Bayesian Bootstrap Spike-and-Slab LASSO

Lizhen Nie\textsuperscript{a} and Veronika Ročková\textsuperscript{b}

\textsuperscript{a}Department of Statistics, University of Chicago, Chicago, IL; \textsuperscript{b}Booth School of Business, University of Chicago, Chicago, IL

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. 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 very well in these comparisons, often providing substantial computational gains. Supplementary materials for this article are available online.

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 (Mitchell and Beauchamp 1988; George and McCulloch 1993). 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 article focuses on the canonical linear regression model, where a vector of responses $Y = (Y_1, \ldots, Y_n)^T$ is stochastically linked to fixed predictors $x_i \in \mathbb{R}^p$ through

$$Y_i = x_i^T \beta_0 + \epsilon_i \quad \text{with} \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{for} \quad 1 \leq i \leq n,$$

where $\sigma^2 > 0$ and where $\beta_0 \in \mathbb{R}^p$ is a possibly sparse vector of regression coefficients. In this work, we assume that $\sigma^2$ is known and we refer to Moran et al. (2019) for elaborations with an unknown variance. We assume that the vector $Y$ and the regressors $X = [X_1, \ldots, X_p]$ have been centered and, thereby, we omit the intercept. In the presence of uncertainty about which subset of $\beta_0$ is in fact nonzero, one can assign a prior distribution over the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$ as well as the pattern of nonzeros $y = (y_1, \ldots, y_p)^T$ where $y_j \in [0, 1]$ for whether or not the effect $\beta_j$ is active. This formalism can be condensed into the usual spike-and-slab prior form

$$\pi(\beta | \gamma) = \prod_{j=1}^p \left[ \gamma_j \psi_0(\beta_j) + (1 - \gamma_j) \psi_1(\beta_j) \right],$$

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 toward 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 nonlocal 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). A separable variant of
the SSL prior is obtained by fixing the mixing weight $\theta$. We refer to Ročková (2018a) and Ročková and George (2018) for more discussion on separable (fixed $\theta$) and nonseparable (random $\theta$) SSL priors.

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; Hans 2009; Bottolo and Richardson 2010; Clyde, Ghosh, and Littman 2011; Welling and Teh 2011; Johndrow, Orenstein, and Bhattacharya 2020; Xu et al. 2014). More recently, several clever computational tricks have been suggested that avoid costly matrix inversions by using linear solvers (Bhattacharya, Chakraborty, and Mallick 2016) or by disregarding correlations between active and inactive coefficients (Narisetty, Shen, and He 2019). Neuronized priors have been proposed (Shin and Liu 2018) that offer computational benefits by using 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 (Carbonetto and Stephens 2012; Ročková and George 2014, 2018). Fast posterior mode detection is effective in structure discovery and data exploration, a little less so for inference. In this article, 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, Polson, and Xu (2020), Fong, Lyddon, and Holmes (2019) and Ng and Newton (2020). The main idea behind WLB is to perform approximate sampling by independently optimizing randomly perturbed likelihood functions. We extend the WLB framework by incorporating perturbations both in the likelihood and in the prior. The main contributions of this work are 2-fold. First, we introduce Bayesian Bootstrap Spike-and-Slab LASSO (BB-SSL), 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 largely concentrated on consistency statements in low dimensions for iid data (Newton and Raftery 1994). More recently, Ng and Newton (2020) established conditional consistency (asymptotic normality) in the context of LASSO regression for a fixed dimensionality and model selection consistency for a growing dimensionality. Our theoretical results also 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. BB-SSL is (a) unapologetically parallelisable, and (b) it does not require costly matrix inversions (due to its coordinate-wise optimization nature), thereby having the potential to meet the demands of large datasets.

The structure of this article 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 article with a discussion in Section 7.

1.1. Notation

With $\phi(y; \mu, \sigma^2)$ we denote the Gaussian density with a mean $\mu$ and a variance $\sigma^2$. We use $\rightarrow$ to denote convergence in distribution. We write $a_n = O_p(b_n)$ if for any $\epsilon > 0$, there exist finite $M > 0$ and $N > 0$ 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 $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. 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 $X_A$ a sub-matrix consisting of columns of $X$’s indexed by a subset $A \subset \{1, \ldots, p\}$ and with $P_A$ the orthogonal projection to the range of $X_A$ (Zhang and Zhang 2012), that is, $P_A = X_A^* X_A$, where $X_A^*$ is the Moore-Penrose inverse of $X_A$. We denote with $\|X\|$ the matrix operator norm of $X$.

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 nonconvex penalized likelihood methods. Various SSL incarnations have spawned since its introduction, including a version for group shrinkage (Tang et al. 2018; Bai et al. 2020), survival analysis (Tang et al. 2017), varying coefficient models (Bai et al. 2020) and/or Gaussian graphical models (Despande, Ročková, and George 2019; Li, Mccormick, and Clark 2019). The original procedure proposed for Gaussian regression targets a posterior mode

$$
\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \phi(Y_i; \begin{bmatrix} X_i \hat{X}_i \end{bmatrix}^T \beta; \sigma^2) \times \int_{\theta} \prod_{j=1}^{p} \pi(\beta_j | \theta) d\pi(\theta) \right\},
$$

where $\pi(\beta_j | \theta) = \theta \psi_1(\beta_j) + (1 - \theta) \psi_0(\beta_j)$ is obtained from (2) by integrating out the missing indicator $\gamma_j$ and by deploying $\psi_1(\beta_j) = \lambda_1 e^{-\beta_j^2/2}$ and $\psi_0(\beta_j) = \lambda_0 e^{-\beta_j^2/2}$ with $\lambda_0 \gg$
Ročková and George (2018) develop a coordinate-ascent strategy which targets $\beta$ 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 $\hat{\beta}_j$ using an implicit equation\footnote{Here we are not necessarily assuming that $\|X\|_2^2 = n$ and the above formula is hence, slightly different from Theorem 3.1 of Ročková and George (2018).}

$$\hat{\beta}_j = \frac{1}{\|X\|_2^2} \|z_j - \sigma^2 \lambda_{0j}^2(\hat{\beta}_j)\|^2 + \text{sign}(z_j) \times 1(|z_j| > \Delta_j)$$

