Model Based Screening Embedded Bayesian Variable Selection for Ultra-high Dimensional Settings

Dongjin Li, Somak Dutta, and Vivekananda Roy
Department of Statistics, Iowa State University, Ames, IA

ABSTRACT
We develop a Bayesian variable selection method, called SVEN, based on a hierarchical Gaussian linear model with priors placed on the regression coefficients as well as on the model space. Sparsity is achieved by using degenerate spike priors on inactive variables, whereas Gaussian slab priors are placed on the coefficients for the important predictors making the posterior probability of a model available in explicit form (up to a normalizing constant). Embedding a unique model based screening and using fast Cholesky updates, SVEN produces a highly scalable computational framework to explore gigantic model spaces, rapidly identify the regions of high posterior probabilities and make fast inference and prediction. A temperature schedule is used to further mitigate multimodal posterior distributions. The temperature value is guided by our model selection consistency results which hold even when the norm of mean effects solely due to the unimportant variables diverges. An appealing byproduct of SVEN is the construction of novel model weight adjusted prediction intervals. The performance of SVEN is demonstrated through a number of simulation experiments and a real data example from a genome wide association study with over half a million markers. Supplementary materials for this article are available online.

1. Introduction
In almost every scientific discipline, rapid collection of sophisticated data has been booming due to recent advancements in technology. In biology, for example, automated sequencing tools have made whole genome sequencing possible in a cost effective manner, thus, providing variations of millions of single nucleotides between individuals. On the other hand, because phenotypic data are typically collected via carefully conducted scientific experiments or other observational studies, number of observations remains on the smaller size, giving rise to regression problems where the number of variables $p$ far exceeds the sample size $n$. Nevertheless, only a few of these variables are believed to be associated with the response. Thus, variable selection plays a crucial role in the modern scientific discoveries.

Classical approaches to deal with the variable selection problems are through regularization methods. A variety of methods using different penalization techniques have been proposed for variable selection in the linear models, such as the lasso (Tibshirani 1996; Datta and Zou 2017), SCAD (Fan and Li 2001; Kim, Choi, and Oh 2008), elastic net (Zou and Hastie 2005), adaptive lasso (Zou 2006), the octagonal shrinkage and clustering algorithm for regression (Bondell and Reich 2008), $L_0$-penalty for best subset regression (Bertsimas, King, and Mazumder 2016; Huang et al. 2018) and others. These methods achieve sparsity by either penalizing the effect sizes or the model sizes but rarely both. Several Bayesian variable selection methods exploit the connection between the penalized estimators and the modes of Bayesian posterior densities under suitably chosen prior distributions on the regression coefficients. Example includes the lasso-Laplace prior connection (Tibshirani 1996), the hierarchical Bayesian lasso (Park and Casella 2008) and other works by Kyung et al. (2010), Xu and Ghosh (2015), and Roy and Chakraborty (2017).

Another popular approach to Bayesian variable selection is integrating the penalties on the effect size and the model size via priors distributions. To that end, auxiliary indicator variables indicating the presence or absence of each variable are introduced to obtain a “spike and slab” prior on the regression coefficients. Here the “spike” corresponds to the probability mass concentrated at zero or around zero for the variables vulnerable to deletion and the “slab” specifies prior uncertainty for coefficients of other variables. Analysis using such models determines (selects) the most promising variables by summarizing the posterior density of the indicator variables and/or the regression coefficients. The seminal works of Mitchell and Beauchamp (1988) and George and McCulloch (1993, 1997) developed a hierarchy of priors over the regression coefficients and the latent indicators and used Gibbs sampler to identify promising models in low dimensional setup (see also Yuan and Lin 2005; Ishwaran and Rao 2005; Liang et al. 2008; Johnson and Rossell 2012). Hans (2010, 2011) discuss the integration of lasso priors into the spike-and-slab setup. Several of these methods have been recently modified and extended to the ultra-high dimensional setup. Narisetty and He (2014) pioneered the theoretical study of Bayesian variable selection in the ultra-high dimensional setup. }
dimensional setup, Ročková and George (2014) introduced the EM algorithm for fast exploration of high-posterior models. Yang, Wainwright, and Jordan (2016) studied model selection consistency and computational complexity when g-prior is placed on the regression coefficients. Shin, Bhattacharya, and Johnson (2018) extended the popular nonlocal priors to model selection and modified the stochastic shotgun model search algorithm (Hans, Dobra, and West 2007), while Zhou and Guan (2019) and Zanella and Roberts (2019) implemented Metropolis Hastings algorithms with an iterative complex factorization and a tempered Gibbs sampler, respectively, for estimating posterior model probabilities.

From a practical standpoint, in the ultra-high dimensional set up, where the number of variables ($p$) is much larger than the sample size ($n$), generally variable screening is performed to reduce the number of variables before applying any of the aforementioned variable selection methods for choosing important variables. The classical approaches as well as Narisetty and He (2014) resort to a two-stage procedure where they first use frequentist screening algorithms (Fan and Lv 2008; Wang and Leng 2016) to reduce the dimension of the problem and then perform variable selection. Shin, Bhattacharya, and Johnson (2018) as well as Cao, Khare, and Ghosh (2020) fuse the frequentist iterated sure independent screening in their stochastic search algorithm. However, these screening methods are frequentist procedures that are not guaranteed to be fideltious to the Bayesian model in practice.

In this work, we extend the classical variable selection model of Mitchell and Beauchamp (1988) to the ultra-high dimensional setting. In particular, we develop a novel methodology for variable selection in the spirit of the stochastic shotgun search algorithm (Hans, Dobra, and West 2007) with embedded screening that is faithful to the hierarchical Bayesian model. We develop sophisticated computational framework that allows us to consider larger search neighborhoods and compute exact unnormalized posterior probabilities in contrast to Shin, Bhattacharya, and Johnson (2018). Furthermore, in order to recover models with large posterior probabilities and mitigate posterior multimodality associated with variable selection models, we use a temperature schedule. We call this Bayesian method and the computational framework selection of variables with embedded screening (SVEN). Keeping prediction of future observations in mind, we develop novel methods for computing approximate posterior predictive distribution and prediction intervals. In particular, using SVEN we construct two prediction intervals, called Z-prediction intervals and Monte Carlo prediction intervals. Furthermore, the temperature schedule in SVEN is guided by our posterior model selection asymptotics. Indeed, following the path laid by Narisetty and He (2014) we derive posterior consistency results when the number of predictors grows nearly exponentially with the sample size. The use of zero (exact spike) inflated mixture priors for regression coefficients allows us to introduce sparsity and relax some assumptions of Narisetty and He (2014).

The rest of the article is laid out as follows. In Section 2 we describe the hierarchical Bayesian variable selection model under the independent normal priors (Section 2.1); develop the SVEN framework (Section 2.2) and extend SVEN to other priors (Section 2.3). Two prediction interval methods are considered in Section 3. We perform detailed simulation studies in Section 4 and compare our methods to several other popular Bayesian and frequentist methods. In Section 5 we analyze a massive dataset from an agricultural experiment with $n = 3951$ and $p = 546,034$ where among the Bayesian methods used for comparison only our method is able to perform variable selection on the whole data. We also show the practical usefulness of our method in obtaining posterior predictive distribution and prediction intervals for the yield of novel crop varieties. We conclude in Section 6 with some discussion and future research directions. The strong model selection consistency results are relegated to the Appendix 6. A supplementary materials containing the proofs of the theoretical results and some computational details is available with sections referenced here with the prefix “S.” The methodology proposed here is implemented by the authors in the R package “bravo” for Bayesian screening and variable selection available on CRAN.

2. Bayesian Variable Selection with Screening

2.1. Hierarchical Mixture Models

Let $y = (y_1, \ldots, y_n)$ denote a $n \times 1$ vector of response values, $Z = (Z_1, \ldots, Z_p)$ an $n \times p$ design matrix of $p$ potential predictors, with vector of partial regression coefficients $\mu \equiv (\mu_1, \ldots, \mu_p)$. We assume latent indicator vector $\gamma = (\gamma_1, \ldots, \gamma_p) \in \{0, 1\}^p$ to denote a model such that the $j$th predictor is included in the regression model if and only if $\gamma_j = 1$. Corresponding to the binary vector, the size of a model is denoted as $|\gamma|$, where $|\gamma| = \sum_{j=1}^{p} \gamma_j$. Also, with model $\gamma$, let $Z_\gamma$ be the $n \times |\gamma|$ submatrix of $Z$ that consists of columns of $Z$ corresponding to model $\gamma$ and $\mu_\gamma$ be the vector that contains the regression coefficients for model $\gamma$. In the first hierarchy of the Bayesian hierarchical mixture model we assume that the conditional distribution of $y$ given $Z, \gamma, \mu_0, \mu$, and $\sigma^2$ is $n$-dimensional Gaussian and is given by

$$
y | Z, \gamma, \mu_0, \mu, \sigma^2 \sim N_n(\mu_0 1_n + Z_\gamma \mu_\gamma, \sigma^2 I_n),$$

where $\mu_0$ is the intercept term and $\sigma^2 > 0$ is the conditional variance. Thus, (1) indicates that each $\gamma$ corresponds to a Gaussian linear regression model $y = \mu_0 1 + Z_\gamma \mu_\gamma + \epsilon$ where the residual vector $\epsilon \sim N_n(0, \sigma^2 I_n)$. However, because the original covariates could have unbalanced scales, a common approach is to reparameterize the above model using a scaled covariate matrix. To that end, suppose $Z$ is the vector of column means of $Z$ and $D$ is the $p \times p$ diagonal matrix whose $i$th diagonal entry is the sample standard deviation of $Z_i$ (the $i$th column of $Z$) and let $X = (Z - 1_nZ^\top)D^{-1}$ denote the scaled covariate matrix. Also we assume that $\beta = D\mu$ and $\beta_0 = \mu_0 + Z_\gamma^\top \mu$. The Bayesian hierarchical regression model after reparameterization is given by

$$
y | \beta, \beta_0, \sigma^2, \gamma \sim N_n(1_n \beta_0 + X_\gamma \beta_\gamma, \sigma^2 I),$$

$$
\beta | \beta_0, \sigma^2, \gamma \sim N(0, \frac{\gamma_j \sigma^2}{k}) \text{ for } j = 1, \ldots, p,
$$

$$
(\beta_0, \sigma^2) | \gamma \sim f(\beta_0, \sigma^2) \propto 1/\sigma^2,
$$

$$
y | w \sim f(y | w) = w^{y \gamma}(1 - w)^{p-|\gamma|}.
$$
In this hierarchical setup a popular noninformative prior is set for \((\beta_0, \sigma^2)\) in (2c) and a conjugate independent normal prior is used on \(\beta\) given \(\gamma\) in (2b) with \(\lambda > 0\) controlling the precision of the prior independently from the scales of measurements. Note that under this prior, if a covariate is not included in the model, the prior on the corresponding regression coefficient degenerates at zero. In (2d) an independent Bernoulli prior is set for \(\gamma\), where \(w \in (0, 1)\) reflects the prior inclusion probability of each predictor. We assume \(\lambda\) and \(w\) are known nonrandom functions of \(n\) and \(p\).

