Taylor & Francis
Taylor & Francis Group



Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Variational Bayes for High-Dimensional Linear Regression With Sparse Priors

Kolyan Ray & Botond Szabó

To cite this article: Kolyan Ray & Botond Szabó (2022) Variational Bayes for High-Dimensional Linear Regression With Sparse Priors, Journal of the American Statistical Association, 117:539, 1270-1281, DOI: 10.1080/01621459.2020.1847121

To link to this article: https://doi.org/10.1080/01621459.2020.1847121

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 14 Jan 2021.

Submit your article to this journal

Article views: 5568

View related articles

View Crossmark data

Citing articles: 5 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 539, 1270–1281: Theory and Methods
https://doi.org/10.1080/01621459.2020.1847121

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Variational Bayes for High-Dimensional Linear Regression With Sparse Priors

Kolyan Ray[a] and Botond Szabó[b*]

[a]Department of Mathematics, Imperial College London, London, UK; [b]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

ABSTRACT

We study a mean-field spike and slab variational Bayes (VB) approximation to Bayesian model selection priors in sparse high-dimensional linear regression. Under compatibility conditions on the design matrix, oracle inequalities are derived for the mean-field VB approximation, implying that it converges to the sparse truth at the optimal rate and gives optimal prediction of the response vector. The empirical performance of our algorithm is studied, showing that it works comparably well as other state-of-the-art Bayesian variable selection methods. We also numerically demonstrate that the widely used coordinate-ascent variational inference algorithm can be highly sensitive to the parameter updating order, leading to potentially poor performance. To mitigate this, we propose a novel prioritized updating scheme that uses a data-driven updating order and performs better in simulations. The variational algorithm is implemented in the R package `sparsevb`. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2019
Accepted November 2020

KEYWORDS
Model selection; Oracle inequalities; Sparsity; Spike-and-slab prior; Variational Bayes

1. Introduction

Inference under sparsity constraints has found many applications in statistics and machine learning (Mitchell and Beauchamp 1988; Tipping 2001). Perhaps the most widely applied such model is sparse linear regression, where we observe

$$Y = X\theta + Z, \tag{1}$$

where $Y \in \mathbb{R}^n$, X is a given, deterministic $n \times p$ design matrix, $\theta \in \mathbb{R}^p$ is the parameter of interest and $Z \sim N_n(0, I_n)$ is additive Gaussian noise. We are interested in the *sparse high-dimensional* setting, where $n \leq p$ and typically $n \ll p$, and many of the coefficients θ_i are (close to) zero.

From a Bayesian perspective, perhaps the most natural way to impose sparsity is through a *model selection* prior, which assigns probabilistic weights to each potential model, that is, each subset of $\{1, \ldots, p\}$ corresponding to selecting the nonzero coordinates of $\theta \in \mathbb{R}^p$. This is one of the most widely used approaches within the Bayesian community (Mitchell and Beauchamp 1988; George and McCulloch 1993; West 2003; Efron 2008) and includes the popular spike-and-slab prior, which is often considered the gold standard in sparse Bayesian linear regression. Such priors have been shown to perform well for estimation and prediction (Johnstone and Silverman 2004; Castillo and van der Vaart 2012; Castillo, Schmidt-Hieber, and van der Vaart 2015; Chae, Lin, and Dunson 2019), uncertainty quantification (Ray 2017; Castillo and Szabó 2020), and multiple hypothesis testing (Castillo and Roquain 2020), see Banerjee, Castillo, and Ghosal (2020) for a recent review.

However, while these priors perform excellently both empirically and theoretically, the discrete model selection component of the prior can make computation hugely challenging. For $\theta \in \mathbb{R}^p$, inference using the spike-and-slab prior generally involves a combinatorial search over all 2^p possible models, a hugely expensive task for even moderate p. Fast algorithms for exact posterior computation are thus usually restricted to the diagonal design case (Castillo and van der Vaart 2012; van Erven and Szabo 2020), while Markov chain Monte Carlo methods are known to have problems mixing for typical problem sizes of interest (Griffin, Latuszynski, and Steel 2017).

A popular scalable alternative is variational Bayes (VB), which recasts posterior approximation as an optimization problem. One minimizes the VB objective function, consisting of the Kullback–Leibler (KL) divergence between a family of tractable distributions, called the variational family, and the posterior. Though the resulting approximation does not provide exact Bayesian inference, picking a computationally convenient variational class can dramatically increase scalability (see, e.g., Blei, Ng, and Jordan 2003; Hoffman et al. 2013). An especially popular variational family consists of distributions under which the model parameters are independent, so called *mean-field VB*. For a nice recent review of VB, see Blei, Kucukelbir, and McAuliffe (2017).

In this work, we consider a mean field family consisting of distributions independently assigning each coordinate of θ an independent mixture of a Gaussian and Dirac mass at zero, thereby mirroring the form of the spike-and-slab prior (but

CONTACT Botond Szabó botond.szabo@unibocconi.it Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands.
*Current affiliation: Bocconi University

This article was originally published with errors, which have now been corrected in the online version. Please see Correction (http://dx.doi.org/10.1080/01621459.2021.1957322)

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

crucially not the form of the posterior). Such a computational relaxation is significant, reducing the posterior dimension to a much more tractable $O(p)$. This is a natural approximation since it keeps the discrete model selection aspect and many of the interpretable features of the original posterior, for example, access to posterior probabilities of submodels and inclusion probabilities of particular covariates. This sparse variational family has been applied in practice (Logsdon, Hoffman, and Mezey 2010; Titsias and Lázaro-Gredilla 2011; Carbonetto and Stephens 2012; Huang, Wang, and Liang 2016; Ormerod, You, and Müller 2017), but comes with few theoretical guarantees.

