends at 50. Since WBB1 and WBB2 produce a point mass at zero, we exclude them from density comparisons.

R package "FNN" (Beygelzimer et al. (2013)), where all parameters are set to their default values. The Hamming distance is calculated using an R package "e1071" (Meyer et al. (2019)). All methods do well in terms of MIP and the selected model (based on the median probability model rule). For β_i's, all methods estimate the mean accurately. Taking into account the shape of the posterior for β_i's, the performance is divided among coordinates and methods. For all methods, the approximability of active coordinates is less accurate than for the inactive ones. And we find in the settings we tried, the different methods can be ranked as follows: BB-SSL $>$ Skinny Gibbs $>$ WBB1 $\approx$ WBB2.

Conclusion We found that BB-SSL is a reliable approximate method for posterior sampling that achieves a close-to-exact (SSVS) performance but is computationally cheaper. Additional speedups can be obtained with parallelization. In addition, while MCMC-based methods are sensitive of the initialization and can fall into a local trap, BB-SSL seems to be less susceptible to this problem.

6 Data Analysis

6.1 Life Cycle Savings Data

The Life Cycle Savings data (Belsley et al., 2005) consists of $n = 50$ observations on $p = 4$ highly correlated predictors: "pop15" (percentage of population under 15 years old), "pop75" (percentage of population over 75 years old), "dpi" (per-capita disposable income), "ddpi" (percentage of growth rate of dpi). According to the life-cycle savings hypothesis proposed by Ando and Modigliani (1963), the savings ratio (y) can be explained by these four predictors and a linear model can be used to model their relationship.

We preprocess the data in the following way. First, we standardize predictors so that each column of $\boldsymbol{X}$ is centered and rescaled so that $||\boldsymbol{X}_j||_2 = \sqrt{n}$. Next, we estimate the noise variance σ^2 using an ordinary least squares regression. We then divide y by the estimated noise standard deviation and estimate θ by fitting SSL with $\lambda_0 = 20, \lambda_1 = 0.05$. For BB-SSL we set $\alpha = 2 \log \frac{(1-\theta)\lambda_0}{\theta\lambda_1} \approx 14$ and set $a = 1$, $b = 4$. We run SSVS1 and Skinny