The hierarchical model (2) with centered \(X\) allows us to obtain the distribution of \(y\) given \(\gamma\) in a closed form by integrating out \(\beta_0, \beta_\gamma\) and \(\sigma^2\) (Roy, Tan, and Flegel 2018, sec. 56). Consequently, the marginal likelihood function of \(\gamma\) is given by

\[
L(\gamma | y) = \int_{\mathbb{R}^+} \int_{\mathbb{R}^+} \int_{\mathbb{R}} f(y | \gamma, \sigma^2, \beta_\gamma) f(\beta_\gamma | \gamma, \sigma^2, \beta_0) \, d\beta_\gamma \, d\beta_0 \, d\sigma^2
\]

\[
= c_n \, \lambda^{n/2} |A_\gamma| \lambda^{-1/2} R_\gamma^{-(n-1)/2},
\]

where \(A_\gamma = X_\gamma^T X_\gamma + \lambda I, |A_\gamma|\) is the determinant of \(A_\gamma\),

\[
R_\gamma = \tilde{y}^T \tilde{y} - \tilde{y}^T X_\gamma A_\gamma^{-1} X_\gamma^T \tilde{y} - \tilde{y}^T \tilde{y} - \beta_\gamma^T A_\gamma \beta_\gamma
\]

\[
= \tilde{y}^T \left( I + \lambda^{-1} X_\gamma X_\gamma^T \right)^{-1} \tilde{y}
\]

is the ridge residual sum of squares, \(\tilde{y} = y - \tilde{y} 1_n, \tilde{y} = \sum_{i=1}^{n} y_i/n, \beta_\gamma = A_\gamma^{-1} X_\gamma^T \tilde{y}\) and \(c_n = \Gamma((n-1)/2)/\pi^{(n-1)/2}\) is the normalizing constant.

In order to identify the important variables, we use the (marginal) posterior distribution of \(\gamma\). Thanks to the explicit form of the marginal likelihood (3), this posterior density is given by

\[
f(\gamma | y) \propto f(y | \gamma) f(\gamma) \propto \lambda^{n/2} |A_\gamma| \lambda^{-1/2} R_\gamma^{-(n-1)/2} w^{\gamma_\gamma - \gamma - |\gamma|} /
\]

It’s often convenient to work with \(\log f(\gamma | y)\) which is given by

\[
\log f(\gamma | y) = \text{const} + \frac{1}{2} |\gamma| \log \lambda - \frac{1}{2} \log |A_\gamma| - \frac{1}{2} (n-1) \log R_\gamma - |\gamma| \log(w/(1-w))
\]

**Remark 1.** It is important to note that the regression model (1) on the original covariate scale should be used for prediction, instead of (2) because the hierarchical prior (2b) is defined under the assumption that \(1/n X = 0\) and \(X_j^T X_j = n\) for all \(j\).

**Remark 2.** In this work we assume \(w\) is fixed. However, a popular alternative is to assign a Beta prior on \(w\), that is, let \(w \sim f(w) \propto w^{a-1}(1-w)^{b-1}\) for some \(a, b > 0\). Then it is possible to integrate out \(w\) from (2d) to obtain the marginal prior distribution of \(\gamma\) given by \(f(\gamma) = B(\gamma | a, p - |\gamma| + b) / B(a, b)\), where \(B(\cdot | \cdot)\) is the beta function. This will replace the last term in (5) by \(\log f(\gamma)\).

Ideally, as the sample size increases we would like the posterior of \(\gamma\) to concentrate more and more on the important variables. Several works have alluded to asymptotic guarantees for strong model selection consistency in the ultra-high dimensional regression where \(p\) is allowed to vary subexponentially with \(n\), that is, \(\min(n, p) \to \infty \text{ and } (\log p)/n \to 0\). Here, the strong model selection consistency implies that the posterior probability of the true set of variables converge to 1 as \(n\) tends to infinity. Under shrinking and diffusing priors, Narisetty and He (2014) developed explicit scaling laws for hyper parameters that are sufficient for strong model selection consistency. On the other hand, Shin, Bhattacharya, and Johnson (2018), and more recently, Cao, Khare, and Ghosh (2020) established sufficient conditions for strong model selection consistency under nonlocal type priors (Johnson and Rossell 2012). The Bayesian hierarchical model (2) is similar to Narisetty and He’s (2014) model with the crucial distinction that the spike prior is degenerate: \(P(\beta_i = 0 | y_i = 0) = 1\). Consequently, although most assumptions used here for selection consistency are similar to those made by Narisetty and He (2014), we are able to relax some of the conditions to allow for more noisy unimportant variables. In the Appendix we describe strong model selection consistency results for (2). Indeed, denoting the true model by \(T\), in Theorem 2 in Section 6 we show that under appropriate conditions, posterior probability of the true model, \(f(T | y) \to 1\) in probability as the sample size \(n\) approaches \(\infty\).

Note that strong consistency results also imply that with probability tending to one, the true model is the posterior mode, that is, \(P(t = \arg\max_y f(y | y)) \to 1\) as \(n \to \infty\). However, in finite sample this need not be true. Furthermore, when the regularity conditions do not hold, there may be multiple models with large posterior probabilities even for large \(n\). Thus, we would like to discover models with practically large posterior probability values. However, in ultra-high dimensional problems, traditional computational methods based on Markov chain Monte Carlo (MCMC) algorithms have poor performance. Thus, next we describe SVEN to explore the posterior distribution \(f(y | y)\). In particular, SVEN will be used to discover high probability regions and find the maximum a posteriori (MAP) model arg max_y f(y | y).

### 2.2. Stochastic Search for High Posterior Probability Models

#### 2.2.1. Stochastic Shotgun Search Algorithms

Hans, Dobra, and West (2007) proposed the stochastic shotgun search (SSS) algorithm for recovering models with large posterior probabilities. To that end, for a given model \(y\) let \(\text{nbd}(\gamma) = \gamma^+ \cup \gamma^- \cup \gamma^0\) denote a neighborhood of \(\gamma\), where \(\gamma^+\) is an “added” set containing all the models with one of the \(p - |\gamma|\) remaining covariates added to the current model \(\gamma\), \(\gamma^-\) is a “deleted” set obtained by removing one variable from \(\gamma\); and \(\gamma^0\) is a “swapped” set containing the models with one of the variables from \(\gamma\) replaced by one variable from \(\gamma^c\). The SSS algorithm then starts with an initial model \(g(0)\), and for \(k = 1, 2, \ldots\)

- (SSS1) Compute \(f(\gamma | y)\) for all \(\gamma \in \text{nbd}(g(k-1))\).
- (SSS2) Separate sample \(s^+\) from \(g(k-1)\), \(s^0\) from \(g(k-1)\) and \(s^-\) from \(g(k-1)\) with probabilities proportional to \(f(\cdot | y)\).
- (SSS3) Sample \(g(k)\) from \(s^+, s^0\) and \(s^-\) with probability proportional to \(f(s^+ | y), f(s^0 | y)\) and \(f(s^- | y)\), respectively.
After running for some prespecified large number of iterations, the algorithm then declares the model discovered with the largest (unnormalized) posterior probability as the MAP model. Hans, Dobra, and West (2007) notes that the sampling probabilities in (SSS1) and (SSS2) can be replaced by the Bayesian information criteria (BIC) and the sampling weights can be computed in parallel.

Following the success of SSS, Shin, Bhattacharya, and Johnson (2018) propose further improvement. Note that, Shin, Bhattacharya, and Johnson (2018) use nonlocal priors, and so the posterior probabilities \( f(y|γ) \) are not available analytically. In fact, they resort to using computationally expensive Laplace approximations which suggests exact numerical computations of these quantities are also not straightforward (see also Cao, Khare, and Ghosh 2020). Also in ultra-high dimensional problems, SSS may not be scalable due to its implementation. Thus Shin, Bhattacharya, and Johnson (2018) propose a simplified stochastic shotgun search with screening (SSS5) by dropping the “swapped” set from consideration and moreover, by screening out variables from the “added” set. (Note that, in high dimension, the number of “swapped” models is much larger than the numbers of “added” and “deleted” models.) For screening, borrowing ideas from frequentist correlation screening of Fan and Lv (2008), they propose computing the least squares residuals from a regression of \( y \) on \( X_γ \), and compute the absolute values of the residuals. They then propose keeping models in the “added” set corresponding to the largest few of the absolute correlations. This greatly reduces the burden of computing \( f(y|γ) \) for all \( γ \) in the “added” set. However, in their R package BayesS5, the authors have used ridge residuals with unit ridge penalty instead of the least squares residuals. Nevertheless, the S5 algorithm has been useful for exploring the posterior distribution of \( y \) (Cao, Khare, and Ghosh 2020).

In the variable selection model (2), the Gaussian conjugacy provides analytically tractable forms for \( f(y|γ) \) up to a normalizing constant. We also show that \( f(y|γ) \) can be rapidly computed for the swapped models, thereby allowing us to include the swapped models in the neighborhood. We thus develop a stochastic shotgun algorithm with (posterior) model based screening and develop scalable statistical computations for drawing fast Bayesian inference and prediction.