We study this VB procedure under the frequentist assumption that the data Y has been generated according to a given sparse parameter θ_0. Under standard conditions on the design matrix, we obtain refined oracle type contraction rates for the mean-field VB approximation of model selection priors. As a consequence, these imply that the VB posterior performs optimally regarding both estimation of a sparse θ and for prediction of the response vector. This provides a theoretical justification for this attractive approximation algorithm in a sparsity context.

While similar VB approaches have been applied in the methodological literature (Logsdon, Hoffman, and Mezey 2010; Titsias and Lázaro-Gredilla 2011; Carbonetto and Stephens 2012; Huang, Wang, and Liang 2016; Ormerod, You, and Müller 2017), our contribution also possesses a crucial methodological difference. These existing works typically use Gaussian slabs for the prior, which allows analytic evaluation of certain formulas in the variational algorithm leading to fast optimization. However, Gaussian slabs are inappropriate for recovering the true signal θ_0 since the *true* underlying posterior performs excessive shrinkage causing poor performance (Castillo and van der Vaart 2012). One cannot typically expect a VB approximation based on a poorly performing underlying posterior to perform well for recovery. We instead consider Laplace slabs for the prior, which result in optimal recovery when using the true posterior (Castillo and van der Vaart 2012; Castillo, Schmidt-Hieber, and van der Vaart 2015). We are thus using a similar variational family to estimate a *different posterior distribution* compared to previous works. Another way to correct the original posterior is to explicitly shift the posterior mean using an empirical Bayes approach (Martin, Mess, and Walker 2017; Martin and Tang 2019; Belitser and Ghosal 2020; Belitser and Nurushev 2020).

We provide the methodological details for applying the widely used coordinate-ascent variational inference (CAVI) algorithm (Blei, Kucukelbir, and McAuliffe 2017) with Laplace slabs and investigate our method numerically on both simulated and real world ozone interaction data. As predicted by the theory, our method performs well in a number of settings and typically outperforms VB approaches with prior Gaussian slabs. In fact, we find that our approach generally performs at least as well as other state-of-the-art Bayesian variable selection methods. We have implemented our algorithm in the R-package `sparsevb` (Clara, Szabo, and Ray 2020).

Our simulations also show that the CAVI algorithm is highly sensitive to the updating order of the parameters. Since the VB objective function is nonconvex and typically has multiple local minima, a poorly chosen updating order can trap the algorithm near a highly suboptimal local minimum causing poor performance. To resolve this, we propose a novel *prioritized* update scheme where we base the CAVI parameter update order on the estimated size of the coefficients via a preliminary estimator. Our simulations indicate that such a data-driven updating order performs better than using either a naive or random update order and provides more robustness against being trapped at a suboptimal local minimum. This idea is applicable beyond the present setting and may be useful for other CAVI approaches.

1.1. Related Work

While VB has found increasing usage in practice, its theoretical understanding is still in the early stages. In low-dimensional settings, some Bernstein–von Mises type results have been derived (Lu, Stuart, and Weber 2017; Wang and Blei 2019), while in high-dimensional and nonparametric settings, first results have only recently appeared (Pati, Bhattacharya, and Yang 2018; Zhang and Gao 2020; Zhang and Zhou 2020). There has also been theoretical work on studying variational approximations to *fractional posteriors*, which down-weight the likelihood (Alquier and Ridgway 2020; Yang and Martin 2020; Yang, Pati, and Bhattacharya 2020). The articles (Zhang and Gao 2020; Pati, Bhattacharya, and Yang 2018; Yang, Pati, and Bhattacharya 2020) provide general proof methods which employ the classical prior mass and testing approach of Bayesian nonparametrics (Ghosal, Ghosh, and van der Vaart 2000). However, since it is known that posterior convergence rates, let alone oracle rates as we derive here, for model selection priors cannot easily be established using this approach (Castillo and van der Vaart 2012; Castillo, Schmidt-Hieber, and van der Vaart 2015), their results do not apply to our setting. We have extended some of the present results to high-dimensional logistic regression in follow up work (Ray, Szabo, and Clara 2020).

1.2. Organization

In Section 2, we give details of the prior, variational approximation and conditions on the design matrix. We present our main results in Section 3, details of the VB algorithm in Section 4, numerical results in Section 5, and conclusions in Section 6. In the supplementary materials, we give additional numerical results in Section A, full oracle results and proofs in Section B, additional methodological details in Section C, and further discussion of the design matrix assumptions in Section D.

1.3. Notation

Let P_θ be the probability distribution of the observation Y arising in model (1) and let E_θ denote the corresponding expectation. For two probability distributions P, Q, $\mathrm{KL}(P||Q) = \int \log \frac{dP}{dQ} dP$ denotes the KL divergence. For $x \in \mathbb{R}^d$, we write $||x||_2 = (\sum_{i=1}^d |x_i|^2)^{1/2}$ for the Euclidean norm. For a vector $\theta \in \mathbb{R}^p$ and a subset $S \subseteq \{1, \ldots, p\}$ of indices, set θ_S to be the vector $(\theta_i)_{i \in S}$ in $\mathbb{R}^{|S|}$, where $|S|$ denotes the cardinality of S. Further let $S_\theta = \{i : \theta_i \neq 0\}$ be the set of nonzero coefficients of θ. We will often write $S_0 = S_{\theta_0}$ and $s_0 = |S_{\theta_0}|$, where θ_0 is the true vector. For $X_{\cdot i}$ the ith column of X, set

$$||X|| := \max_{1 \leq i \leq p} ||X_{\cdot i}||_2 = \max_{1 \leq i \leq p} (X^T X)_{ii}^{1/2}. \tag{2}$$