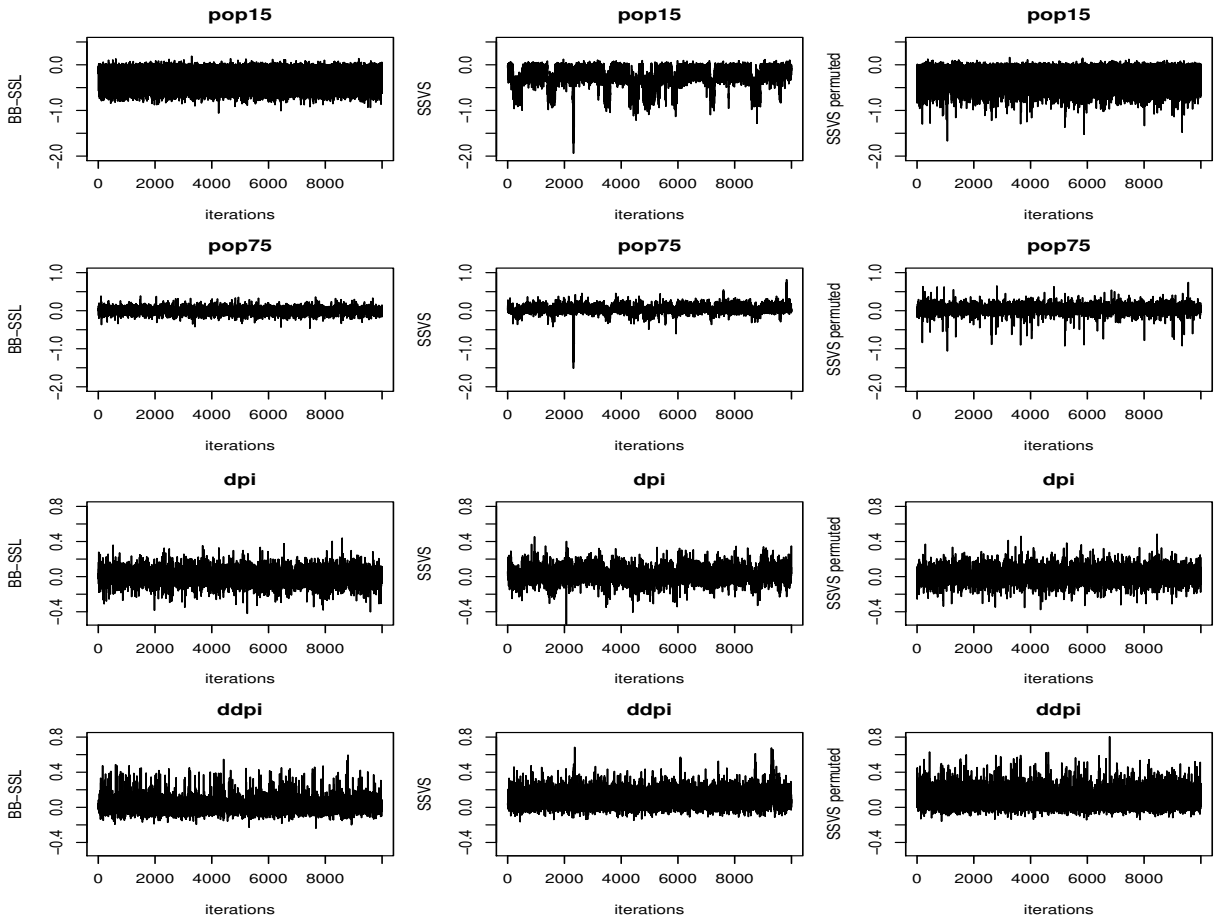
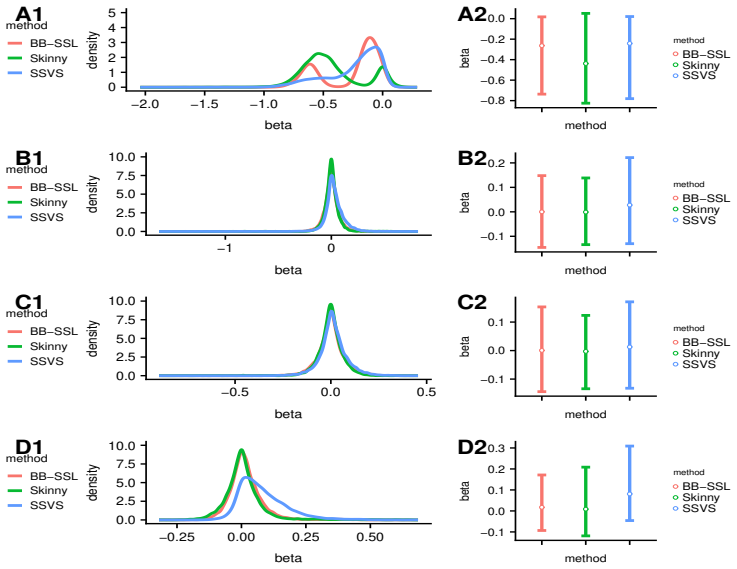
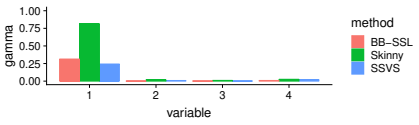


Figure 6: Trace plots for the Life Cycle Savings data. We choose $\lambda_0 = 20, \lambda_1 = 0.05$. The first column is the BB-SSL traceplot with weight distribution $\alpha = 2 \log \frac{(1-\theta)\lambda_0}{\theta \lambda_1} = 14$, the second column is thinned SSVS traceplot chain with a LASSO initialization (regularization parameter chosen by cross validation). The third column is the same SSVS chain only with samples permuted.



(a) Density of β_i's..



(b) Comparison of mean of γ_i's.

Figure 7: Plots for Life Cycle Savings Data. We choose $\lambda_0 = 20, \lambda_1 = 0.05$. SSVS is initialized at the LASSO solution (with the regularization parameter chosen by cross-validation). The weight distribution for BB-SSL uses $\alpha = 2 \log \frac{(1-\theta)\lambda_0}{\theta\lambda_1} = 14$.