(4)

where $\Delta_j = \mathbb{E}[\theta | \hat{\beta}_j]$, $z_j = X_j^T(Y - X \hat{\beta}_j)$ and $\Delta_j = \inf_{\rho > 0} (\|X\|^2_2 t/2 - \sigma^2 \rho(t(\hat{\beta}_j)/t)$ with $\rho(t(\theta) = -\lambda_1|t| + \log(p_n^0(t)/p_n^*(t))$, where

$$p_n^*(t) = \frac{\theta \psi_1(t)}{\theta \psi_0(t)} + (1 - \theta) \psi_0(t)$$
$$\lambda_0^*(t) = \lambda_1 p_n^*(t) + \lambda_0 (1 - p_n^*(t)).$$

(5)

Ročková and George (2018) also provide fast updating schemes for $\Delta_j$ and $\hat{\beta}_j$. 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:

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

(6)

where $\tau^{-1} = (1/\tau_1^2, \ldots, 1/\tau_p^2)^T$ is the vector of variances. The conditional conjugacy of the SSL prior enables direct Gibbs sampling for $\beta$ (see Algorithm 1 below). 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, Chakraborty, and Mallick (2016) proposed a clever trick. By recasting the sampling step as a solution to a linear system, one can circumvent a Cholesky factorization which would otherwise have a complexity $O(np^2)$ per iteration. Building on this development, Johndrow, Orenstein, and Bhattacharya (2020) developed a blocked Metropolis-within-Gibbs algorithm to sample from horseshoe posteriors (Carvalho, Polson, and Scott 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, Shen, and He (2019) also bypasses large matrix inversions by independently sampling from active and inactive $\beta_i$’s. While the method is only approximate, it has a rather favorable computational complexity $O(np)$.

A referee suggested another Gibbs sampler implementation with a complexity $O(np)$ which can be obtained by updating $(\hat{\beta}_j, \gamma_j)$ one at a time while conditioning on the remaining $(\hat{\beta}_j, \gamma_j)$’s (Geweke 1991). While this implementation is very fast for point-mass spikes, the Spike-and-Slab LASSO prior requires sampling from a half-normal distribution which can be inefficient in practice. One-site Gibbs samplers also generally lead to slower mixing due to increased autocorrelation. In simulations, we find the performance of this method to be comparable with SSVS using Bhattacharya, Chakraborty, and Mallick’s (2016) trick. The detailed description of this algorithm is included in the Section C, supplementary material.

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 $\pi(\beta | Y)$ 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 $\hat{\beta}_t$ by first sampling random weights $w_t = (w_{t1}, w_{t2}, \ldots, w_{tn})^T$ from some weight distribution $\pi(w)$ and then maximizing a
reweighted likelihood
\[ \tilde{\beta}_i = \arg \max_{\beta} \tilde{L}^w_i(\beta, \sigma^2; X^{(n)}, Y^{(n)}) \] (6)

where
\[ \tilde{L}^w_i(\beta, \sigma^2; X^{(n)}, Y^{(n)}) = \prod_{i=1}^{n} \phi(Y_i; X_i^T \beta; \sigma^2)^{w_i}. \]

Newton and Raftery (1994) argue that for certain weight distributions \( \pi(w) \), the conditional distribution of \( \tilde{\beta} \) is given the data can provide a good approximation to the posterior distribution of \( \beta \). 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, the WLB posterior is "consistent" (i.e., concentrating on any arbitrarily small neighborhood around MLE) and asymptotically first-order correct (normal with the same centering) for almost every realization of the data. Note that this Bayesian notion of consistency is fundamentally different from typical bootstrap consistency statements which usually refer to the convergence of the bootstrap distribution to the sampling distribution (see, e.g., Præstgaard and Wellner 1993 and Remark 4.2 below). 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 Jeffreys' prior. However, for more general prior distributions (such as shrinkage priors considered here), the correspondence between the prior \( \pi(\beta) \) and \( \pi(w) \) 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, Polson, and Xu (2020) suggested blending the prior directly into WLB by including a weighted prior term, that is, replacing (6) with
\[ \tilde{\beta}_i = \arg \max_{\beta} \tilde{L}^w_i(\beta, \sigma^2; X^{(n)}, Y^{(n)}) \pi(\beta)^{w_i}, \]

where \( w_i \) id \( \exp(1)^2 \). This so called Weighted Bayesian Bootstrap (WBB) method treats the prior weight \( w_i \) 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 \( \pi(\beta) \) 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 (Park and Casella 2008; Hans 2009). Deploying the WBB method thus, does not guarantee that uncertainty be properly 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
\[ y_i = \beta_i^0 + \epsilon_i / \sqrt{n} \quad \text{for} \quad i = 1, 2, \ldots, n. \] (7)