2. Prior, Variational Families, and Design Matrix

2.1. Model Selection Priors

We first present the desirable, but computationally challenging, model selection priors that underlie our VB approximation. Consider a prior for $\theta \in \mathbb{R}^p$ that first selects a *dimension* s from a prior π_p on $\{0, \ldots, p\}$, then uniformly selects a random subset $S \subset \{1, \ldots, p\}$ of cardinality $|S| = s$ and lastly a set of nonzero values $\theta_S = \{\theta_i : i \in S\}$ from a prior density g_S on $\mathbb{R}^{|S|}$. Since it is known that the "slab" distribution should have exponential tails or heavier to achieve good recovery (Castillo and van der Vaart 2012), we restrict to the case where $g_S = \prod_{i \in S} \text{Lap}(\lambda)$ is a product of centered Laplace densities with parameter $\lambda > 0$ on $\mathbb{R}^s$. This yields the hierarchical prior:

$$\begin{aligned} s &\sim \pi_p(s) \\ S||S| = s &\sim \text{Unif}_{p,s} \\ \theta_i &\overset{\text{ind}}{\sim} \begin{cases} \text{Lap}(\lambda), & i \in S, \\ \delta_0, & i \notin S, \end{cases} \end{aligned} \tag{3}$$

where $\text{Unif}_{p,s}$ selects S from the $\binom{p}{s}$ possible subsets of $\{1, \ldots, p\}$ of size s with equal probability and δ_0 denotes the Dirac mass at zero. Since we wish the prior to perform model selection via the prior π_p on the dimension s rather than via shrinkage of the Laplace distribution, the choice of prior π_p is crucial. The aim is to select a distribution which sufficiently downweights large models while simultaneously placing enough mass to the true model. Following Castillo, Schmidt-Hieber, and van der Vaart (2015), we select an exponentially decreasing prior: we assume that there are constants $A_1, A_2, A_3, A_4 > 0$ with

$$A_1 p^{-A_3} \pi_p(s-1) \le \pi_p(s) \le A_2 p^{-A_4} \pi_p(s-1), \tag{4}$$

$s = 1, \ldots, p$. Assumption (4) is satisfied by a variety of priors, including those of the form $\pi_p(s) \propto a^{-s} p^{-bs}$ for constants $a, b > 0$ ("complexity priors," Castillo and van der Vaart 2012) and binomial priors. The spike-and-slab prior, where we model $\theta_i \overset{\text{iid}}{\sim} r\text{Lap}(\lambda) + (1-r)\delta_0$, falls within this framework by taking π_p to be $\text{Bin}(p, r)$. The value r is the prior inclusion probability of the coordinate i and controls the model selection. Taking a hyperprior $r \sim \text{Beta}(1, p^u)$ for $u > 1$ also satisfies (4) (Castillo and van der Vaart 2012, Example 2.2), allows mixing over the sparsity level r and gives a prior that does not depend on unknown hyperparameters.

The regularization parameter λ in the slab distribution in (3) is allowed to vary with p within the range

$$\frac{||X||}{p} \le \lambda \le 2\bar{\lambda}, \qquad \bar{\lambda} = 2||X||\sqrt{\log p}, \tag{5}$$

where the norm $||X||$ is the maximal column norm defined in (2). The quantity $\bar{\lambda}$ is the usual value of the regularization parameter of the LASSO (Bühlmann and van de Geer 2011, chap. 6). Large values of λ may shrink many coordinates θ_i in the slab toward zero, which is undesirable in our Bayesian setup since we wish to induce sparsity via π_p instead. Indeed, since the slab component identifies the nonzero coordinates, it is unnatural to further shrink these values. It is natural to take fixed values of λ or $\lambda \to 0$, both of which are typically allowed by (5) depending on the specific design matrix and regression setting. Specific values of $||X||$ for some examples of design matrices are given in Section D in the supplementary materials.

The theoretical frequentist behavior of the full posterior arising from prior (3) has been studied in Castillo and van der Vaart (2012) and Castillo, Schmidt-Hieber, and van der Vaart (2015), who obtain oracle contraction rates amongst other things. We build on their work to show that these results extend to the scalable variational approximation.

We briefly comment on the more realistic situation that the model has unknown variance ς^2, in which case we instead observe $Y = X\theta + \varsigma Z$. Since then

$$Y/\varsigma = (X/\varsigma)\theta + Z, \tag{6}$$

one may first rescale the data using an estimate $\hat{\varsigma}$ of ς and as before endow θ with the prior (3), thereby obtaining an empirical Bayes approach. We investigate this empirical Bayes approach numerically in Section 5.2, showing that our method continues to perform well in the more realistic scenario of unknown noise level. One can alternatively use a hierarchical Bayesian approach by endowing ς with a hyperprior, common choices including the inverse Gamma distribution, c/ς^2 or the improper prior $1/\varsigma$.

2.2. Variational Approximations

The posterior $\Pi(\cdot|Y)$ arising from the prior (3) and data (1) assigns weights to all the 2^p possible models, except for very special instances of the design matrix X and prior. Since the posterior is difficult to compute for even moderate p, we take a VB approximation using the mean-field variational family

$$\mathcal{P}_{\text{MF}} = \Big\{ P_{\mu,\sigma,\gamma} = \bigotimes_{i=1}^{p} \big[\gamma_i N(\mu_i, \sigma_i^2) + (1-\gamma_i)\delta_0\big] : \mu_i \in \mathbb{R},\ \sigma_i \in \mathbb{R}^+,\ \gamma_i \in [0,1] \Big\}, \tag{7}$$

with corresponding VB posterior

$$\tilde{\Pi} = \underset{P_{\mu,\sigma,\gamma} \in \mathcal{P}_{\text{MF}}}{\text{argmin}}\ \text{KL}(P_{\mu,\sigma,\gamma} || \Pi(\cdot|Y)), \tag{8}$$

the minimizer of the KL divergence with respect to the posterior. Under $P_{\mu,\sigma,\gamma}$, we have $\theta_i \sim \gamma_i N(\mu_i, \sigma_i^2) + (1-\gamma_i)\delta_0$ independent. We thus approximate the posterior with a spike-and-slab distribution with Gaussian slabs under which every coordinate is independent. Note that while the prior may take the form (7), the posterior will in general not. The key reduction here is that we replace the 2^p model weights with the p VB inclusion probabilities (γ_i), thereby dramatically shrinking the posterior dimension. The VB approximation (8) forces (substantial) additional independence into the resulting distribution, breaking dependencies between the variables. For instance, pairwise information that two coefficients θ_i and θ_j are likely to be selected simultaneously or not at all is lost.