	SSVS	Skinny Gibbs	BB-SSL
Effective sample size	2716	11188	15000

Table 3: Average effective sample size (out of 15 000 samples) for Life Cycle Saving Data. Effective sample size is calculated using R package coda (Plummer et al. (2006)).

Gibbs for $T = 100\,000$, $B = 5\,000$ and BB-SSL for $T = 10\,000$.

Figure 6 shows the trace plots on the four predictors. BB-SSL (first column) has the same mean and spread as SSVS (third column). We also observe that raw samples from SSVS (second column) are correlated, so more iterations are needed in order to fully explore the posterior. In contrast, each sample from BB-SSL is independent and thereby fewer samples will be needed in practice. See Table 3 for effective sample size comparisons. Figure 7a shows the marginal density of β_i's and 7b shows the marginal mean of γ_i's. In both figures BB-SSL achieves good performance.

6.2 Durable Goods Marketing Data Set

Our second application examines a cross-sectional dataset from Ni et al. (2012) (ISMS Durable Goods Dataset 2) consisting of durable goods sales data from a major anonymous U.S. consumer electronics retailer. The dataset features the results of a direct-mail promotion campaign in November 2003 where roughly half of the $n = 176\,961$ households received a promotional mailer with 10$ off their purchase during the promotion time period (December 4-15). The treatment assignment ($tr_i = \mathbb{I}\,(\text{promotional mailer}_i)$) was random. The data contains 146 descriptors of all customers including prior purchase history, purchase of warranties etc. We will investigate the effect of the promotional campaign (as well as other covariates) on December sales. In addition, we will interact the promotion mail indicator with customer characteristics to identify the "mail-deal-prone" customers. To be more specific, we adopt the following model

$$Y_i = \alpha \times tr_i + \boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{\gamma} \times tr_i \times \boldsymbol{x}_i + \epsilon_i \tag{20}$$

where $\boldsymbol{x}_i$ refers to the 146 covariates, tr_i is the treatment assignment, and the noise ϵ_i is iid normally distributed. And our aim is to (1) estimate the coefficients $\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}$; (2) identify those customers with $\mathbb{E}\,[Y_i \,|\, \boldsymbol{x}_i, tr_i = 1] > \mathbb{E}\,[Y_i \,|\, \boldsymbol{x}_i, tr_i = 0]$.

For preprocessing, we first remove all variables that contain missing values or that are all 0's. We also create new predictors by interacting the treatment effect with the descriptor variables. After that the total number of predictors becomes $p = 273$. We standardize $\boldsymbol{X}$ such that each column has a zero mean and a standard deviation $\sqrt{n}$ and we use the maximum likelihood estimate of the standard deviation to rescale the outcome. We run BB-SSL for $T = 1\,000$ iterations and SSVS1 for $T = 20\,000$ iterations with a $B = 1\,000$ burnin period, initializing MCMC at the origin. We set $\lambda_0 = 100, \lambda_1 = 0.05, a = 1, b = p$. Estimating $\hat{\theta} = \frac{\#\text{ of selected variables}}{p}$ by fitting the Spike-and-Slab LASSO, we then set $\alpha = 2 \log \frac{(1-\hat{\theta})\theta_0}{\hat{\theta}\lambda_1}$.

Figure 9 depicts estimated posterior density of selected coefficients in the model (20), showing that BB-SSL estimation is very close to the gold standard (SSVS). Further, BB-SSL identified 67.3% of customers as "mail-deal-prone", reaching accuracy 98.2% and a

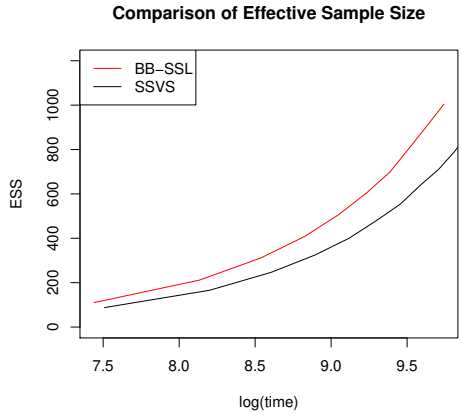


Figure 8: Effective sample size comparison for ISMS Durable Goods Dataset 2. We choose $\lambda_0 = 100, \lambda_1 = 0.05$. Red line is BB-SSL with $\alpha = 2 \log \frac{(1-\theta)\lambda_0}{\theta\lambda_1} \approx 15$ and black line is SSVS initialized at origin.