### 2.2.2. Selection of Variables with Embedded Screening

In order to describe the SVEN algorithm, we first describe how to compute the unnormalized posterior probabilities in the (SSS1) step. To that end, compute \( ζ = X^T \hat{y} \) as \( D^{-1} Z^T \hat{y} \) once and for all. Next, suppose we have a current model \( γ \) and we want to compute the unnormalized posterior probabilities of each model in \( γ^+ \). Suppose \( U_γ \) is the upper triangular Cholesky factor of \( X_γ^T X_γ + \lambda I \) and \( ν_γ = U_γ^{-1} X_γ^T \hat{y} \). In the algorithm below, scalar addition to vector, division between two vectors and other arithmetical and algebraic operations on vectors are interpreted as entry-wise operations, as implemented in most statistical software (e.g., R). Then

1. Compute \( S_1 ← U_γ^T X_γ^T \) by using forward substitution.
2. Update \( S_2 ← S_1 D^{-1} \). [No need to center \( Z \) because \( S_1 1 = 0 \).]
3. Compute \( S_3 \) as the sum of squares of each column of \( S_2 \). Note that \( S_3 \) is a \( |γ| \times p \) matrix and so these sums of squares should be computed without storing another \( |γ| \times p \) matrix.
4. Set \( S_4 ← √n + λ − S_3 \) where the arithmetic operations are performed entry-wise on the vector. Also in this operation, the entries corresponding to the variables in \( γ \) are ignored.
5. Compute \( S_5 ← (ξ − S_2 ν_γ)/S_4 \).
6. Compute \( S_6 ← \log \det U_γ + \log S_4 \).
7. Compute \( S_7 ← ||y||^2 − ||νy||^2 − S_2^2 \).
8. Compute \( S_8 ← 0.5(|γ| + 1) \log λ − S_6 − 0.5(n − 1) \log S_7 + ((|γ| + 1) \log(1/w)). \)

Then for all \( i \not\in γ \), the \( i \)-th entry of \( S_8 \) above contains the unnormalized posterior probability of the model obtained by including \( i \) in \( γ \). The other entries are ignored. For each model in \( γ − \), its unnormalized posterior probability can be computed easily because typically \( |γ| \) is small. In order to compute the unnormalized posterior probabilities of the models in \( γ + \), we consider each \( γ' \in γ − \) and compute the unnormalized posterior probabilities of each \( γ'' \in γ + \) using the above algorithm. Thus, we can compute the unnormalized posterior probabilities of each model in \( \text{nbld}(γ) \).

Given the current model \( γ \), the complexity for computing (unnormalized) \( f(y|γ) \) for all \( γ \in \text{nbld}(γ) \) by the above algorithm is \( O(|γ|^p + |γ|^2|Z|) \). In addition, \( Z \) is sparse, as in the genome-wide association study example in Section 5, the complexity for computing all unnormalized posterior probabilities in \( \text{nbld}(γ) \) is linear in both \( n \) and \( p \). Finally, note that, the additional memory requirement for the above algorithm except storing the \( Z \) matrix is practically \( O(n \times p) \). Also, different steps including step 2 of the above algorithm can be performed in parallel using distributed computing architecture.

Using the above algorithm as the foundation, we now discuss the SVEN algorithm. Suppose \( γ = T_1 < T_2 < \cdots < T_m \) is a given temperature schedule. Let \( g^{(0)} \) denote the empty model (i.e., the model without any predictor included). Then, for \( i = 1, 2, \ldots, m \)

- Set \( g^{(i,0)} \) to be the empty model. Then for \( k = 1, \ldots, N \)
  - (SVEN1) [Same as (SSS1)] Compute \( f(g'|γ) \) for all \( g' ∈ \text{nbld}(g^{(i−1)}) \).
  - (SVEN2) [Screening step] Consider at least \( L \) probability neighboring models. That is, construct the set \( M_k \subseteq \text{nbld}(g^{(i−1)}) \) with \( |M_k| ≥ L \) such that \( g' ∈ M_k \) if
    
  \[ f(g'|γ) > ϕ \max_{g'' ∈ \text{nbld}(g^{(i−1)})} f(g''|γ) \]

  and
  
  \[ f(g'|γ) ≥ f(g''|γ), \quad ∀ g'' ∈ \text{nbld}(g^{(i−1)}) \cap M_k \], where \( ϕ \) is some prespecified number.

  - (SVEN3) [Shotgun step] Assign the weight \( f(g'|γ)^{1/T} \) to a model \( g' ∈ M_k \). Sample a model from \( M_k \) using these weights and set it as \( g^{(i,k)} \).

In our simulation studies and data analyses, we use \( L = 20 \) and \( ϕ = \exp(-6) \). Our ability to efficiently compute the unnormalized posterior probability of all neighboring models allows us
to implement the screening (SVEN2) directly using the objective function $f(y')|y)$. This is a key difference between SVEN and S5 of Shin, Bhattacharya, and Johnson (2018). Because models with large probabilities could be separated by models with very low probabilities, a temperature schedule has been used. Such tempering is quite common in simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983) and has also been used in Shin, Bhattacharya, and Johnson (2018). In order to choose a temperature schedule, we turn to our asymptotic results from Section S6 of the supplementary materials. In particular, the theory indicates that the log-posterior probabilities of good models with small model size are separated by roughly $O(\log p)$. Thus, in order to facilitate jumps between these models we set $T_m = \log p + \log \log p$ where the additional log log $p$ is a heuristic adjustment common in numerical computations. Also the remaining temperatures are chosen to be equally spaced between 1 and $T_m$.

Note that at every temperature we start the SVEN algorithm at the empty model that are run separately. Because the stochastic shotgun might have a tendency to wander off to obscure valleys containing large number of variables especially under high temperature; running them separately avoids getting trapped in such a valley. Most good models have small size and so they could be explored relatively early when started multiple times from the empty model. A direction for future work could be exploring occasional communication across the temperature ladder in SVEN.

Note that our algorithm does not require explicitly storing the matrix $X$. Indeed, in many applications, $Z$ could be sparse and efficiently stored in the memory. The matrix $X$ on the other hand is always dense. Overall our method is extremely memory efficient, and we are able to directly perform variable selection with significantly larger $p$ than the other methods may handle.

In addition to the MAP model, our method also provides the unnormalized posterior probability of all the models explored by the algorithm and facilitates approximate Bayesian model unnormalized posterior probability of all the models explored and efficiently stored in the memory. The matrix $X$ ladder in SVEN.

Indeed, in many applications, a temperature schedule has been be explored relatively early when started multiple times from the empty model. A direction for future work could be exploring occasional communication across the temperature ladder in SVEN.

2.3. SVEN for Other Priors

In this section, we describe how SVEN can be extended to accommodate other popular priors on the regression coefficients and $y$.

Zellner’s g-prior (Zellner 1986) is a popular alternative to the independent normal prior (2b) and is given by $\beta_s|\gamma, \sigma^2 \sim N_\gamma(0, g\sigma^2 (X_j^T X_j)^{-1})$ provided that for every $k \leq n - 1$, all $n \times k$ submatrices of $X$ have full column rank and we restrict the support of the prior distribution on $\gamma$ to models of size at most $n - 1$. Assuming that $g$ is a nonrandom function of $n$ and $p$, the marginal posterior of $\gamma$ is then given by

$$f_{\mu}(\gamma|y) \propto \left(\frac{1}{1 + g}\tilde{y}^T X_\gamma(X_\gamma^T X_\gamma)^{-1}X_\gamma \tilde{y}\right)^{-(n-1)/2} \frac{f(y)}{(1 + g)^{n/2}} I(|y| < n),$$

(7)

where the priors on $\beta_0$ and $\sigma^2$ have been assumed to be the same as (2c).

Recently, Kojima and Komaki (2016) have proposed a class of discrete determinantal point process priors on the model space that discourages simultaneous selection of collinear predictors. The founding member of this class of priors is given by

$$f_{\mu}(\gamma) = \left|\omega X_\gamma^T X_\gamma + I_p\right|,$$

(8)

where $\omega > 0$ controls the prior interaction of the model size. A value of $\omega > 1$ promotes larger models, while $\omega \leq 1$ promotes smaller models. Although Kojima and Komaki (2016) have studied the prior when $p < n$, it can be also used when $p > n$. In particular, (8) puts zero prior probability on all models of size greater than $n$.

In order to compute (7) and (8) for all models in the neighborhood of a model $\gamma$, we modify SVEN as follows. Suppose $\tilde{U}_\gamma$ is the upper triangular Cholesky factor of $X_\gamma^T X_\gamma$ and $\tilde{v}_\gamma = \tilde{U}_\gamma^{-1} X_\gamma^T \tilde{y}$. Then we compute,

1. Compute $\tilde{S}_1 = \tilde{U}_\gamma^{-T} X_\gamma^T$ by using forward substitution.
2. Update $\tilde{S}_2 = \tilde{S}_1 Z D^{-1}$.
3. Compute $\tilde{S}_3$ as the sum of squares of each column of $\tilde{S}_2$.
4. Set $\tilde{S}_4 = \sqrt{\tilde{n} - \tilde{S}_3}$, where for a vector $x$, $x_+$ is defined as the vector whose ith entry is $\max(x_i, 0)$. This is required to guard against singularity.
5. Compute $\tilde{S}_5 = (\tilde{S}_3 + \tilde{S}_4)/\tilde{S}_4$, where the ith entry of $\tilde{S}_5$ is set to be zero whenever the ith entry of $\tilde{S}_4$ is zero. This step is needed only when the Zellner’s $g$-prior is used.

Consequently, for every $i \neq \gamma$ the ith entry of $|\tilde{U}_\gamma| \tilde{S}_4$ is equal to $\left|X_\gamma^T X_\gamma\right|^{1/2}$ where $\gamma'$ is obtained by including $i$ in $\gamma$.

Moreover, the $ith$ entry of $\tilde{S}_7 = \|	ilde{y}\|^2 - (g/(g + 1))((\tilde{S}_2 + \tilde{S}_4))$ is equal to $\|	ilde{y}\|^2 - (g/(g + 1))\tilde{y}^T X_\gamma(X_\gamma^T X_\gamma)^{-1}X_\gamma^T \tilde{y}$ required in the computation of (7). This modified steps for computing unnormalized posterior probabilities of models in $\gamma^+$ can then be used to compute the unnormalized posterior probabilities of all models in $\gamma^-$ and $\gamma^0$ as described in Section 2.2.2 and (SVEN1)–(SVEN3) can be used to search for models with high posterior probabilities.
In the literature, mostly the MAP (more precisely the discovered MAP model) model is used for prediction. In the next section we develop methods for point and interval predictions using the top models $\gamma^{(k)}$'s with associated weights $w_k$'s.