While we use Gaussian slabs in our variational family, it is crucial the *true prior* has slab distributions with at least exponential tails (e.g., Laplace) (Castillo and van der Vaart 2012). The reason a Gaussian approximation works well here is that the likelihood induces Gaussian tails in the posterior. We emphasize

that we use the same variational family to estimate a different posterior compared to previous works (Logsdon, Hoffman, and Mezey 2010; Titsias and Lázaro-Gredilla 2011; Carbonetto and Stephens 2012; Huang, Wang, and Liang 2016; Ormerod, You, and Müller 2017), which use Gaussian *prior* slabs. While using Gaussian prior slabs is particularly efficient computationally, it can yield poor performance due to excessive shrinkage of the estimated coefficients, as we demonstrate numerically in Section A.2 in the supplementary materials. Computing the VB estimate (8) is an optimization problem that can be tackled using CAVI, see Section 4 for details.

While the family $\mathcal{P}_{\mathrm{MF}}$ is our main object of interest, our proofs yield similar theoretical results for two other closely related variational families. Consider the family of distributions consisting of products of a single multivariate normal distribution with a Dirac measure:

$$\begin{aligned}\mathcal{Q} = \{&N_S(\mu_S, \Sigma_S) \otimes \delta_{S^c} : S \subseteq \{1, 2, \ldots, p\}, \mu_S \in \mathbb{R}^{|S|}, \\ &\Sigma_S \in \mathbb{R}^{|S|\times|S|} \text{ a positive definite covariance matrix}\},\end{aligned} \tag{9}$$

where δ_{S^c} denotes the Dirac measure on the coordinates S^c. This family is more rigid on the model selection level than $\mathcal{P}_{\mathrm{MF}}$, selecting a distribution with a single fixed support set S. On this set, however, the family permits a richer representation for the nonzero coefficients, allowing nonzero correlations. Next consider the mean field subclass of $\mathcal{Q}$:

$$\begin{aligned}\mathcal{Q}_{\mathrm{MF}} = \{&N_S(\mu_S, D_S) \otimes \delta_{S^c} : S \subseteq \{1, 2, \ldots, p\}, \mu_S \in \mathbb{R}^{|S|}, \\ &D_S \in \mathbb{R}^{|S|\times|S|} \text{ a positive definite diagonal matrix}\}.\end{aligned} \tag{10}$$

This family again allows distributions with only a single fixed support set S, but further forces independence of the nonzero coefficients. This class is contained in $\mathcal{P}_{\mathrm{MF}}$ by considering distributions $P_{\mu,\sigma,\gamma}$ with inclusion probabilities restricted to $\gamma_i \in \{0, 1\}$. We define the corresponding VB posteriors by

$$\begin{aligned}\hat{Q} &= \operatorname*{argmin}_{Q\in\mathcal{Q}} \mathrm{KL}(Q||\Pi(\cdot|Y)), \\ \tilde{Q} &= \operatorname*{argmin}_{Q\in\mathcal{Q}_{\mathrm{MF}}} \mathrm{KL}(Q||\Pi(\cdot|Y)).\end{aligned} \tag{11}$$

While all our theoretical results also apply to the VB posteriors $\hat{Q}$ and $\tilde{Q}$, these seem to perform worse in practice than $\tilde{\Pi}$, see Section A.2 in the supplementary materials. This is potentially due to the discrete constraint $\gamma_i \in \{0, 1\}$ for these two families, which renders the highly nonconvex optimization problems (11) difficult to solve.

2.3. Design Matrix

The parameter θ in model (1) is not estimable without further conditions on the regression matrix X. For the high-dimensional case $p > n$, which is of most interest to us, θ is not even identifiable without additional assumptions. We thus assume that there is some "true" sparse θ_0 generating the observation (1) with at most s_n nonzero coefficients:

$$\theta_0 \in \{\theta : \#(j : \theta_j \neq 0) \leq s_n\}, \quad \text{for some } s_n = o(n).$$

In the sparse setting, it suffices for estimation to have "local invertibility" of the Gram matrix X^TX. The notion of invertibility can be made more precise using the following definitions, which are based on the sparse high-dimensional literature (e.g., Bühlmann and van de Geer 2011), and have been adapted to the Bayesian setting in Castillo, Schmidt-Hieber, and van der Vaart (2015). We provide only a brief description, referring the interested reader to Section 2.2 of Castillo, Schmidt-Hieber, and van der Vaart (2015) for further discussion.

Definition 1 (Compatibility). A model $S \subseteq \{1, \ldots, p\}$ has *compatibility number*

$$\phi(S) = \inf\left\{\frac{||X\theta||_2|S|^{1/2}}{||X||||\theta_S||_1} : ||\theta_{S^c}||_1 \leq 7||\theta_S||_1, \theta_S \neq 0\right\}.$$

A model is considered "compatible" if $\phi(S) > 0$, in which case $||X\theta||_2|S|^{1/2} \geq \phi(S)||X||||\theta_S||_1$ for all θ in the above set. The number 7 is not important and is taken in Definition 2.1 of Castillo, Schmidt-Hieber, and van der Vaart (2015) to provide a specific numerical value; since we use several results from Castillo, Schmidt-Hieber, and van der Vaart (2015), we employ the same convention. The compatibility number does not directly require sparsity, but reduces the problem to approximate sparsity by considering only vectors θ whose coordinates are small outside S. Conversely, the following two definitions deal only with sparse vectors.