false positive rate 2.1% (treating SSVS estimation as the truth). Despite the comparable performance to SSVS, BB-SSL is advantageous in terms of computational efficiency. As shown in Figure 8, within the same amount of time, BB-SSL obtains more effective samples compared with SSVS and its advantage becomes even more significant as time increases. This experiment confirms our hypothesis that BB-SSL has a great potential as an approximate method for large datasets.

7 Discussion

In this paper we developed BB-SSL, a computational approach for approximate posterior sampling under Spike-and-Slab LASSO priors based on Bayesian bootstrap ideas. The fundamental premise of BB-SSL is the following: replace sampling from conditionals (which can be costly when either n or p are large) with fast optimization of randomly perturbed (reweigthed) posterior densities. We have explored various ways of performing the perturbation and looked into asymptotics for guidance about perturbing (weighting) distributions. We have concluded that with suitable conditions on the weights distribution, the pseudo-posterior distribution attains the same rate as the actual posterior in high-dimensional estimation problems (sparse normal means and high-dimensional regression). These theo-

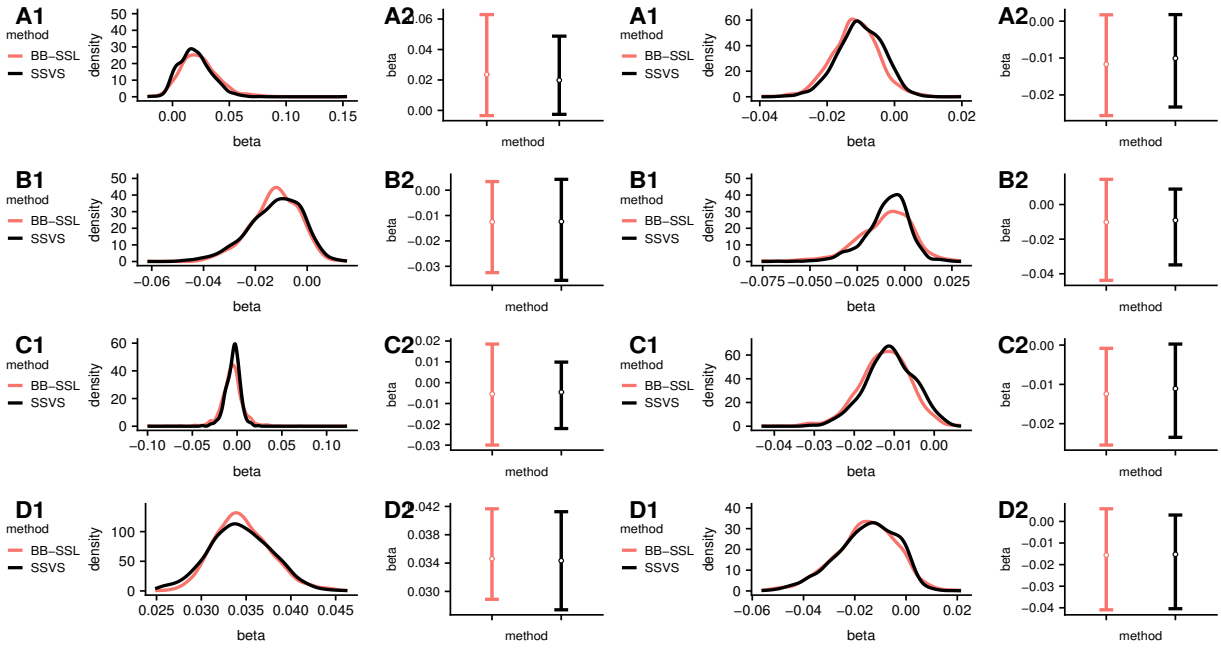


Figure 9: Posterior density and credible intervals for the selected β_i's. From left to right, top to bottom they correspond to "S-SAL-TOT60M", "S-U-CLS-NBR-12MO", "PH-HOLIDAY-MAILER-RESP-SA", "PROMO-NOV-SALES", "S-SAL-FALL-24MO", "S-TOT-CAT × treatment", "C-ESP-RECT × treatment", "S-CNT-TOT24M × treatment". We set $\lambda_0 = 100, \lambda_1 = 0.05, a = 1, b = 273$. We set $\alpha = 2 \log \frac{(1-\theta)\lambda_0}{\theta \lambda_1} \approx 15$.