3. Posterior Predictive Distribution and Intervals

The posterior predictive distribution of the response $y^*$ at a new covariate vector $z^* \in \mathbb{R}^p$, conditional on the observed covariate matrix $Z$ and hyper parameters $\lambda$ and $\omega$ is given by

$$f(y^* | y, \mu, \sigma^2) = \int f(y^* | z^*, \gamma, \mu, \sigma^2) f(\gamma, \mu, \sigma^2 | y, Z) \, d\gamma \, d\mu \, d\sigma^2,$$

where $f(y^* | z^*, \mu, \sigma^2)$ is the density of $N(\mu, \sigma^2)$ as given in (1), $f(\gamma, \mu, \sigma^2 | y, Z)$ is the joint posterior density of $(\gamma, \mu, \sigma^2)$ given $(y, Z)$ deduced from the hierarchical model (2), and $\mathcal{S}_y = (0, \infty) \times \mathbb{R}^{|y|} \times \mathbb{R}$. Note that, the distribution (9) is not tractable. However, as shown later in this section, posterior predictive mean and variance of $y^*$ can be expressed as (posterior) expectations of some analytically available functions of $y$. Also, samples from an approximation of (9) can be drawn using our SVEN algorithm. Using these approaches, we now propose two methods for computing approximate posterior prediction intervals for $y^*$ under the hierarchical model setup (2a)–(2d).

3.1. A Z-prediction Interval

In this section we describe some approximations to $E(y^* | y)$ and var($y^* | y$) and use those to construct an interval for $y^*$. To that end, from (2) we observe that $\beta_0$ and $\beta_\gamma$ are conditionally independent given $y, \gamma, \sigma^2$, and $Z$ with

$$\beta_0 | y, Z, \gamma, \sigma^2 \sim N(\bar{y}, \sigma^2 / n),$$

and

$$\beta_\gamma | y, Z, \gamma, \sigma^2 \sim N(A^{-1}\bar{y}, \sigma^2 A^{-1}),$$

where $A_\gamma = X_\gamma^T X_\gamma + \lambda I$ as defined in Section 2.1. Consequently, the full conditional distribution of $(\mu_0, \beta_\gamma)$ is a $(|y| + 1)$-dimensional multivariate Gaussian distribution given by

$$
\begin{pmatrix}
\mu_0 \\
\beta_\gamma
\end{pmatrix}
\sim N
\left(
\begin{pmatrix}
\bar{y} - Z_\gamma^T F_y D_y X_\gamma^T y \\
F_y D_y X_\gamma^T y
\end{pmatrix},
\begin{pmatrix}
\sigma^2 I & \sigma^2 Z_\gamma^T F_y D_y X_\gamma^T \\
\sigma^2 F_y D_y X_\gamma^T & \sigma^2 F_y D_y X_\gamma^T D_y F_y
\end{pmatrix}
\right),
$$

where $Z_\gamma$ and $D_\gamma$ are subvector of $Z$ and submatrix of $D$, respectively, corresponding to the model $\gamma$, and $F_\gamma = D_\gamma^{-1} A_\gamma^{-1} D_\gamma^{-1}$. Also,

$$\sigma^2 | y, \gamma \sim IG((n - 1)/2, R_y/2),$$

where $IG(a, b)$ denotes a inverse gamma random variable with density $f(\sigma^2) \propto (\sigma^2)^{-a-1} \exp(-b/\sigma^2)$, and $R_y$ is defined in (4). Next, let $\bar{z}_\gamma = z_\gamma^* - Z_\gamma$ and note that $E(\sigma^2 | y, \gamma) = R_y / (n - 3)$. Thus, using iterated expectation and variance formulas, we have

$$E(y^* | y) = E \left[ E \left( y^* | y, \sigma^2, \mu_0, \mu, \gamma \right) | y \right]$$

$$= E \left[ \mu_0 + \mu_\gamma^T z_\gamma^* | \sigma^2, y, \gamma \right] | y$$

$$= \bar{y} + E \left\{ z_\gamma^T F_y D_y X_\gamma^T y \right\} | y$$

and,

$$\text{var}(y^* | y) = \text{var} \left( y^* | y, \sigma^2, \mu_0, \mu, \gamma \right)$$

$$+ \text{var} \left( E \left( y^* | y, \sigma^2, \mu_0, \mu, \gamma \right) | y \right)$$

$$= \text{var} \left( \mu_0 + \mu_\gamma^T z_\gamma^* | \sigma^2, y, \gamma \right)$$

$$+ \text{var} \left( E \left( \mu_0 + \mu_\gamma^T z_\gamma^* | \sigma^2, y, \gamma \right) | y \right)$$

$$= E \left[ \mu_0 + \mu_\gamma^T z_\gamma^* | \sigma^2, y, \gamma \right]$$

$$+ \left[ \frac{R_y}{n - 3} \left( 1 + \frac{1}{n} - \frac{\bar{z}_\gamma^T F_y D_y X_\gamma^T y}{F_y D_y X_\gamma^T D_y F_y} \right) \right] | y.\right)$$

From (13a) and (13b) we see that both $E(y^* | y)$ and $\text{var}(y^* | y)$ can be expressed as posterior expectations of analytically available functions of $y$. However, because the posterior of $\gamma$ is not entirely available, we propose using the models $\gamma^{(1)}, \ldots, \gamma^{(K)}$ obtained from SVEN as described in Section 2.2.2 with weights $w_1, \ldots, w_K$, respectively, to approximate these expectations and variances. We can use these approximate posterior predictive mean and variance of $y^*$ to obtain a $(1 - \alpha)$ prediction interval for $y^*$ as $\hat{E}(y^* | y) \pm z_{\alpha/2} \sqrt{\text{var}(y^* | y)^{1/2}}$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$th standard normal quantile. We call this interval Z-prediction interval (Z-PI). Also, the posterior predictive mean is used as a point estimate of $y^*$. In the next section, we describe an alternative method for computing a prediction interval for $y^*$ using Monte Carlo simulation.

3.2. A Monte Carlo Prediction Interval

A prediction interval for $y^*$ can also be constructed using Monte Carlo (MC) samples generated from the posterior predictive distribution (9). Specifically, a $(1 - \alpha)$ prediction interval for $y^*$ is given by $[F_{y^*}^{-1}(\alpha/2), F_{y^*}^{-1}(1 - \alpha/2)]$, where $F_{y^*}^{-1}(\alpha)$ denotes the $\alpha$th quantile of the distribution (9). Now, we describe a method for sampling from an approximation of (9) using SVEN. To that end, we consider $\tilde{f}(y^* | y)$ given by

$$\tilde{f}(y^* | y) = \sum_{i=1}^{K} w_i \int_{\mathcal{S}_y^{(0)}} f(y^* | z^*, \gamma^{(i)}, \mu_0, \mu_\gamma^{(i)}, \sigma^2) \, d\gamma \, d\mu \, d\sigma^2,$$

where $w_i$’s are defined in (6), and $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(K)}$ are the $K$ highest probability models obtained by SVEN as described in Section 2.2.2. Thus, $\tilde{f}(y^* | y)$ is the posterior predictive pdf $f(y^* | y)$ given in (9) except that the marginal posterior of $\gamma$ is replaced by a mixture distribution of models chosen by SVEN. Samples from (14) can be drawn as follows. First, we sample $y$ from the top $K$ models with $P(\gamma \equiv \gamma^{(k)}) = w_k, \ (1 \leq k \leq K)$. Given $y$, we then sample $\sigma^2$ from (12). Next given $y$ and $\sigma^2$, we sample $\beta_0$ and $\beta_\gamma$ from (10). Then we compute $\mu_\gamma = D_\gamma^{-1} \beta_\gamma$ and $\mu_0 = \beta_0 - Z_\gamma^T \mu_\gamma$, which are samples from (11). Finally generate $y^*$ from $N(\mu_0 + \mu_\gamma^T z_\gamma^*, \sigma^2)$. We repeat the above process
a large number of times and construct a \((1 - \alpha)\) MC prediction interval (MC-PI) for \(y^*\) as \(\tilde{F}^{-1}(\alpha/2), \tilde{F}^{-1}(1 - \alpha/2)\), where \(\tilde{F}^{-1}(\cdot)\) denotes the empirical quantiles based on these samples. In practice, generally one wants prediction intervals at several new covariate vectors \(z^*\)'s. In Section S1 of the supplementary materials, we describe a computationally efficient way of drawing multiple samples from (14) using the above method and thus simultaneously computing prediction intervals at several new covariate vectors \(z^*\)'s.

4. Simulation Studies

In this section, we study the performance of our SVEN method through several numerical experiments, and compare it with some other existing methods. The competing variable selection methods we consider are S5 (Shin, Bhattacharya, and Johnson 2018, R package: BayesS5), EMVS (Rocková and George 2014, R package: EMVS), fastBVS (Zhou and Guan 2019) and three penalization methods, LASSO, Elastic Net with elastic mixing parameter \(\alpha = 0.5\) (R package: glmnet) and SCAD (R package: ncveqreg). As also noted in Shin, Bhattacharya, and Johnson (2018), we could not include BASAD (Narisetty and He 2014) for its high computational burden and our ultra-high dimensional examples. As used in Table 1 of Rocková and George (2014) we run EMVS with \(v_1 = 1000\) and three choices for \(v_0\), namely, \(v_0 = 0.2\) (EMVS\(^1\)), \(v_0 = 0.6\) (EMVS\(^2\)) and \(v_0 = 1\) (EMVS\(^3\)). For fastBVS, the results are obtained using 100,000 MCMC iterations after a burn-in of 10,000 steps. For S5 the hyperparameters are tuned using a function provided in BayesS5. Moreover, we denote by piMOM and peMOM, respectively, the product inverse-moment and the product exponential moment nonlocal priors used under S5. In addition, for piMOM and peMOM, we use MAP and LS to denote the MAP estimator and the least squares estimator from the MAP model, respectively. Under SVEN, both MAP and WAM models, as described in Section 2.2 are considered. For SVEN, we use \(N = 200\) and the temperature schedule described in Section 2.2.2 with \(m = 9\). Also, for SVEN, the ridge estimator \(\hat{\beta}_\gamma\) is used to estimate the regression coefficients for the MAP and the WAM models.

4.1. Setup of Experiments

Our numerical studies are conducted in six different simulation settings described below.