Under the separable SSL prior (i.e., \( \theta \) fixed), the true posterior is a mixture
\[ \pi(\beta_i | y_i) = w_1 \pi(\beta_i | y_i, \gamma_i = 1) + w_0 \pi(\beta_i | y_i, \gamma_i = 0) \] (8)

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

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 Proposition 1, supplementary material that \( w_0 \to 0 \) and \( w_1 \to 1 \). The true posterior \( \pi(\beta_i | y_i) \) is hence, dominated by the component \( \pi(\beta_i | y_i, \gamma_i = 1) \), which takes the following form
\[ \pi(\beta_i | y_i, \gamma_i = 1) = \frac{\mathbb{I}(\beta_i \geq 0) \phi(-1/2) + \mathbb{I}(\beta_i < 0) \phi(1/2)}{\phi(1/2) - \phi(-1/2)} \]

where
\[ \phi^{(-)}(x) = \phi(x; y_i - 1/n, 1/n) \quad \text{and} \quad \phi^{(+)}(x) = \phi(x; y_i + 1/n, 1/n). \] (9)

Intuitively, \( \lambda_1/n \) vanishes when \( n \) is large, so both \( \phi^{(-)}(\beta_i) \) and \( \phi^{(+)}(\beta_i) \) will be close to \( \phi(\beta_i; y_i, 1/n) \). This intuition is proved rigorously in the Section A.6.2, supplementary material, 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 \( \pi(\sqrt{n}(\beta_i - y_i) | y_i, \gamma_i = 1) \) 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 \( \hat{\beta}_i \) equals
\[ \hat{\beta}_i = \begin{cases} 0, & \text{if } |y_i| \leq \Delta_{w_i} \\ \left[ |y_i| - \frac{1}{w_i} \lambda^* \hat{\beta}_i \right]_+, & \text{sign}(\sqrt{w_i} y_i), \text{otherwise} \end{cases} \] (11)

where \( \Delta_{w_i} = \inf_{t \geq 0} |t/2 - \rho(t/\theta)/(nw_i t)| \) is the analogue of \( \Delta_t \) defined below (4) for the regression model and where \( \rho(t/\theta) \) was also defined below (4). When \( \hat{\beta}_i \neq 0 \), we show in the Section A.6.2, supplementary material that \( n(\hat{\beta}_i - y_i) \to -1/w_1 \lambda_1 \). Under the condition \( |y_i| > |\beta_i^0|/2 > 0 \), it can be shown (Section A.6.2) that \( P_{w_i}(\hat{\beta}_i = 0 | y_i) \to 0 \). For active coordinates, the distribution of the WBB samples \( \hat{\beta}_i \) is thus, purely determined by that of \(-1/w_1 \lambda_1 \). The shape of this posterior can be very different from the standard normal one, as can be seen from Figure 1. In particular, Figure 1(a) shows how WBB (a) assigns a nonnegligible prior mass to zero (in spite of evidence of signal)
In fact, under the uniform Dirichlet distribution, the marginal distribution

\[ y_i = x_i + \epsilon_i / \sqrt{n} \]

with \( n = 10 \). 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 with \( \lambda_0 = 10 \).

(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 \( |y_i| = O_p(1/n) \) from (7). The following arguments will be under the assumption \( |y_i| \approx \frac{1}{\sqrt{n}} \). One can show (Section A.6.3, supplementary material) that the true posterior \( \pi(\beta_i | y_i) \) is dominated by the component \( \pi(\beta_i | y_i, y_i = 0) \) since \( w_0 \rightarrow 1 \) and \( w_1 \rightarrow 0 \). When \( n \) is sufficiently large, one can then approximate this distribution with the Laplace spike, indeed \( \pi(\lambda_0 \beta_i | y_i, y_i = 0) \) converges to \( \frac{1}{2} e^{-|\lambda_0 \beta_i|} \) in total variation (Section A.6.3, supplementary material). 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 (Section A.6.3, supplementary material) that the posterior converges to a point mass at 0, that is, \( P_{\beta_i}(\beta_i = 0 | y_i) \rightarrow 1 \). This is a misleading approximation of the actual posterior (Figure 1(b)). To conclude, since SSL is always shrinking the estimates toward 0, WBB samples will often be true. The zero posterior, however, follows roughly a spike 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, Polson, and Xu (2020), we argue that the random perturbation should affect both the prior and the data. Instead of inflating the prior contribution by a 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 \( \pi(\beta | y) \) in (2) which is centered around zero, we consider a variant that uses hierarchical jittered Laplace distributions.

3In fact, under the uniform Dirichlet distribution, the marginal distribution becomes \( w_i \sim \text{Beta}(1, n-1) \). Since \( n \times \text{Beta}(1, n-1) \vdash \text{Gamma}(1, 1) \), 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.
\[ \hat{\beta}^*_i = \arg \max_{\beta \in \mathbb{S}^p} \left\{ \tilde{L}^{w_i=1,\ldots,1}(\beta^*, \sigma^2; X^*, Y^*) \times \int \tilde{\pi}(\beta \mid 0, \theta) d\tilde{\pi}(\theta) \right\}, \]

and then get \( \hat{\beta}_i \) through a post-processing step \( \hat{\beta}_i = \hat{\beta}^*_i + \mu_t \).

### 4.1. Theory for BB-SSL

This section is dedicated to understanding the asymptotic behavior of posterior approximations obtained with BB-SSL when \( \theta \) is fixed. Asymptotic properties of the global model and the exact posterior for this separable SSL prior were characterized previously in Roˇcková (2018a) and Roˇcková and George (2018).