retical results are reassuring and significantly extend existing knowledge about Weighted Likelihood Bootstrap (Newton et al. (2018)), which was shown to be consistent for iid data in finite-dimensional problems. We have shown in simulations and on real data that BB-SSL can approximate the true posterior well and can be computationally beneficial.

8 Supplementary Materials

appendix: File "appendix" containing proofs, discussion of connections to NPL, details of computational complexity analysis and additional experimental results mentioned in the article. (pdf file)

code: File "code" containing R scripts to perform the simulations and experiments described in the article. (zipped file)

References

Ando, A. and Modigliani, F. (1963). The "life cycle" hypothesis of saving: Aggregate implications and tests. *The American economic review*, 53(1):55–84.

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2020). Spike-and-slab group lassos for grouped regression and sparse generalized additive models. *Journal of the American Statistical Association*, (just-accepted):1–41.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons.

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013). FNN: fast nearest neighbor search algorithms and applications. *R package version*, 1(1).

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042.

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(512):1479–1490.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Deshpande, S. K., Ročková, V., and George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, 28(4):921–931.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The annals of applied statistics*, 6(4):1971.

Fong, E., Lyddon, S., and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. *arXiv:1902.03175*.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373.

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

Ishwaran, H. and Rao, J. S. (2005). Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association*, 100(471):764–780.

Johndrow, J. E., Orenstein, P., and Bhattacharya, A. (2017). Bayes shrinkage at GWAS scale: Convergence and approximation theory of a scalable MCMC algorithm for the horseshoe prior. *arXiv:1705.00841*.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Li, Z., Mccormick, T., and Clark, S. (2019). Bayesian joint spike-and-slab graphical lasso. In *International Conference on Machine Learning*, pages 3877–3885. PMLR.

Lyddon, S., Walker, S., and Holmes, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems*, pages 2071–2081.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C., and Meyer, M. D. (2019). Package 'e1071'. *The R Journal*.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moran, G. E., Ročková, V., and George, E. I. (2018). Variance prior forms for high-dimensional Bayesian variable selection. *Bayesian Analysis*, pages 1091–1119.

Narisetty, N. N., Shen, J., and He, X. (2019). Skinny Gibbs: A consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association*, 114(527):1205–1217.

Newton, M., Polson, N. G., and Xu, J. (2018). Weighted Bayesian Bootstrap for Scalable Bayes. *arXiv:1803.04559*.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26.

Ni, J., Neslin, S. A., and Sun, B. (2012). Database submission—The ISMS durable goods data sets. *Marketing Science*, 31(6):1008–1013.

Papandreou, G. and Yuille, A. L. (2010). Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems*, pages 1858–1866.

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

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11.

Ročková, V. (2018a). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437.

Ročková, V. (2018b). Particle EM for variable selection. *Journal of the American Statistical Association*, 113(524):1684–1697.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438.

Shin, M. and Liu, J. S. (2018). Neuronized priors for Bayesian sparse linear regression. *arXiv:1810.00141*.

Tang, Z., Shen, Y., Li, Y., Zhang, X., Wen, J., Qian, C., Zhuang, W., Shi, X., and Yi, N. (2018). Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics*, 34(6):901–910.

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017). The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics*, 33(18):2799–2807.

Vivo, P., Majumdar, S. N., and Bohigas, O. (2007). Large deviations of the maximum eigenvalue in Wishart random matrices. *Journal of Physics A: Mathematical and Theoretical*, 40(16):4317.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Xu, M., Lakshminarayanan, B., Teh, Y. W., Zhu, J., and Zhang, B. (2014). Distributed Bayesian posterior sampling via moment sharing. In *Advances in Neural Information Processing Systems*, pages 3356–3364.

Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593.