Definition 2 (Uniform compatibility for sparse vectors). The compatibility number for vectors of dimension s is

$$\overline{\phi}(s) = \inf\left\{\frac{||X\theta||_2|S_\theta|^{1/2}}{||X||||\theta||_1} : 0 \neq |S_\theta| \leq s\right\}.$$

Definition 3 (Smallest scaled sparse singular value). The *smallest scaled sparse singular value* of dimension s is

$$\tilde{\phi}(s) := \inf\left\{\frac{||X\theta||_2}{||X||||\theta||_2} : 0 \neq |S_\theta| \leq s\right\}.$$

We shall require that these numbers are bounded away from zero for s a multiple of the true model size. If $||X|| = 1$, then $\tilde{\phi}(s)$ is simply the smallest scaled singular value of a submatrix of X of dimension s. Note that Definitions 1–3 are Definitions 2.1–2.3 of Castillo, Schmidt-Hieber, and van der Vaart (2015). Such compatibility conditions are standard for sparse recovery problems, see Sections 6.13 and 7.15 of Bühlmann and van de Geer (2011) for further discussion.

These compatibility type constants are bounded away from zero for many standard design matrices, such as diagonal matrices, orthogonal designs, iid (including Gaussian) random matrices, and matrices satisfying the "strong irrepresentability condition" of Zhao and Yu (2006). Details of these examples are provided in Section D in the supplementary materials.

3. Main Results

We now provide the main theoretical results of this article concerning the frequentist behavior of the VB posterior $\tilde{\Pi}$ in the asymptotic regime $n, p \to \infty$. While the results are obtained assuming Gaussian noise in model (1), they are in fact robust

to misspecification of the error distribution, see Remark B.1 in Section B in the supplementary materials. This robustness to misspecification is reflected in practice, see Section A.4 in the supplementary materials for numerical results.

Our first result establishes contraction rates for the VB posterior to a sparse truth in ℓ_1-loss, ℓ_2-loss and *prediction error* $||X(\theta-\theta_0)||_2$. Apart from the sparsity level, the rate also depends on compatibility. For $M > 0$, set

$$\begin{aligned}\overline{\psi}_M(S) &= \overline{\phi}\left(\left(2+\frac{4M}{A_4}\left(1+\frac{16}{\phi(S)^2}\frac{\lambda}{\bar{\lambda}}\right)\right)|S|\right),\\ \tilde{\psi}_M(S) &= \tilde{\phi}\left(\left(2+\frac{4M}{A_4}\left(1+\frac{16}{\phi(S)^2}\frac{\lambda}{\bar{\lambda}}\right)\right)|S|\right).\end{aligned} \tag{12}$$

In the natural case $\lambda \ll \bar{\lambda}$, these constants are asymptotically bounded from below by $\overline{\phi}((2+\frac{4M}{A_4})|S|)$ and $\tilde{\phi}((2+\frac{4M}{A_4})|S|)$ if $\phi(S)$ is bounded away from zero. Our results are uniform over vectors in sets of the form

$$\Theta_{\rho_n,s_n} := \{\theta_0 \in \mathbb{R}^p : \phi(S_0) \geq c_0,\ |S_0| \leq s_n,\ \tilde{\psi}_{\rho_n}(S_0) \geq c_0\}, \tag{13}$$

for $S_0 = S_{\theta_0}$, $s_n \geq 1$, $c_0 > 0$ and $\rho_n \to \infty$ (arbitrarily slowly).

Theorem 1 (Recovery). Suppose the model selection prior (3) satisfies (4), (5), and $\lambda = O(||X||\sqrt{\log p}/s_n)$. Then the VB posterior $\tilde{\Pi}$ satisfies, with $S_0 = S_{\theta_0}$,

$$\sup_{\theta_0\in\Theta_{\rho_n,s_n}} E_{\theta_0}\tilde{\Pi}\left(\theta : ||X(\theta-\theta_0)||_2 \geq \frac{M\rho_n^{1/2}}{\overline{\psi}_{\rho_n}(S_0)}\frac{\sqrt{|S_0|\log p}}{\phi(S_0)}\right) \to 0,$$

$$\sup_{\theta_0\in\Theta_{\rho_n,s_n}} E_{\theta_0}\tilde{\Pi}\left(\theta : ||\theta-\theta_0||_1 > \frac{M\rho_n}{\overline{\psi}_{\rho_n}(S_0)^2}\frac{|S_0|\sqrt{\log p}}{||X||\phi(S_0)^2}\right) \to 0,$$

$$\sup_{\theta_0\in\Theta_{\rho_n,s_n}} E_{\theta_0}\tilde{\Pi}\left(\theta : ||\theta-\theta_0||_2 > \frac{M\rho_n^{1/2}}{||X||\tilde{\psi}_{\rho_n}(S_0)^2}\frac{\sqrt{|S_0|\log p}}{\phi(S_0)}\right) \to 0$$

for any $\rho_n \to \infty$ (arbitrarily slowly), Θ_{ρ_n,s_n} defined in (13) and where $M > 0$ depends only on the prior. Moreover, the same holds true for the VB posteriors $\hat{Q}$ and $\tilde{Q}$.

Theorem 1 follows directly from the oracle type Theorem 3 below upon setting $\theta_* = \theta_0$. Recall that we are working under the frequentist model where there is a "true" θ_0 generating data Y of the form (1). Since the above rates equal the minimax estimation rates over $|S_0|$-sparse vectors, Theorem 1 states that the VB posterior puts most of its mass in a neighborhood of optimal size around the truth with high P_{θ_0}-probability in terms of ℓ_1, ℓ_2, and prediction loss. Thus for estimating θ_0, the VB approximation behaves optimally from a theoretical frequentist perspective. This backs up the empirical evidence that VB can provide excellent scalable estimation.