For the uniform Dirichlet weights, Newton and Raftery (1994) (Theorems 1 and 2) show first-order correctness, that is, 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, Polson, and Xu 2020; Ng and Newton 2020) 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. A more recent theoretical development in this direction is the work of Ng and Newton (2020) who find that WBB first-order correctness holds for a wide class of random weight distributions in low-dimensional LASSO regressions. They also theoretically assess the influence of assigning random weights to the penalty term. 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 \( \pi(w) \) 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 \( Y^{(n)} = (Y_1, \ldots, Y_n)^T \) is observed as a noisy version of a sparse mean vector \( \beta_0 = (\beta_0^1, \ldots, \beta_0^0)^T \), that is,

\[ Y_i = \beta_0^i + \epsilon_i, \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \quad \text{for} \quad 1 \leq i \leq n. \quad (14) \]

**Theorem 4.1 (Normal Means).** Consider the normal means model (14) with \( \|\beta_0\|_0 \) such that \( \|\beta_0\|_0 = o(n) \) as \( n \to \infty \). Assume the SSL prior with \( 0 < \lambda_1 < \frac{1}{2} \) and \( \theta \sim \left(\frac{3}{2}\right)^0, \lambda_0 \sim \left(\frac{3}{2}\right)^0 \) with \( \eta, \gamma > 0 \) such that \( \eta + \gamma > 1 \). Assume that \( w = (w_1, \ldots, w_n)^T \) are nonnegative and arise from \( \pi(w) \) 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 \( \mathbb{E} \left( \frac{1}{w_i} \right) \leq C_2 \) 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, that is,

\[ \lim_{n \to \infty} \mathbb{E}_{\beta_0} \mathbb{E}_{w, \mu} \left[ \|\hat{\beta}_w^\mu - \beta_0\|_2^2 \geq M_n n q \log \left( \frac{n}{q} \right) \right] = 0. \quad (15) \]

**Proof.** See Section A.1.2, supplementary material.

In Theorem 4.1, \( \hat{\beta}_w^\mu \) denotes the BB-SSL sample whose distribution, for each given \( Y^{(n)} \), is induced by random weights \( w \) arising from \( \pi(w) \) and random recentering \( \mu \) arising from \( \psi_0(\cdot) \). Despite the approximate nature of BB-SSL, the concentration rate (15) is minimax optimal and it is the same rate achieved by the actual posterior distribution under the same prior assumptions (Roˇcková 2018a). Condition (1) in Theorem 4.1 is not surprising and aligns with considerations in Newton, Polson, and Xu (2020). Conditions (2) and (3) can be viewed as regularizing the tail behavior of \( w_i \)’s (left and right, respectively). While Newton, Polson, and Xu (2020) 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 \( \pi(w) \).

**Corollary 4.1.** Assume the same model and prior as in Theorem 4.1. Next, when \( w = (w_1, w_2, \ldots, w_n)^T \sim n \times \text{Dir}(\alpha, \alpha, \ldots, \alpha) \) with \( \alpha \gtrsim \sigma^2 \log \left( \frac{1-\theta}{\theta \lambda_1} \right) \) or \( w_i \overset{\text{iid}}{\sim} \frac{1}{\sigma} \text{Gamma}(\alpha, 1) \) with \( \alpha \gtrsim \sigma^2 \log \left( \frac{1-\theta}{\theta \lambda_1} \right) \), the BB-SSL posterior satisfies (15).

**Proof.** See the Section A.2, supplementary material.

Theorem 4.1 and Corollary 4.1 give insights into which 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 nonparametric normal means model, the assumption \( w \sim n \times \text{Dir}(\alpha, \ldots, \alpha) \) for \( \alpha < 2 \) yields risk (for active coordinates) that can be arbitrarily large (as we show in Section A.3, supplementary material). 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) with rescaled columns \( \|X_j\|_2 = \sqrt{n} \) for all \( j = 1, 2, \ldots, p \).

**Theorem 4.2 (Regression Model Size).** Consider the regression model (1) with \( p > n \), \( q = \|\beta_0\|_0 \) (unknown). Assume the SSL prior with \( (1-\theta)/\theta \sim \rho^p \) and \( \lambda_0 \sim \rho^q \) where \( \eta, \gamma \geq 1 \). Assume that \( w = (w_1, \ldots, w_n)^T \) are nonnegative and arise from \( \pi(w) \) such that

1. \( \mathbb{E} w_i = 1 \) for each \( 1 \leq i \leq n \),
(2) \( \exists m \in (0, 1) \) s.t. \( \lim_{n \to \infty} P(\min_i w_i > m) = 1 \),
(3) \( \exists M > 1 \) s.t. \( \lim_{n \to \infty} P(\max_i w_i < M) = 1 \),
(4) \( \text{var}(w_i) \lesssim \frac{1}{\log n} \), \( \text{cov}(w_i, w_j) = C_0 \lesssim \frac{1}{n\log n} \) for any \( 1 \leq i, j \leq n \),
(5) \( \max_{i \neq j} |X_i^T x_j| \lesssim \frac{\lambda_2}{\sqrt{n}} \), and \( \xi_0 > 0 \) satisfies
\[
\max\{\frac{\lambda_2}{\sqrt{n}}, \frac{\lambda_2}{\sqrt{Mn}}/n\} \geq 1 - \delta > 0
\]
for some \( \delta > 0 \), where 
\( \eta^* = \max\{\tilde{\eta} + C_n [\lambda_1, \lambda_1/m] / m, \tilde{\eta} \}
\]
\( C_n = \text{a sequence s.t. } C_n \to \infty, d = \frac{c}{c^2(1 + \gamma - 1)} \) and \( c = c(\eta^*; \beta) \).