4.1.1. Independent Predictors

In this example, entries of \(X\) are generated independently from \(N(0, 1)\). The coefficients are specified as \(\beta_1 = 0.5, \beta_2 = 0.75, \beta_3 = 1, \beta_4 = 1.25, \beta_5 = 1.5, \) and \(\beta_j = 0, \forall j > 5\).

4.1.2. Compound Symmetry

This example is taken from Example 3 in Wang (2009) and Example 2 in Wang and Leng (2016). The rows of \(X\) are generated independently from \(N_\rho\left(0, (1 - \rho)I_p + \rho I_p^1I_p^\top\right)\) where we take \(\rho = 0.6\). The regression coefficients are set as \(\beta_j = 5\) for \(j = 1, \ldots, 5\) and \(\beta_j = 0\) otherwise.

4.1.3. Autoregressive Correlation

This autoregressive correlation structure is commonly observed in time series data where the correlation between observations depends on the time lag between them. In this example, we use AR(1) structure where the variables further apart from each other are less correlated. Following Example 2 in Wang and Leng (2016), \(X_j = \rho X_{j-1} + (1 - \rho^2)^{1/2}z_j\), for \(1 \leq j \leq p\), where \(X_0\) and \(z_j\) \((1 \leq j \leq p)\) are iid \(\sim N_\rho(0, I_p)\). We use \(\rho = 0.6\) and set the regression coefficients as \(\beta_1 = 3, \beta_4 = 1.5, \beta_7 = 2\) and \(\beta_j = 0\) for \(j \not\in \{1, 4, 7\}\).

4.1.4. Factor Models

This example is from Meinshausen and Bühlmann (2006) and Wang and Leng (2016). With a fixed number of factors, \(K\), we first generate a \(p \times K\) matrix \(F\) whose entries are iid standard normal. Then the rows of \(X\) are independently generated from \(\mathcal{N}_p(0, FF^\top + I_p)\). We fix \(K = 2\) and the regression coefficients are set to be the same as in Example 4.1.2.

4.1.5. Group Structure

This special correlation structure arises when variables are grouped together in the sense that the variables from the same group are highly correlated. This example is similar to example 5 of Wang and Leng (2016) and example 4 of Zou and Hastie (2005) where 15 true variables are assigned to three groups. We generate the predictors as \(X_{ni} = z_i + \xi_{i1m}\), \(X_{5+m} = z_2 + \xi_{2m}\), \(X_{10+m} = z_3 + \xi_{3m}\) where \(z_i\) are iid \(\sim N_{n}(0, I_n)\) and \(\xi_{im} \sim N_{n}(0, 0.01I_n)\) for \(1 \leq i \leq 3\) and \(m = 0, 1, 2, 3, 4\). The regression coefficients are set as \(\beta_j = 3\) for \(j \in \{1, 2, \ldots, 15\}\) and \(\beta_j = 0\) otherwise.

4.1.6. Extreme Correlation

This challenging example is the example 6 of Wang and Leng (2016). In this example, we first simulate \(z_j, j = 1, \ldots, p\) and \(w_j, j = 1, \ldots, 5\) independently from the multivariate standard normal distribution \(\mathcal{N}_5(0, I_5)\). Then the covariates are generated as \(X_j = (z_j + w_j)/\sqrt{2}\) for \(j = 1, \ldots, 5\) and \(X_j = (z_j + \sum_{i=1}^{5} w_i)/2\) for \(j = 6, \ldots, p\). By setting the number of true covariates to be 5 and let \(\beta_j = 5\) for \(j = 1, \ldots, 5\) and \(\beta_j = 0\) for \(j = 6, \ldots, p\), the correlation between the response and the unimportant covariates is around \(2.5/\sqrt{3}\) times larger than that between the response and the true covariates, making it difficult to identify the important covariates.

Our simulation experiments are conducted using 100 simulated pairs of training and testing datasets. For each of the simulation settings introduced above, we set \(p = 20,000\) and generate training dataset and testing dataset of size \(n = 400\) each. The error variance \(\sigma^2\) is determined by setting theoretical \(R^2 = 90\%\) (Wang 2009). The hyperparameters \(\omega\) and \(\lambda\) are chosen to be \(\sqrt{n}/p\) and \(n/p^2\), respectively, except for group structure where we also use \(\lambda = 200\) and \(\omega = 0.02\) to account for the high within-group correlation and relatively large true model size.

In order to evaluate the performance of the propose method, we compute the following metrics: (a) mean squared prediction error based on testing data (MSPE); (b) mean squared error between the estimated regression coefficients and the true coefficients (MSEC); (c) coverage probability which is defined
Table 1. Independent predictors (Section 4.1.1)

| Method       | MSPE   | MSE_β | Coverage probability (%) | Average model size | FDR (%) | FNR (%) | Jaccard Index (%) |
|--------------|--------|-------|--------------------------|--------------------|---------|---------|------------------|
| SVEN(WAM)    | 0.6387 | 0.0083| 100                      | 5                  | 0       | 0       | 100              |
| SVEN(MAP)    | 0.6387 | 0.0083| 100                      | 5                  | 0       | 0       | 100              |
| piMOM(MAP)   | 0.6384 | 0.0081| 100                      | 5                  | 0       | 0       | 100              |
| peMOM(MAP)   | 0.6384 | 0.0080| 100                      | 5                  | 0       | 0       | 100              |
| piMOM(LS)    | 0.6387 | 0.0083| 100                      | 5                  | 0       | 0       | 100              |
| peMOM(LS)    | 0.6387 | 0.0083| 100                      | 5                  | 0       | 0       | 100              |
| fastBVSR     | 0.6478 | 0.0091| 100                      | 5.09               | 1.45    | 0       | 98.55            |
| EMVS<sup>1</sup> | 1.0087 | 0.3777| 0                        | 3.80               | 0       | 24      | 76               |
| EMVS<sup>2</sup> | 2.5203 | 1.8734| 0                        | 3.60               | 0       | 60.2    | 39.8             |
| EMVS<sup>3</sup> | 5.0909 | 4.3994| 0                        | 0.53               | 0       | 89.4    | 10.6             |
| LASSO        | 0.7489 | 0.1146| 100                      | 18.42              | 9.54    | 0       | 88.45            |
| SCAD         | 0.6454 | 0.0152| 100                      | 9.22               | 0       | 95.86   | 4.14             |
| Elastic Net  | 0.8266 | 0.1898| 100                      | 1.33               | 0       | 98.67   | 1.33             |

Table 2. Compound symmetry (Section 4.1.2) with ρ = 0.6.

| Method       | MSPE   | MSE_β | Coverage probability (%) | Average model size | FDR (%) | FNR (%) | Jaccard Index (%) |
|--------------|--------|-------|--------------------------|--------------------|---------|---------|------------------|
| SVEN(WAM)    | 48.3069| 1.1912| 100                      | 5                  | 0       | 0       | 100              |
| SVEN(MAP)    | 48.3069| 1.1892| 100                      | 5                  | 0       | 0       | 100              |
| piMOM(MAP)   | 48.2277| 1.0018| 100                      | 5                  | 0       | 0       | 100              |
| peMOM(MAP)   | 50.1528| 3.5669| 94                       | 4.96               | 0.37    | 1.2     | 98.5             |
| piMOM(LS)    | 48.3069| 1.1892| 100                      | 5                  | 0       | 0       | 100              |
| peMOM(LS)    | 50.2789| 3.8758| 94                       | 4.96               | 0.37    | 1.2     | 98.5             |
| fastBVSR     | 50.0479| 2.5620| 100                      | 5.78               | 9.54    | 0       | 90.46            |
| EMVS<sup>1</sup> | 49.9839| 5.3218| 100                      | 5.38               | 9.22    | 0       | 90.78            |
| EMVS<sup>2</sup> | 49.6243| 4.5157| 100                      | 5.08               | 1.33    | 0       | 98.67            |
| EMVS<sup>3</sup> | 55.2280| 17.9975| 100                      | 51.02              | 89.94   | 0       | 84.4             |
| LASSO        | 0.6454 | 0.0152| 100                      | 18.42              | 9.54    | 0       | 88.45            |
| SCAD         | 48.3167| 2.1556| 100                      | 11.55              | 0       | 95.86   | 4.14             |
| Elastic Net  | 57.5750| 23.9724| 100                      | 93.76              | 0       | 6.24    |                 |

as the proportion of the selected models containing the true model (d) average model size which is calculated as the average number of predictors included in the selected models over all the replications (e) false discovery rate (FDR); (f) false negative rate (FNR) and (g) the Jaccard index which is calculated as the size of the intersection divided by the size of the union of the selected model and the true model. All computations are done using single-threaded R on a workstation with two 2.6 GHz 8-Core Intel®E5-2640 v3 processors and 128GB RAM.

4.2. Simulation Results and Main Findings

The average of the metrics of our simulation results are presented in Tables 1–6. Additional simulation studies for $R^2 = 60\%$ and $R^2 = 75\%$ are presented in Tables S1–S12 in Section S4 of the supplementary materials. For peMOM and piMOM priors, the difference between the MAP and the LS only arise in the MSPE and the MSE_β but not in the other metrics. We cannot observe from the tables that SVEN and S5 perform much better than EMVS, fastBVSR and the three frequentist penalized methods in general. In most settings, the penalized methods result in many false discoveries, yet attaining similar or worse coverage probabilities compared to the Bayesian methods. Since the estimates of $\beta$ from EMVS are not sparse, it has higher MSE_β than SVEN and S5. As observed in Tables 5 and 6, fastBVSR results in large values of MSPE and MSE_β due to poor estimates of $\beta$. Also, SVEN yields competitive prediction errors and has better FDR and Jaccard indices in every case other than the group structure. As observed in Tables S1–S12, a larger value for $\lambda$ results in lower prediction error and higher Jaccard index when the signal strength is weak. Furthermore, we observe that model averaging generally produces lower MSPE and MSE_β values than the MAP models. Finally, we found that SVEN generally hits the MAP model quite quickly. However, in the absence of temperature, it jumps back and forth between the MAP model and a couple of other models. On the other hand, we observed that the use of temperature allows SVEN to explore many more models with high posterior probabilities.