The VB posterior mean often provides a good point estimator and the VB posterior is known to typically underestimate the marginal posterior variance (see, e.g., Blei, Kucukelbir, and McAuliffe 2017—this is a result of using the KL divergence as optimization criterion). The combination of good centering point and the posterior shrinking at least as fast as the true posterior explains why the VB posterior still provides optimal recovery, despite the loss of information from using a mean-field approximation.

Since the prior and variational family do not depend on the unknown sparsity level $|S_0|$ and the VB estimate contracts around the truth at the minimax rate, the procedure is *adaptive*. That is, the procedure can recover an $|S_0|$-sparse truth nearly as well as if we knew the exact level of sparsity of the unknown θ_0. However, the choice of tuning parameters still has an effect on the finite-sample performance, see Section A.3 for a numerical investigation of the effect of the hyperparameter λ. Note that Theorem 1 does not imply that the VB posterior $\tilde{\Pi}$ converges to the true posterior $\Pi(\cdot|Y)$. Indeed, this is neither a typical situation nor a necessary property since the VB estimate should be substantially simpler than the true posterior to be useful.

Theorem 1 implies the variational families $\mathcal{Q}$ and $\mathcal{Q}_{\mathrm{MF}}$ also provide optimal asymptotic estimation of θ_0 in ℓ_1, ℓ_2, and prediction loss. However, the corresponding optimization routine seems to yield worse performance in practice, see Section A.2 in the supplementary materials.

An important motivation for using model selection priors is their ability to perform variable selection. The following result shows that the variational approximation puts most of its mass on models of size at most a multiple of the true dimension, thereby bounding the number of false positives.

Theorem 2 (Dimension). Suppose the model selection prior (3) satisfies (4), (5), and $\lambda = O(||X||\sqrt{\log p}/s_n)$. Then the VB posterior $\tilde{\Pi}$ satisfies, with $S_0 = S_{\theta_0}$,

$$\sup_{\theta_0\in\Theta_{\rho_n,s_n}} E_{\theta_0}\tilde{\Pi}\left(\theta : |S_\theta| \geq |S_0| + M\rho_n\left(1+\tfrac{16}{\phi(S_0)^2}\tfrac{\lambda}{\bar{\lambda}}\right)|S_0|\right) \to 0,$$

for any $\rho_n \to \infty$ (arbitrarily slowly), Θ_{ρ_n,s_n} defined in (13) and where $M > 0$ depends only on the prior. Moreover, the same holds true for the VB posteriors $\hat{Q}$ and $\tilde{Q}$.

Theorem 2 follows directly from the oracle type Theorem 4 below upon setting $\theta_* = \theta_0$. In the interesting case $\lambda \ll \bar{\lambda}$, the factor in Theorem 2 can be simplified to $(1 + M\rho_n)$ if the true parameter is compatible. Note also that under the conditions of Theorems 1 and 2, it is not possible to consistently estimate the true support S_{θ_0} of θ_0 since one cannot separate small and exactly zero signals.

Since the variational families $\mathcal{Q}$ and $\mathcal{Q}_{\mathrm{MF}}$ contain only distributions with a single support set S, the last statement says the resulting VB posteriors will select such a set of size at most a multiple times $|S_0|$ with high P_{θ_0}-probability. The VB estimates based on these two variational families perform model selection in a hard-thresholding manner, reporting only whether a variable is selected or not. On the other hand, the more flexible family $\mathcal{P}_{\mathrm{MF}}$ quantifies the individual variable selection via the reported nontrivial inclusion probabilities $0 \leq \gamma_i \leq 1$, and in this regard provides a richer approximation of the target posterior. Information on pairwise variable inclusion is obviously lost given the mean-field nature of the approximation. Nevertheless, it is interesting to note that all these families still permit good estimation of θ_0.

We now provide more refined *oracle*-type versions of Theorems 1 and 2 as are known to hold for the true posterior (Castillo, Schmidt-Hieber, and van der Vaart 2015).

Theorem 3 (Oracle recovery). Suppose the model selection prior (3) satisfies (4), (5), and $\lambda = O(||X||\sqrt{\log p}/s_n)$. For $\theta_0 \in \mathbb{R}^p \backslash \{0\}$, let $\theta_* \in \mathbb{R}^p$ be any vector satisfying $1 \le s_* = |S_{\theta_*}| \le |S_{\theta_0}| = s_0$ and $||X(\theta_0 - \theta_*)||_2^2 \le (s_0 - s_*)\log p$. Then the VB posterior $\tilde{\Pi}$ satisfies, for any θ_* as above,

$$\sup_{\theta_0 \in \Theta_{\rho_n, s_n}} E_{\theta_0} \tilde{\Pi}\Bigg(\theta : ||X(\theta - \theta_0)||_2 \ge \frac{M\rho_n^{1/2}}{\overline{\psi}_{\rho_n}(S_0)}\left[\frac{\sqrt{s_* \log p}}{\phi(S_*)} + ||X(\theta_0 - \theta_*)||_2\right]\Bigg) \to 0,$$

$$\sup_{\theta_0 \in \Theta_{\rho_n, s_n}} E_{\theta_0} \tilde{\Pi}\Bigg(\theta : ||\theta - \theta_0||_1 > ||\theta_0 - \theta_*||_1 + \frac{M\rho_n}{\overline{\psi}_{\rho_n}(S_0)^2}\left[\frac{s_*\sqrt{\log p}}{||X||\phi(S_*)^2} + \frac{||X(\theta_0 - \theta_*)||_2^2}{||X||\sqrt{\log p}}\right]\Bigg) \to 0,$$

$$\sup_{\theta_0 \in \Theta_{\rho_n, s_n}} E_{\theta_0} \tilde{\Pi}\Bigg(\theta : ||\theta - \theta_0||_2 > \frac{M\rho_n^{1/2}}{||X||\tilde{\psi}_{\rho_n}(S_0)^2} \times \left[\frac{\sqrt{s_* \log p}}{\phi(S_*)} + ||X(\theta_0 - \theta_*)||_2\right]\Bigg) \to 0$$

for any $\rho_n \to \infty$ (arbitrarily slowly), Θ_{ρ_n, s_n} defined in (13) and where $M > 0$ depends only on the prior. Moreover, the same holds true for the VB posteriors $\hat{Q}$ and $\tilde{Q}$.