Then the BB-SSL posterior satisfies
\[
\lim_{n \to \infty} E_{\beta_n} |\hat{\beta}^n_w - \mu|_0 \lesssim q(1 + K) |Y(n)| = 1
\]
where \( K = 2\pi D \). The definition of \( c(\eta^*; \beta) \) is in the Section A.4.1, supplementary material.

Proof. Section A.4.6, supplementary material.

Theorem 4.3 (Regression model). Under the same conditions as in Theorem 4.2, the BB-SSL posterior concentrates at the near-minimax rate, that is,
\[
\lim_{n \to \infty} E_{\beta_n} \left( \|\hat{\beta}^n_w - \beta\|_2^2 / C_5(\eta^*; \beta) / m \right) \lesssim q(1 + K) \log p / n |Y(n)| = 0
\]
where \( c = c(\eta^*; \beta) \), \( \phi = \phi(\eta^*; \beta) \), whose definition are in the Section A.4.1, supplementary material.

Proof. Section A.4.7, supplementary material.

It follows from Rocková and George (2018) and from (16) 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 concentration as the BB-SSL posterior achieves the same rate of posterior distribution with a proper covariances structure or distributions from Corollary 4.1.

Remark 4.1. In the regression model, when \( w \) arises from the same distribution as in Corollary 4.1, Conditions (1)–(4) in Theorem 4.3 are satisfied by setting \( m = \frac{1}{c} \) and \( M = \frac{3}{2}(\eta + \gamma) \). The detailed proof is in Section A.5, supplementary material.

Remark 4.2. Theorems 4.1 and 4.3 are framed under the typical Bayesian notion of consistence where the posterior distribution concentrates around the truth as \( n \) grows. This notion is different from typical consistency notions in the bootstrap literature. For example, Præstgaard and Wellner (1993) study exchangeable weighted bootstrap which is somewhat related to our work. They provide sufficient conditions for the bootstrap weights to obtain the central limit theorem for the bootstrapped empirical process in finite-parameter situations. Here, we study high-dimensional settings and view the actual posterior distribution, as opposed to the sampling distribution, as a reference point.

4.2. Connections to Other Bootstrap Approaches

Our approach bears a resemblance to Bayesian nonparametric learning (NPL) introduced by Lyddon, Walker, and Holmes (2018) and Fong, Lyddon, and Holmes (2019) which generates exact posterior samples under a Bayesian nonparametric model that assumes less about the underlying model structure. Under a prior on the sampling distribution function \( F_n \), one can use WBB (and also WLB) to draw samples from a posterior of \( F_n \) 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, Lyddon, and Holmes 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 \( \tilde{x}_{1:m}, \tilde{y}_{1:m} \) from the prior \( F_n(x, y) \) defined through
\[
\tilde{x}_k \sim F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i), \quad \tilde{y}_k = \tilde{x}_k + \tilde{x}_k^T \mu
\]
where \( \delta(\cdot) \) is the Dirac measure, \( \mu \) is the Spike, and \( \tilde{y}_k \) satisfies \( \tilde{x}_k = x_i \) for \( i \) satisfies \( \tilde{x}_k = x_i \). A motivation for this prior is derived in the Section B, supplementary material. Under this prior, the NPL posterior samples \( \tilde{\beta} \) generated by Algorithm 3 approximately follow the distribution (see Section B, supplementary material)
\[
\tilde{\beta} \approx \text{arg} \max_{\beta} \left\{ -\frac{1}{2} \sum_{i=1}^{n} w_i(\tilde{Y}_i - \tilde{x}_i^T \beta)^2 \right\}
\]
\[
+ \log \left[ \prod_{j=1}^{p} \pi \left( \beta_j - \frac{c}{c+n} \mu_j^* | \theta \right) \right]
\]
\[
\tilde{\beta} \approx \frac{1}{c+n} \mu^*
\]
Algorithm 3: Posterior Bootstrap Sampling

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

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

for $t = 1, 2, \ldots, T$ do

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

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

(c) Calculate $\tilde{\beta}^t = \arg\max_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2 + \log \left( \int_{\theta} \prod_{j=1}^{p} \pi \left( \tilde{\beta}_j - \mu_j | \theta \right) d\pi(\theta) \right) \right\}$.

end

where $(w_1, w_2, \ldots, w_n)^T \sim n \times \text{Dir}(1 + c/n, \ldots, 1 + c/n)$, each coordinate of $\mu^*$ independently follows the spike distribution, and $c$ represents the strength of our belief in $F_{\pi}$ and can be interpreted as the effective sample size from $F_{\pi}$. In comparison with the BB-SSL estimate

$$
\tilde{\beta}^t = \arg\max_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2 + \log \left( \int_{\theta} \prod_{j=1}^{p} \pi \left( \tilde{\beta}_j - \mu_j | \theta \right) d\pi(\theta) \right) \right\}
$$

where $(w_1, w_2, \ldots, w_n)^T \sim n \times \text{Dir}(\alpha, \ldots, \alpha)$, both (17) and (18) are shrinking toward a random location and both are using Dirichlet weights. The main difference is in the choice of the concentration parameter $c$. When $c = 0$, (17) reduces to WBB (with a fixed weight on the prior) which reflects less confidence in the prior $F_{\pi}$ and thus, less prior perturbation (location shift). When $c$ is large, (17) becomes more similar to (18) where the prior $F_{\pi}$ is stronger and thereby more prior perturbation is induced. Another difference is that (17), although shrinking toward a random location $\frac{c}{c+n} \mu^*$, adds the location back which results in less variance (see Figures in Section B, supplementary material).