For the case of group structure (Table 4) where there is a high correlation between the variables within the same group, SVEN with $w = \sqrt{n}/p$ and $\lambda = n/p^2$ and S5 both pick up only one representative variable from each group, resulting in a high false negative rate and average model size around three. Although elastic net regression successfully includes all the important variables it also includes a large number of unimportant variables and thus leads to a very high false discovery rate. However, by increasing the shrinkage to $\lambda = 200$ and increasing the prior inclusion probability to $w = 0.02$, SVEN stands out from its competitors. In fact, if important predictors are anticipated to be highly correlated, this prior information can be incorporated by choosing a larger value for $\lambda$.

In addition, we compare the computing times between S5 (with both piMOM and peMOM priors) and SVEN and find that SVEN mostly hits the MAP model faster than S5. The details are provided in Section S2 of the supplementary materials.
Table 3. Autoregressive correlation (Section 4.1.3) with $\rho = 0.6$.

| Method       | MSPE  | $\text{MSE}_\beta$ | Coverage probability (%) | Average model size | FDR (%) | FNR (%) | Jaccard index (%) |
|--------------|-------|---------------------|--------------------------|-------------------|---------|---------|------------------|
| SVEN(WAM)    | 2.1521| 0.0173              | 100                      | 3                 | 0       | 0       | 100              |
| SVEN(MAP)    | 2.1521| 0.0173              | 100                      | 3                 | 0       | 0       | 100              |
| piMOM(MAP)   | 2.1519| 0.0172              | 100                      | 3                 | 0       | 0       | 100              |
| peMOM(MAP)   | 2.1515| 0.0168              | 100                      | 3                 | 0       | 0       | 100              |
| piMOM(LS)    | 2.1521| 0.0173              | 100                      | 3                 | 0       | 0       | 100              |
| peMOM(LS)    | 2.1521| 0.0173              | 100                      | 3                 | 0       | 0       | 100              |
| fastBVSR     | 2.1961| 0.0187              | 100                      | 3.03              | 0.75    | 0       | 99.25            |
| EMVS1        | 2.2738| 0.1286              | 100                      | 6.7               | 54.57   | 0       | 45.43            |
| EMVS2        | 2.2803| 0.1419              | 100                      | 5.28              | 41.42   | 0       | 58.58            |
| EMVS3        | 2.2974| 0.1619              | 100                      | 4.33              | 28.40   | 0       | 71.60            |
| LASSO        | 2.3118| 0.1641              | 100                      | 28.16             | 76.82   | 0       | 23.19            |
| SCAD         | 2.1592| 0.0252              | 100                      | 10.33             | 45.43   | 0       | 54.57            |
| Elastic Net  | 2.4590| 0.3754              | 100                      | 54.35             | 71.70   | 0       | 28.30            |

Table 4. Group structure with three groups (Section 4.1.5).

| Method       | MSPE  | $\text{MSE}_\beta$ | Coverage probability (%) | Average model size | FDR (%) | FNR (%) | Jaccard index (%) |
|--------------|-------|---------------------|--------------------------|-------------------|---------|---------|------------------|
| SVEN(WAM)a   | 78.7067| 299.4512           | 0                        | 2.65              | 0       | 82.33   | 17.67            |
| SVEN(MAP)a   | 81.0355| 533.5387           | 0                        | 3                 | 0       | 80      | 20               |
| SVEN(WAM)b   | 82.5443| 1.8816             | 98                       | 14.99             | 0.06    | 0.13    | 99.80            |
| SVEN(MAP)b   | 82.1825| 1.6467             | 98                       | 15.02             | 0.25    | 0.13    | 99.62            |
| piMOM(MAP)   | 81.3345| 528.8252           | 0                        | 3.02              | 0.4     | 80      | 19.98            |
| peMOM(MAP)   | 81.7316| 530.0427           | 0                        | 3.02              | 0.4     | 80      | 19.98            |
| piMOM(LS)    | 81.2392| 528.7916           | 0                        | 3.02              | 0.4     | 80      | 19.98            |
| peMOM(LS)    | 81.6289| 530.1160           | 0                        | 3.02              | 0.4     | 80      | 19.98            |
| fastBVSR     | 81.1029| 326.776            | 87                       | 4.14              | 1.38    | 72.87   | 27.01            |
| EMVS1        | 77.8038| 14.9534            | 99                       | 15.05             | 0.38    | 0.07    | 99.56            |
| EMVS2        | 77.5867| 7.5430             | 100                      | 15.02             | 0.13    | 0       | 99.88            |
| EMVS3        | 80.3466| 11.3581            | 100                      | 15.02             | 0.13    | 0       | 99.88            |
| LASSO        | 84.9837| 111.852            | 0                        | 9.36              | 63.49   | 28.93   | 29.96            |
| SCAD         | 81.2506| 530.2818           | 0                        | 11.59             | 30.54   | 80      | 16.28            |
| Elastic Net  | 85.7435| 9.3598             | 100                      | 68.03             | 65.94   | 0       | 34.06            |

$^a\lambda = n/p^2, w = \sqrt{n/p}; ^b\lambda = 200, w = 0.02.$

Table 5. Factor model with two factors (Section 4.1.4).

| Method       | MSPE  | $\text{MSE}_\beta$ | Coverage probability (%) | Average model size | FDR (%) | FNR (%) | Jaccard index (%) |
|--------------|-------|---------------------|--------------------------|-------------------|---------|---------|------------------|
| SVEN(WAM)    | 42.9106| 0.3892             | 100                      | 5                 | 0       | 0       | 100              |
| SVEN(MAP)    | 42.9103| 0.3891             | 100                      | 5                 | 0       | 0       | 100              |
| piMOM(MAP)   | 42.8731| 0.3724             | 100                      | 5                 | 0       | 0       | 100              |
| peMOM(MAP)   | 42.9491| 0.4211             | 100                      | 5.01              | 0.17    | 0       | 99.83            |
| piMOM(LS)    | 42.9103| 0.3891             | 100                      | 5                 | 0       | 0       | 100              |
| peMOM(LS)    | 42.9361| 0.4083             | 100                      | 5.01              | 0.17    | 0       | 99.83            |
| fastBVSR     | 67.0982| 19.9837            | 87                       | 6.14              | 18.52   | 3.60    | 79.89            |
| EMVS1        | 64.6038| 22.1115            | 95                       | 19.13             | 66.40   | 1.00    | 33.59            |
| EMVS2        | 56.7884| 14.5042            | 95                       | 11.58             | 45.34   | 1.00    | 54.64            |
| EMVS3        | 53.4840| 11.3980            | 94                       | 9.08              | 34.73   | 1.20    | 65.20            |
| LASSO        | 54.2887| 11.2984            | 99                       | 66.37             | 91.81   | 0.20    | 7.03             |
| SCAD         | 43.1155| 0.5743             | 100                      | 11.56             | 27.99   | 0       | 72.01            |
| Elastic Net  | 62.4327| 19.4566            | 99                       | 54.29             | 95.90   | 0.20    | 4.10             |

5. Real Data Analysis

We examine the practical performance of our proposed method by applying it to a real data example. Cook et al. (2012) conducted a genome-wide association study on starch, protein, and kernel oil content in maize. The original field trial at Clayton, NC in 2006 consisted of more than 5000 inbred lines and check varieties primarily coming from a diverse IL panel consisting of 282 founding lines (Flint-Garcia et al. 2005). Because the dataset comes from a field trial, the responses could be spatially autocorrelated. Thus, we use a random row-column adjustment to obtain the adjusted phenotypes of the varieties. However, marker information of only $n = 3951$ of these varieties are available from the panzea project (https://www.panzea.org/) which provide information on 546,034 single nucleotide polymorphisms (SNP) markers after removing duplicates and SNPs with minor allele frequency (MAF) less than 5%. We use the protein content as our phenotype for conducting the association study. Because the inbred varieties are bi-allelic, we
store the marker information in a sparse format by coding the minor alleles by one and major alleles by zero.

### 5.1. Marker Selection After Screening

We compare our method to S5, fastBVSR and the three penalized regression methods (LASSO, Elastic Net and SCAD). Since both R packages *BayesS5* (version 1.31) and *glmnet* (version 2.0-18) do not work on this massive dataset, we perform a screening of these markers before conducting variable selection so as to reduce the dimension of the data. We randomly split the data into a training set of size $n_0 = 3851$ and testing set of size 100. Then we use high dimensional ordinary least squares projection (HOLP) screening method (Wang and Leng 2016) to preserve $p_0 = 3851$ markers. Note that the screening sets are formed by controlling the MAF of each marker to be no less than 1.5%. Because markers tend to be highly correlated, we use SVEN with $w = 1/p_0$ but with two choices of $\lambda : \lambda = \sqrt{n_0}$ (high shrinkage) and $\lambda = n_0/p_0$ (low shrinkage); and with $m = 3$ and $N = 50$ for selecting the markers. In our experience, both the model size and MSPE lie in between the respective reported values for other intermediate values of $\lambda$ that we have tried. We repeat the entire process 50 times—each time computing the MSPE and the model size from each method. The peMOM nonlocal prior in S5 failed to provide any result even after 100 hrs of running, and S5 with the piMOM prior failed to provide any result in three cases. The fastBVSR algorithm ran successfully in only 39 out of the 50 cases, while the complex iterative factorization at the core of fastBVSR encountered floating point errors in the remaining 11 cases and could not produce any result. In contrast, SVEN faced no difficulties and produced the results within reasonable time.

The boxplots of these MSPEs are shown in Figure 1 along with the average model sizes. Overall SVEN, S5 and SCAD perform significantly better than the lasso, the elastic net regression and fastBVSR and produce smaller MSPE with smaller model sizes. Moreover, SVEN and S5 produce comparable MSPE values but SVEN results in more parsimonious models. In addition, the outlying value for SVEN corresponds to one of the three cases where S5 failed to provide any result. Finally, SVEN with high shrinkage produces slightly smaller MSPE but double model size than with low shrinkage.