This can yield better rates than Theorem 1 for certain parameters and choices of θ_*. For example, if $X = I$ is the identity matrix so that $\overline{\psi}_{\rho_n}(S) = \phi(S) = 1$ for all S, setting $\theta_* = 0$ yields

$$\sup_{\theta_0} E_{\theta_0} \tilde{\Pi}\left(\theta : ||\theta - \theta_0||_2 \ge M\rho_n^{1/2}||\theta_0||_2\right) \to 0$$

for any $\rho_n \to \infty$. If $||\theta_0||_2^2 \ll |S_0| \log p$, this improves upon the rate $\sqrt{|S_0| \log p}$ in Theorem 1 by accounting for the size of the coefficients of θ_0 and not only its sparsity level.

The advantage of the oracle bound is it can take into account small nonzero coefficients of θ_0 and capture its "effective sparsity." If $S_* \subset S_0$, as one typically takes, the condition $||X(\theta_0 - \theta_*)||_2^2 = ||X\theta_{0,S_*^c}||_2^2 \le (s_0 - s_*)\log p$ implies that the coordinates of θ_0 in $S_0 \backslash S_*$ contribute on average at most $\log p$ to the squared prediction error. Thus if the coefficient contributes less than $\log p$ to the squared prediction loss, it is preferable to assign it as bias rather than pay the full $\log p$ term required by the squared minimax rate $s_0 \log p$, which accounts only for sparsity irrespective of signal size.

Theorem 4 (Oracle dimension). Suppose the model selection prior (3) satisfies (4), (5), and $\lambda = O(||X||\sqrt{\log p}/s_n)$. For $\theta_0 \in \mathbb{R}^p \backslash \{0\}$, let $\theta_* \in \mathbb{R}^p$ be any vector satisfying $1 \le s_* = |S_{\theta_*}| \le |S_{\theta_0}| = s_0$ and $||X(\theta_0 - \theta_*)||_2^2 \le (s_0 - s_*)\log p$. Then the VB posterior $\tilde{\Pi}$ satisfies, for any θ_* as above,

$$\sup_{\theta_0 \in \Theta_{\rho_n, s_n}} E_{\theta_0} \tilde{\Pi}\Bigg(\theta : |S_\theta| \ge |S_*| + M\rho_n\left[\left(1 + \tfrac{16}{\phi(S_*)^2}\tfrac{\lambda}{\bar{\lambda}}\right)|S_*| + \tfrac{||X(\theta_0 - \theta_*)||_2^2}{\log p}\right]\Bigg) \to 0,$$

for any $\rho_n \to \infty$ (arbitrarily slowly), Θ_{ρ_n, s_n} defined in (13) and where $M > 0$ depends only on the prior. Moreover, the same holds true for the VB posteriors $\hat{Q}$ and $\tilde{Q}$.

Theorems 3 and 4 are special cases of the finite-sample Theorems B.1 and B.2 in the supplementary materials. Our proofs are based on the following crucial result, which allows one to exploit exponential probability bounds for the posterior to control the corresponding probability under the variational approximation.

Theorem 5. Let Θ_n be a subset of the parameter space, A be an event and Q be a distribution for θ. If there exist $C > 0$ and $\delta_n > 0$ such that

$$E_{\theta_0} \Pi(\theta \in \Theta_n | Y) 1_A \le Ce^{-\delta_n}, \tag{14}$$

then

$$E_{\theta_0} Q(\theta \in \Theta_n) 1_A \le \frac{2}{\delta_n}\left[E_{\theta_0} KL(Q||\Pi(\cdot|Y)) 1_A + Ce^{-\delta_n/2}\right].$$

Proof. Recall the duality formula for the KL divergence (Boucheron, Lugosi, and Massart 2013, Corollary 4.15)

$$\mathrm{KL}(Q||P) = \sup_f \left[\int f dQ - \log \int e^f dP\right],$$

where the supremum is taken over all measurable f such that $\int e^f dP < \infty$. In particular,

$$\int f(\theta) dQ(\theta) \le \mathrm{KL}\big(Q \big\| \Pi(\cdot|Y)\big) + \log \int e^{f(\theta)} d\Pi(\theta|Y).$$

Applying this inequality with $f(\theta) = \frac{1}{2}\delta_n 1_{\Theta_n}(\theta)$ and using that $\log(1 + x) \le x$ for $x \ge 0$,

$$\begin{aligned} &\tfrac{1}{2}\delta_n Q(\theta \in \Theta) 1_A \\ &\quad \le \mathrm{KL}(Q||\Pi(\cdot|Y)) 1_A + \log\left(1 + \Pi(\theta \in \Theta_n|Y) e^{\delta_n/2}\right) 1_A \\ &\quad \le \mathrm{KL}(Q||\Pi(\cdot|Y)) 1_A + e^{\delta_n/2} \Pi(\theta \in \Theta_n|Y) 1_A. \end{aligned}$$

Taking E_{θ_0}-expectations on both sides and using (14) gives the result. □

When deriving oracle rates for the original posterior, the exponent $e^{-\delta_n}$ in (14) depends on the oracle quantity, see Section B.3 in the supplementary materials. To apply Theorem 5, we must thus develop novel oracle type bounds on the KL divergence $\mathrm{KL}(\tilde{\Pi}||\Pi(\cdot|Y))$, which is the main technical difficulty in establishing our results, see Section B.2 in the supplementary materials. The proof uses an iterative structure, using successive posterior localizations to eventually bound the KL divergence (see, e.g., Nickl and Ray 2020 for a similar idea).