5. Simulations

We compare the empirical performance of our BB-SSL with several existing posterior sampling methods including WBB (Newton, Polson, and Xu 2020), SSVS (George and McCulloch 1993), and Skinny Gibbs (Narisetty, Shen, and He 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, Chakraborty, and Mallick (2016). Comparisons are based on the marginal posterior distributions for $\beta$'s, marginal inclusion probabilities (MIP) $P(\gamma = 1 | Y^{(n)})$ as well as the joint posterior distribution $\pi(\phi | Y^{(n)})$. As the benchmark gold standard for comparisons, we run SSVS initialized at the truth for a sufficiently large 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 $\beta$ at the origin. For BB-SSL, we draw weights $w \sim n \times \text{Dir}(\alpha, \ldots, \alpha)$ where $\alpha$ depends on $(n, p, \sigma^2)^T$. When solving the optimization problem

(13) using coordinate-ascent, the default initialization for $\beta$ in the SSLASSO R package (Ročková and Moran 2017) is at the origin. In high-dimensional correlated settings when $\lambda_0 \gg \lambda_1$, however, the performance of BB-SSL can be greatly enhanced by using a warm start reinitialization strategy for a sequence of increasing $\lambda_0$'s where the last value is the target $\lambda_0$ value. The target value should scale polynomially in $p$ in order to achieve the near-minimax rate of convergence, according to Theorem 4.2. This dynamic posterior exploration strategy, that is, warm start reinitialization, was recommended by Ročková (2018a) and Ročková and Moran (2017) in the context of the spike-and-slab LASSO. It is computationally more economical to perform such annealing only once on the original data and then use the output (for the target value $\lambda_0$) for each BB-SSL iteration. We apply this strategy using an output obtained from the R package SSLASSO using an equispaced sequence of $\lambda_0$'s of length 50, starting at $\lambda_1$ and ending at $\lambda_0$. We then run WB1, WBB2, and BB-SSL for $T$ iterations. Throughout the simulations we set $\sigma^2 = 1$ and assume the prior $\theta \sim B(1, p)$. Computational complexity of each algorithm is summarized in Table 1 with actual running times (for varying $p$ and $n$) in Figure 2.

5.1. The Low-Dimensional Case

Similarly to the experimental setting in Ročková (2018b), we generate $n = 50$ observations on $p = 12$ predictors with $\beta_0 = (1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0, 1.3, 0, 0)^T$, where the predictors have been grouped into four blocks. Within each block, predictors have an equal correlation $\rho$ and there is only one active predictor. All the other correlations are set to 0. We choose a single value for $\lambda_0 \propto p$ and generate Dirichlet weights assuming $\alpha = 1$ (for WB1, 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 $B = 5000$. For WBB1, WBB2, and BB-SSL we use $T = 5000$ iterations. All methods perform very well under various metrics in this setting. We refer the reader to Section E.1, supplementary material 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 (see Figure 4(b)). For SSVS1 and Skinny Gibbs, we set $T =$
5.2. The High-Dimensional Case

In terms of the marginal densities of \( \beta_i \)'s, Figure 3 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 with the opposite being true for Skinny Gibbs, WBB1, and WBB2. In terms of the marginal means of \( \gamma_i \) (Figure 4(a)) 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 \( \pi(\gamma | Y^{(n)}) \) we identify over 60 unique models using SSVS1 where the true model accounts for most of the posterior mass. In Figure 4(b), 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 and minor predictors. BB-SSL tracked down 99% of the posterior probability, followed by WBB1 (92%), WBB2 (91%), and Skinny Gibbs (73%). The average times (reported in seconds and ordered from fastest to slowest) spent on generating 1000 effective samples for \( \beta_i \)'s are WBB2 (0.68 s) < WBB1 (0.72 s) < BB-SSL (0.74 s) < SSVS2 (0.82 s) < SSVS1 (0.85 s) < Skinny Gibbs (1.19 s).

We now consider a higher-dimensional case with \( n = 100 \) and \( p = 1000 \), assuming \( \lambda_0 = 50, \lambda_1 = 0.05 \) and \( \omega = 2 \) for BB-SSL. For SSVS1 and Skinny Gibbs we set \( T = 15,000 \) and \( B = 5000 \) while for WBB1, WBB2, and BB-SSL we set \( T = 1000 \). We consider two correlation structures: (a) block-wise correlation, and (b) equi-correlation. In the setting (a), the active predictors have regression coefficients \((1,2,-2,3)^T\) and all predictors are grouped into blocks of size 10, where each group has exactly one active coordinate and where predictors have a within-group correlation \( \rho \). We consider \( \rho \in \{0, 0.6, 0.9\} \) and an extreme case \( \rho = 0.99 \) with a larger signal \((2, 4, -4, 6)^T\). For the equi-correlation setting (with a correlation coefficient \( \rho \)), active predictors have regression coefficients \((2, 3, -3, 4)^T\). We consider \( \rho \in \{0.6, 0.9\} \).