![Figure 1. Boxplots for MSPE using SVEN, S5, fastBVSR, LASSO and Elastic net after screening. $w = 1/p_0$, $\lambda = \sqrt{n_0}$; $w = 1/p_0$, $\lambda = n_0/p_0^2$.](image)

### 5.2. Marker Selection on the Entire Dataset

Unlike other variable selection methods, SCAD and SVEN can be successfully directly applied to the whole dataset without any prescreening. We ran the SVEN 50 times again with the temperature schedule described in Section 2.2.2 with $m = 3$, with $N = 100$ iterations per temperature, each time starting with a different random seed. Initially, we use $w = 1/p$ and try several values of $\lambda$ as done in Section 5.1. The best models from these 50 runs vary suggesting the posterior surface is severely multimodal. With $\lambda = n/p^2$, we find that although the sizes of these best models remain around nine, the number of unique markers included in at least one of these 50 best models is over 30 (for SCAD these numbers were $> 40$ and $> 60$, respectively). Other larger values of $\lambda$ produce even larger models and more unique variables. Interestingly, by taking a further look into the markers it identified, we discovered that the presence of some of these markers in a model is always accompanied by the absence of certain other markers. More specifically, some pairs and triplets of the markers are never included simultaneously in the MAP models but the frequencies at which they are selected add up to 50. Thus, to achieve more parsimonious models, we reduce $w$ to $1/p^2$ and use $\lambda = n/p^2$. Using such a small $w$, the sizes of the best models from each run reduce to around four and the number of unique markers that are included at least once in the 50 best models comes down to eight. To verify our conjecture on the correlations between these markers, we calculated the pairwise partial correlations between these eight markers. It turns out that the pairs of markers that are never included in the same model are indeed relatively highly partially correlated with values ranging from 29% to 60% (Figure S2 of the supplementary materials) than other pairs where the partial

---

**Table 6.** Extreme correlation (Section 4.1.6).

| Method          | MSPE | MSE $\beta$ | Coverage probability (%) | Average model size | FDR (%) | FNR (%) | Jaccard Index (%) |
|-----------------|------|-------------|---------------------------|--------------------|---------|---------|------------------|
| SVEN(WAM)       | 14.0754 | 0.1571     | 100                       | 5                  | 0       | 0       | 100              |
| SVEN(MAP)       | 14.0757 | 0.1569     | 100                       | 5                  | 0       | 0       | 100              |
| piMOM(MAP)      | 14.0732 | 0.1547     | 100                       | 5                  | 0       | 0       | 100              |
| peMOM(MAP)      | 14.0750 | 0.1562     | 100                       | 5                  | 0       | 0       | 100              |
| piMOM(LS)       | 14.0757 | 0.1569     | 100                       | 5                  | 0       | 0       | 100              |
| peMOM(LS)       | 14.0757 | 0.1569     | 100                       | 5                  | 0       | 0       | 100              |
| fastBVSR        | 31.2771 | 32.1993    | 97                        | 5.03               | 18.65   | 0.0     | 91.54            |
| EMVS1           | 14.7568 | 2.6871     | 100                       | 5.6                | 8.44    | 0       | 95.63            |
| EMVS2           | 14.4561 | 1.5340     | 100                       | 5.09               | 1.45    | 0       | 98.55            |
| EMVS3           | 14.4218 | 1.3793     | 100                       | 5.03               | 0.5     | 0       | 99.5             |
| LASSO           | 15.3893 | 2.8732     | 100                       | 13.77              | 61.13   | 0       | 23.68            |
| SCAD            | 14.0793 | 0.1547     | 100                       | 5.49               | 5.29    | 0       | 94.71            |
| Elastic Net     | 15.5365 | 3.7949     | 100                       | 65.87              | 86.75   | 0       | 13.25            |
correlation values are always $< 18\%$. Thus, SVEN is able to identify pairs of markers that have similar effect on the response.

Next, we study the performance and the widths of the 90% and 95% Z-PIs and MC-PIs described in Section 3. To that end, we randomly split the entire data into a training set of size $n = 3751$ under the constraint that the MAF of each marker is at least 1.5% and a testing set of size 200. We also remove any duplicated markers from the training set, which results in a smaller $p = 544,211$. We generate 10,000 samples from the approximate posterior predictive distribution (14) to compute the MC-PIs. We find that the Z-PIs and the MC-PIs attain identical coverage rates and these are found to be 91% and 95% for the 90% and 95% prediction intervals, respectively. The boxplots of the widths of the 200 intervals from each method are presented in Figure 2. We find that widths of the the Z-PIs are less variable compared to the same for the MC-PIs. It is encouraging to see that despite nonnormality of the posterior prediction distribution, the Z-PIs are better than simulation based intervals.

6. Conclusion
In this article, we introduce a Bayesian variable selection method with embedded screening for ultrahigh-dimensional settings. The model used here is a hierarchical model with well-known spike and slab priors on the regression coefficients. Use of the degenerate spike prior for inactive variables not only results in sparse estimates of regression coefficients and (much) lesser computational burden, it also allows us to establish strong model selection consistency under somewhat weaker conditions than Narisetty and He (2014). In particular, we prove that the posterior probability of the true model converges to one even when the norm of mean effects solely due to the unimportant variables diverge. On the other hand, our method crucially hinges on the fact that model probabilities are available in closed form (up to a normalizing constant) which is due to the use of Gaussian slab priors on active covariates. We propose a scalable variable selection algorithm with an inbuilt screening method that efficiently explores the huge model space and rapidly finds the MAP model. The screening is actually model based in the sense that it is performed on a set of candidate models rather than the set of potential variables. The algorithm also incorporates the temperature control into a neighbor based stochastic search method. We use fast Cholesky update to efficiently compute the (unnormalized) posterior probabilities of the neighboring models. Since mean and variance of the posterior predictive distribution are shown to be means of analytically available functions of the models, a derivative of the proposed method is construction of novel prediction intervals for future observations. Both Z based intervals and simulation based intervals are derived and compared. In the context of the real data analysis, we observe that Z based prediction intervals lead to the same coverage rates and their widths are less variable. The extensive simulations studies in Section 4 and the real data analysis in Section 5 demonstrate the superiority of the proposed method compared with the other state of the art methods, even though the hyperparameters in the proposed method are not carefully tuned. Among the Bayesian methods used for comparison, the package associated with the proposed algorithm seems to be the only one that can be directly applied to datasets of dimension as high as the one analyzed here with the computing resources mentioned before.

Variable selection and consistency of the resulting posteriors for high dimensional generalized linear models are considered in Liang, Song, and Yu (2013). It would be interesting to extend our method to the generalized linear regression model setup. The dataset we have used comes from an agricultural field trial and hence the observations are expected to be spatially autocorrelated. Although we have used a two stage procedure by first obtaining spatially adjusted genotypic effects, our model can be extended to include spatial random effects (Dutta and Mondal 2014). Also, in many applications, the covariates may have a nonlinear effect on the response and our method could be extended to additive models.

Appendix: Model Selection Consistency

We consider the ultra-high dimensional setting where the number of variables $p$ is allowed to vary subexponentially with the sample size. As established by Narisetty and He (2014) the slab precision $\lambda$ also needs to vary with $n$ for strong model selection consistency. In order to state the assumptions and the main results, we use the following notations. Abusing notation, we interchangeably use a model $\gamma$ either as a $p$-dimensional binary vector or as a set of indices of nonzero entries of the binary vector. For models $\gamma$ and $s$, $\gamma \prec s$ and $\gamma \lor s$ denote the union and intersection of $\gamma$ and $s$, respectively. For two real sequences $(a_n)$ and $(b_n)$, $a_n \sim b_n$ means $a_n/b_n \to c$ for some constant $c > 0$. $a_n \geq b_n$ (or $b_n \leq a_n$) means $b_n = O(a_n)$; $a_n \geq b_n$ (or $b_n < a_n$) means $b_n = o(a_n)$. Also for any matrix $A$, let $\sigma_{\text{min}}(A)$ and $\sigma_{\text{max}}(A)$ denote its minimum and maximum eigenvalues, respectively, and let $\sigma_{\text{min}}^*(A)$ be its minimum nonzero eigenvalue. Again, abusing notations, for two real numbers $a$ and $b$, $a \lor b$ and $a \land b$ denote $\max(a,b)$ and $\min(a,b)$, respectively. Define $r_\gamma = \text{rank}(X_\gamma)$ and for $v > 0$, $r_\gamma^v = r_\gamma \land u_n(v)$ where $u_n(v) = p\land \frac{n}{(2 + v)\log p}$ and $n^*_{\gamma}(v) = \inf_{|\gamma| \leq u_n(v)} \sigma^*_{\text{min}}(X_\gamma^TX_\gamma/n)$. Finally for any fixed positive integer $J$, define $\Delta_n(I) = \inf_{|\gamma| \leq J} ||(I - P_\gamma)X_\gamma \beta_\gamma||^2$.
where $P_Y = X_Y^T (X_Y^T X_Y)^{-1} X_Y^T$ is the orthogonal projection matrix onto the column space of $X_Y$ and $\| \cdot \|$ denotes the $L_2$ norm. Here, $A^\dagger$ denotes the Moore-Penrose inverse of $A$. We assume the following set of conditions.

**Condition 1.** $p = e^{a_0 n}$ for some $a_0 \to 0$ as $n \to \infty$, that is, $\log p = o(n)$.

**Condition 2.** $n / x \sim (n \vee p^{2+\varepsilon})$ for some $\delta > 0$, and $w \sim p^{-1}$.

**Condition 3.** $y = \beta_0 I_n + X_T \beta + X_T \beta_p + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I_n)$, the true model $t$ is fixed and $|X_T \beta_p| \leq \sqrt{\log p}$.

**Condition 4.** For $\delta$ given in C2, there exists $J > 1 + 8/\delta$ such that $\Delta_n(f) > \log(\sqrt{n} \vee p)$, and for some $v < \delta, k < (J-1)/2$,

$$n \min(v) \geq \left(n \min(v) \sqrt{n} \vee \sqrt{p}^{-k}\right).$$

**Condition 5.** For some positive constants $a_0$ and $b_0, a_0 < \min \left(\frac{X_Y^T X_Y}{n} \right) < b_0 \forall n$.