For brevity, we only show results for \( \rho = 0.6 \) in the equi-correlation setting with the rest postponed until the Section E, supplementary material. In the setting (b) with \( \rho = 0.6 \), in terms of the marginal density of \( \beta_i \)'s (shown in Figure 6), 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 5 shows that BB-SSL, WBB1 and WBB2 accurately reproduce the MIPs, while Skinny Gibbs tends to slightly overestimate the MIP as \( \rho \) increases.

To better quantifying the performance of each method, we gauge the quality of the posterior approximation using various metrics in Table 2. The KL divergence is calculated using an R package "FNN" (Beygelzimer et al. 2013), where all parameters are set to their default values. We also report the Jaccard distance of 90% credible intervals relative to the SSVS benchmark. The Jaccard distance (Jaccard 1912) of two intervals \( A \) and \( B \) is defined as \( d_j(A, B) = 1 - j(A, B) = \frac{|A \cap B|}{|A \cup B|} \) and \(|\cdot|\) denotes the length. The Hamming distance is calculated using an R package “e1071” (Meyer et al. 2014). We also compare the \( \ell_1 \) distance of posterior means (i.e., “bias” relative to the SSVS standard) for \( \beta_i \)'s as well as \( \gamma_i \)'s. All methods do well in terms of MIP and the selected model (based on the median probability model rule). For \( \beta_i \)'s, all methods estimate the mean accurately. Taking into account the shape of the posterior for \( \beta_i \)'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. In the settings we tried, we rank the performance of various methods as follows: BB-SSL > Skinny Gibbs > WBB1 \( \approx \) WBB2. The average times (in seconds) when \( \rho = 0.6 \) (in the equi-correlated design) spent on generating 100 effective samples for each \( \beta_i \) are: BB-SSL (0.69 s) > Skinny Gibbs > WBB1 (17.25 s) < WBB2 (13.53 s) < SSVS1 (34.67 s).

**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. The most expensive step in BB-SSL is solving the optimization problem (13) at each iteration. This could be potentially circumvented by using the Generative Bootstrap Sampler (GBS) (Shin, Wang, and Liu 2020) which constructs a generator function that can transform weights into samples from the posterior distribution. This strategy could be particularly beneficial when both \( n \) and \( p \) are large and when many posterior samples are needed. While MCMC-based methods are sensitive to the initialization and can fall into a local trap (e.g., when predictors are highly correlated), we have seen BB-SSL to be less susceptible to this problem. BB-SSL, in some sense, relies on the optimization procedure not finding the global mode at all times. Indeed, we want to provide...
Figure 3. Estimated posterior density (left panel) and credible intervals (right panel) of $\beta_j$ in the low-dimensional correlated case. We have $n = 50, p = 12, \beta_{\text{active}} = (1.3, 1.3, 1.3, 1.3, 1.3, 1.3)^T, \lambda_0 = 7, \lambda_1 = 0.15, \rho = 0.9$. 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.
Figure 4. The low-dimensional correlated case with \( n = 50, p = 12, \beta_{\text{active}} = (1.3, 1.3, 1.3, 1.3)' \) where predictors are grouped into 4 correlated blocks with \( \rho = 0.9 \). We choose \( \lambda_0 = 7, \lambda_1 = 0.15 \).

Figure 5. Posterior means of \( \gamma_i \)’s (i.e., a marginal inclusion probabilities) in high-dimensional settings with \( n = 100, p = 1000 \). We set \( \lambda_0 = 50, \lambda_1 = 0.05 \).

6. Data Analysis

6.1. Life Cycle Savings Data

The Life Cycle Savings data (Belsley, Kuh, and Welsch 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 \( X \) is centered and rescaled so that \( ||X||_2 = \sqrt{n} \). Next, we obtain an estimate of the noise variance \( \hat{\sigma}^2 \) using an ordinary least squares regression. We then divide \( y \) by the estimated noise standard deviation and obtain an estimate \( \hat{\theta} \) by fitting the Spike-and-Slab LASSO with \( \lambda_0 = 20, \lambda_1 = 0.05 \). For the actual BB-SSL we then assume that \( \theta \) is random (i.e., the SSL prior is nonseparable) and set \( \alpha = 2 \log \left( \frac{1 - \hat{\theta}}{\hat{\theta}} \right) \lambda_2 \approx 14 \) and \( a = 1, b = 4 \). We run SSVS and Skinny Gibbs for \( T = 100,000, B = 5000 \), and BB-SSL for \( T = 10,000 \).

Figure 7 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 8a shows the marginal density of \( \beta_i \)’s and 8(b) shows the marginal mean of \( \gamma_i \)’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, Neslin, and Sun (2012) (ISMS Durable Goods Dataset 2) consisting of durable goods sales data from a major anony-
Figure 6. Estimated posterior density (left panel) and 90% credible intervals (right panel) of $\beta_j$'s when all covariates are correlated with $\rho = 0.6$. We have $n = 100, p = 1000, \beta_{\text{active}} = (2, 3, -3, 4)', \lambda_0 = 50, \lambda_1 = 0.05$. BB-SSL is fitted using a single $\lambda_0$ and initialized at SSLASSO solution on the original $X, y$. Since WBB1 and WBB2 produce a point mass at zero, we exclude them from density comparisons.
Table 2. Evaluation of approximation properties (relative to SSVS) in the high-dimensional setting with $n = 100$ and $p = 1000$ based on 10 independent runs.

| Setting         | Block-wise $\rho = 0.9$, $\beta_{\text{active}} = (1,2,-2,3)'$ | Block-wise $\rho = 0.99$, $\beta_{\text{active}} = (2,4,-4,6)'$ |
|-----------------|-----------------------------------------------------------------|-----------------------------------------------------------------|
| Metric          | $\beta$'s            | $\gamma$'s           | Model | Metric          | $\beta$'s            | $\gamma$'s           | Model |
|                 | $\gamma$'s            | $\text{J/D of 90}\%\text{ CI}$ | $\text{Bias}$ | $\gamma$'s           | $\text{J/D of 90}\%\text{ CI}$ | $\text{Bias}$ | $\text{HD}$ | $\gamma$'s           | $\text{J/D of 90}\%\text{ CI}$ | $\text{Bias}$ | $\text{HD}$ |
| Skinny Gibbs    | 0.19  0.009  0.30  0.10  0.04  0.003 | *  *  | 0  | 2.00  0.02  0.62  0.10  0.68  0.005 | *  *  | 0  | 0.15  0.002  2.2 |
| WBB1            | 0.41  3.09  0.45  1  0.03  0.003 | *  *  | 0  | 1.89  3.09  0.51  1  0.74  0.006 | *  *  | 0  | 0.25  0.001  2 |
| WBB2            | 0.21  3.09  0.37  1  0.03  0.003 | *  *  | 0  | 1.89  3.09  0.52  1  0.74  0.006 | *  *  | 0  | 0.25  0.001  2 |
| BB-SSL          | 0.02  0.003  0.14  0.10  0.04  0.003 | *  *  | 0  | 1.73  0.01  0.37  0.10  0.74  0.006 | *  *  | 0  | 0.25  0.001  2 |

NOTE: The best performance is marked in bold font. KL is the Kullback-Leibler divergence, JD is the Jaccard distance of credible intervals (CI), HD is the Hamming distance of the median models. ‘Bias’ refers to the $l_1$ distance of estimated posterior means. We denote with $+$ all numbers smaller than 0.0001, with $\gamma$ an average over active coordinates, and with $-$ an average over inactive coordinates.

Figure 7. Trace plots for the Life Cycle Savings data. We choose $\lambda_0 = 0$, $\lambda_1 = 0.05$. The first column is the BB-SSL traceplot with weight distribution $\alpha = 2 \log \left( \frac{1}{1+\rho_{\text{active}}} \right) = 14$, 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.

mous 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 period (December 4–15). The treatment assignment ($tr_i = I(\text{promotional mailer}_i)$) was random. The data contains 146 descriptors of all customers including prior purchase history, purchase of warranties etc. We will investigate the
Figure 8. 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 $\alpha = 2 \log \left( \frac{1 - \hat{\theta}}{\theta \lambda_1} \right) \approx 14$.

Table 3. Average effective sample size (out of 15,000 samples) for Life Cycle Saving Data.

|       | SSVS  | Skinny Gibbs | BB-SSL |
|-------|-------|--------------|--------|
| Effective sample size | 2,716 | 11,188       | 15,000 |

NOTE: Effective sample size is calculated using R package coda (Plummer et al. 2006).

Figure 9. 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 $\alpha = 2 \log \left( \frac{1 - \hat{\theta}}{\theta \lambda_1} \right) \approx 15$ and black line is SSVS initialized at origin.

Figure 10 depicts estimated posterior density of selected coefficients in the model (19), 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 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 9, 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 article 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 $p$ is large) with an approximation that computes effective posterior samples.
Figure 10. Posterior density and credible intervals for the selected \( \beta_i \)'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 \times \) treatment", "C-ESP-RECT \times \) treatment", "S-CNT-TOT24M \times \) treatment". We set \( \lambda_0 = 100, \lambda_1 = 0.05, a = 1, b = 273 \). We set \( \alpha = 2 \log \left( \frac{1-\theta}{\theta} \right) \lambda_0 \approx 15 \).

\( p \) is large) with fast optimization of randomly perturbed (rewighted) 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 theoretical results are reassuring and significantly extend existing knowledge about Weighted Likelihood Bootstrap (Newton, Polson, and Xu 2020), 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. The GBS method of Shin, Wang, and Liu (2020) could potentially greatly improve the scalability of BB-SSL. We leave this direction for future research.

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)

Funding

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).

References

Ando, A., and Modigliani, F. (1963), “The "Life Cycle" Hypothesis of Saving: Aggregate Implications and Tests,” The American Economic Review, 53, 55–84. [2023]
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, 1–4. [2014]
Belsley, D. A., Kuh, E., and Welsch, R. E. (2005), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity (Vol. 571), Hoboken: Wiley. [2023]
Beygelzimer, A., Kakade, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013), "FNN: Fast Nearest Neighbor Search Algorithms and Applications," R package version, 1(1). [2021]
Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016), “Fast Sampling with Gaussian Scale Mixture Priors in High-Dimensional Regression,” Biometrika, 103, 985–991. [2014,2015,2020]
Bottolo, L., and Richardson, S. (2010), “Evolutionary Stochastic Search for Bayesian Model Exploration,” Bayesian Analysis, 5, 583–618. [2014]
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, 73–108. [2014]
Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), “The Horseshoe Estimator for Sparse Signals,” Biometrika, 97, 465–480. [2015]
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, 80–101. [2014]
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, 921–931. [2014]
Efron, B. (2012), “Bayesian Inference and the Parametric Bootstrap,” The Annals of Applied Statistics, 6, 1971–1997. [2016]