The condition C2 states that the conditional distribution of $\beta_i$ given $\gamma_i = 1$ is diffused in the sense that it’s conditional prior variance goes to infinity at a particular rate. The condition C3 greatly relaxes the boundedness assumption on $|X_T \beta_p|$ in Narisetty and He (2014), by slightly strengthening the identifiability condition C4. Yang, Wainwright, and Jordan (2016) obtained similar results under $g$-priors on $\beta$ but as mentioned by them our independence prior is “a more realistic choice.” Moreover, Yang, Wainwright, and Jordan (2016) assumed that $\alpha_{\min}(X_Y^T X_Y/n)$ is bounded away from zero for all models $\gamma$ of size at most $O(n/\log p)$, which is unrealistic for example, even when entries of $X$ are iid $N(0,1)$, inf $\leq x \leq x$ $X_Y^T X_Y/n$ converges to zero in probability. Because of the degenerated form of the spike priors, the regularity assumptions on the submatrices of the design matrix $X$ in C4 relax the assumptions on the bound on their largest eigenvalues. Narisetty and He (2014) showed that if the rows of $X$ are independent isotropic sub-Gaussian random vectors then C4 holds with overwhelmingly large probability (see also Chen and Chen 2008; Kim, Kwon, and Choi 2012; Shin, Bhattacharya, and Johnson 2018). The regularity assumption for the true model C5 is standard and has been used in both Narisetty and He (2014) and Cao, Khare, and Ghosh (2020) without being explicitly stated.

Note that the condition C3 does not explicitly specify the true model $t$ and the relaxation to allow higher noise $|X_T \beta_p|$ warrants a validation of the identifiability of $t$. To that end, suppose on the contrary that it is possible to include some variables, say $s$ from $t'$ into the true model and still maintain the conditions C1–C5 for both $t$ and $t' \neq s$ as true models for every $n$. Then condition C4 with $t' = t$ (now excluding the apparently true variable $s$) would imply $\| (I - P_t) X_T \beta_p \|^2 = \| (I - P_t) X_T \beta_p + X_T \beta_p \|^2 \geq \| (I - P_t) X_T \beta_p \|^2 > \log(p \vee \sqrt{n})$. Here, the first equality follows from the fact that $P_t X_T = X_T$. But because $I - P_t$ is symmetric and idempotent,

$$\| X_T \beta_p \|^2 \geq \| (I - P_t) X_T \beta_p \| > \sqrt{\log(p \vee \sqrt{n})} \geq \sqrt{\log(p)}.$$  \hspace{1cm} (15)

However, condition C3 for $t \neq s$ implies $\| X_T \beta_p \|^2 \leq \sqrt{\log(p)}$. This with (15) implies that

$$\| X_T \beta_p \| = \| X_T \beta_p + X_T \beta_p \|^2 \geq \| X_T \beta_p \| > \sqrt{\log(p)},$$

which contradicts condition C3. We now present the strong model selection consistency results.

**Theorem 1.** Assume conditions C1–C5 hold and that $\sigma^2$ is known. Then the posterior probability of the true model, $f(t | y, \sigma^2) \to 1$ in probability as the sample size $n$ approaches $\infty$.

**Proof.** The proof is given in Section S6 of the supplementary materials.

Note that the statement of Theorem 1 is equivalent to $1 - f(t | y, \sigma^2) / f(t | y, \sigma^2) \to 0$ in probability as $n \to \infty$. The proof of Theorem 1 also provides the rate of convergence given by,

$$1 - f(t | y, \sigma^2) / f(t | y, \sigma^2) \leq \exp\left\{ -\rho_n + \rho_n (J-1)/2 \right\} + \exp\left\{ -\rho_n \log(\sqrt{n} \vee p) \right\}$$

with probability greater than $1 - [2 \exp(-cn) + 2 \exp(-c \log p)]$ for some positive constants $c, v', v'', c', c''$, where $\rho_n = p^{-3/2} \wedge (p^{1+\delta/2} / \sqrt{n})$. It is encouraging that despite relaxing the boundedness condition on $|X_T \beta_p|$, the rate of convergence remains the same as in Narisetty and He (2014).

However, in practice $\sigma^2$ is typically never known. In this case, we need a further assumption that assigns a prior probability of zero on $\hat{M} = \{ y : r_Y > r_t + n/(2 + v') \log p \}$ for some $v' > v \vee (2/\delta)$.

**Condition 6.** For some $v > 0$ and $v' > v \vee (2/\delta)$, $P (y \in \hat{M}) = 0$.

This condition is same as in Narisetty and He (2014) and also equivalent to the assumptions on the prior model sizes in Shin, Bhattacharya, and Johnson (2018) and Cao, Khare, and Ghosh (2020).

**Theorem 2.** Assume conditions C1–C6 hold. Then the posterior probability of the true model, $f(t | y) \to 1$ in probability as the sample size $n$ approaches $\infty$.

**Proof.** The proof is given in Section S7 of the supplementary materials.

**Supplemental Materials**

The supplemental materials contain computer code, additional details on computations and the real data analysis, further simulation studies, and proofs of the theoretical results stated in the article.

**Acknowledgments**

The authors acknowledge the very helpful comments of the Editor, the Associate Editor and a referee on an earlier version of this article. This work was partially supported by the USDA/NIFA Hatch Project IOW03717. The content presented in this article are those of the authors and do not necessarily reflect the views of USDA.

**Funding**

This work was partially supported by the USDA/NIFA Hatch Project IOW03717. The content presented in this article are those of the authors and do not necessarily reflect the views of USDA.

**References**

Barbiere, M. M., and Berger, J. O. (2004), “Optimal Predictive Model Selection,” *The Annals of Statistics*, 32, 870–897. [65]
Bertsimas, D., King, A., and Mazumder, R. (2016), "Best Subset Selection via a Modern Optimization Lens," *The Annals of Statistics*, 44, 813–852. [61]

Bondell, H., and Reich, B. (2008), "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with Oscar," *Biometrics*, 64, 115–123. [61]

Cao, X., Khare, K., and Ghosh, M. (2020), "High-dimensional Posterior Consistency for Hierarchical Non-local Priors in Regression," *Bayesian Analysis*, 15, 241–262. [62,63,64,72]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771. [72]

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., Buckler, E. S., and Flint-Garcia, S. A. (2012), "Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels," *Plant Physiology*, 158, 824–834. [78]

Datta, A., and Zou, H. (2017), "Cocolasso for High-Dimensional Error-in-Variables Regression," *The Annals of Statistics*, 45, 2400–2426. [61]

Dutta, S., and Mondal, D. (2014), "An h-likelihood Method for Spatial Mixed Linear Model Based on Intrinsic Autoregressions," *Journal of the Royal Statistical Society*, Series B, 77, 699–726. [71]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [61]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [62,64]

Flint-Garcia, S. A., Thuilliet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2005), "Maize Association Population: A High-Resolution Platform for Quantitative Trait Locus Dissection," *The Plant Journal*, 44, 1054–1064. [69]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [61] (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [61]

Hans, C. (2010), "Model Uncertainty and Variable Selection in Bayesian Lasso Regression," *Statistics and Computing*, 20, 221–229. [61] (2011), "Elastic Net Regression Modeling with the Orthant Normal Prior," *Journal of the American Statistical Association*, 106, 1383–1393. [61]

Hans, C., Dobra, A., and West, M. (2007), "Shotgun Stochastic Search for "large p" Regression," *Journal of the American Statistical Association*, 102, 507–516. [62,63,64]

Huang, J., Jiao, Y., Liu, Y., and Lu, X. (2018), "A Constructive Approach to L0 Penalized Regression," *The Journal of Machine Learning Research*, 19, 403–439. [61]

Ishwaran, H., and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773. [61]

Johnson, V. E., and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660. [61,63]

Kim, Y., Choi, H., and Oh, H.-S. (2008), "Smoothly Clipped Absolute Deviation on High Dimensions," *Journal of the American Statistical Association*, 103, 1665–1673. [61]

Kim, Y., Kwon, S., and Choi, H. (2012), "Consistent Model Selection Criteria on High Dimensions," *Journal of Machine Learning Research*, 13, 1037–1057. [72]

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983), "Optimization by Simulated Annealing," *Science*, 220, 671–680. [65]

Kojima, M., and Komaki, F. (2016), "Determinantal Point Process Priors for Bayesian Variable Selection in Linear Regression," *Statistica Sinica*, 26, 97–117. [65]

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *Bayesian Analysis*, 5, 369–412. [61]

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [61]

Liang, E., Song, Q., and Yu, K. (2013), "Bayesian Subset Modeling for High-Dimensional Generalized Linear Models," *Journal of the American Statistical Association*, 108, 589–606. [71]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection with the Lasso," *The Annals of Statistics*, 34, 1436–1462. [67]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [61,62]

Narisetty, N. N., and He, X. (2014), "Bayesian Variable Selection with Shrinking and Diffusing Priors," *The Annals of Statistics*, 42, 789–817. [61,62,63,67,71,72]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [61]

Rocková, V., and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [62,67]

Roy, V., and Chakraborty, S. (2017), "Selection of Tuning Parameters, Solution Paths and Standard Errors for Bayesian Lassos," *Bayesian Analysis*, 12, 753–778. [61]

Roy, V., Tan, A., and Flegal, J. (2018), "Estimating Standard Errors for Importance Sampling Estimators with Multiple Markov Chains," *Statistica Sinica*, 28, 1079–1101. [63]

Shin, M., Bhattacharya, A., and Johnson, V. E. (2018), "Scalable Bayesian Variable Selection Using Nonlocal Prior Densities in Ultrahigh-Dimensional Settings," *Statistica Sinica*, 28, 1053–1078. [62,63,64,65,67,72]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [61]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [67]

Wang, X., and Leng, C. (2016), "High Dimensional Ordinary Least Squares Projection for Screening Variables," *Journal of the Royal Statistical Society*, Series B, 78, 589–611. [62,67,70]

Xu, X., and Ghosh, M. (2015), "Bayesian Variable Selection and Estimation for Group Lasso," *Bayesian Analysis*, 10, 909–936. [61]

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016), "On the Computational Complexity of High-Dimensional Bayesian Variable Selection," *The Annals of Statistics*, 44, 2497–2532. [62,72]

Yuan, M., and Lin, Y. (2005), "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models," *Journal of the American Statistical Association*, 100, 1215–1225. [61]

Zanella, G., and Roberts, G. (2019), "Scalable Importance Tempering and Bayesian Variable Selection," *Journal of the Royal Statistical Society*, Series B, 81, 489–517. [62]

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, eds. P. K. Goel and A. Zellner, pp. 233–243, Amsterdam: Elsevier Science. [65]

Zhou, Q., and Guan, Y. (2019), "Fast Model-Fitting of Bayesian Variable Selection Regression Using the Iterative Complex Factorization Algorithm," *Bayesian Analysis*, 14, 573–594. [62,67]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [61]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [61,67]