4. VB Algorithm

4.1. Coordinate Update Equations

We now provide a CAVI algorithm (see, e.g., Blei, Kucukelbir, and McAuliffe 2017) to compute the mean-field VB posterior $\tilde{\Pi}$ based on the spike-and-slab prior with Laplace slabs. Since in the literature (Logsdon, Hoffman, and Mezey 2010; Carbonetto and Stephens 2012; Huang, Wang, and Liang 2016) the VB

approximation is typically considered for Gaussian prior slabs, and can therefore take advantage of explicit analytic formulas, our algorithm requires modification.

Introducing binary latent variables $(z_i)_{i=1}^p$, the spike and slab prior can be rewritten as

$$\begin{aligned} w_i &\overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0), \\ z_i|w_i &\overset{\text{ind}}{\sim} \text{Bernoulli}(w_i), \\ \theta_i|z_i &\overset{\text{ind}}{\sim} z_i\text{Lap}(\lambda) + (1 - z_i)\delta_0. \end{aligned} \tag{15}$$

The prior inclusion probability equals $\Pi(z_i = 1) = \int w d\pi(w) = a_0/(a_0 + b_0)$, the expectation of a beta random variable. In CAVI, we sequentially update the parameters $\gamma_i, \sigma_i, \mu_i$, $i = 1, \ldots, p$, of the VB posterior by minimizing the KL divergence between the variational class with the rest of the parameters kept fixed and the true posterior. We iterate this algorithm until convergence, measured by the change in entropy.

We now give the component-wise variational updates in the algorithm. Fixing the latent variable $z_i = 1$ and all variational factors except μ_i or σ_i (i.e., using vector notation, $\boldsymbol{\mu}_{-i}, \boldsymbol{\sigma}, \boldsymbol{\gamma}$ or $\boldsymbol{\mu}, \boldsymbol{\sigma}_{-i}, \boldsymbol{\gamma}$ are all fixed), the minimizer of the conditional KL divergence between $\mathcal{P}_{\text{MF}}$ and the posterior is the same as the minimizer of

$$\begin{aligned} f_i(\mu_i|\boldsymbol{\sigma}, \boldsymbol{\mu}_{-i}, \boldsymbol{\gamma}, z_i = 1) &= \mu_i \sum_{k \neq i} (X^T X)_{ik} \gamma_k \mu_k + \frac{1}{2}(X^T X)_{ii} \mu_i^2 \\ &\quad - (Y^T X)_i \mu_i + \lambda \sigma_i \sqrt{2/\pi} e^{-\mu_i^2/(2\sigma_i^2)} \\ &\quad + \lambda \mu_i (1 - 2\Phi(-\mu_i/\sigma_i)), \\ g_i(\sigma_i|\boldsymbol{\sigma}_{-i}, \boldsymbol{\mu}, \boldsymbol{\gamma}, z_i = 1) &= \tfrac{1}{2}(X^T X)_{ii} \sigma_i^2 + \lambda \mu_i \sigma_i \sqrt{2/\pi} e^{-\mu_i^2/(2\sigma_i^2)} \\ &\quad + \lambda \mu_i (1 - \Phi(\mu_i/\sigma_i)) - \log \sigma_i, \end{aligned} \tag{16}$$

respectively (see Section C.1 of the supplementary materials for the proof of the above assertion), where Φ denotes the cdf of the standard normal distribution. The minimizers of these functions do not have closed form expressions and hence must be computed by optimization; in our R implementation, we used the built-in optimize() function.

The minimizer γ_i of the conditional KL divergence given $\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}_{-i}$ solves

$$\begin{aligned} \log \frac{\gamma_i}{1 - \gamma_i} &= \log \frac{a_0}{b_0} + \log \frac{\sqrt{\pi} \sigma_i \lambda}{\sqrt{2}} + (Y^T X)_i \mu_i + \frac{1}{2} \\ &\quad - \mu_i \sum_{k \neq i} (X^T X)_{ik} \gamma_k \mu_k - \tfrac{1}{2}(X^T X)_{ii}(\sigma_i^2 + \mu_i^2) \\ &\quad - \lambda \sigma_i \sqrt{2/\pi} e^{-\mu_i^2/(2\sigma_i^2)} - \lambda \mu_i (1 - 2\Phi(-\mu_i/\sigma_i)) \\ &=: \Gamma_i(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}_{-i}), \end{aligned} \tag{17}$$

see Section C.1 of the supplementary materials for the proof.

Following Huang, Wang, and Liang (2016), we terminate the procedure once the coordinate-wise maximal change in binary entropy of the posterior inclusion probabilities falls below a prespecified small threshold ε (e.g., $\varepsilon = 10^{-3}$), that is, stop when $\Delta_H := \max_{i=1,\ldots,p} |H(\gamma_i) - H(\gamma_i')| \leq \varepsilon$, where $H(p) = -p \log p - (1 - p) \log(1 - p)$, $p \in (0, 1)$, and γ_i, γ_i' are the ith coordinate of the starting and updated parameters $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}'$, respectively. The full algorithm is present in Algorithm 1.

4.2. Prioritized Updating Order

The VB objective function is generally nonconvex and so CAVI can be sensitive to initialization (Blei, Kucukelbir, and McAuliffe 2017). It turns out the algorithm is also highly sensitive to the order of the component-wise updates. In fact, naively updating the coordinates in lexicographic order $i = 1, \ldots, p$ is typically suboptimal in our setting. We demonstrate in the next section on various simulated datasets that, unless the significant nonzero coefficients are located at the beginning of the signal, the procedure typically converges to a poor local minimum and gives misleading, inconsistent answers. In particular, CAVI returns a solution that is far from the desired VB posterior it is trying to compute. It is clearly undesirable that the algorithm's performance depends on the arbitrary ordering of the parameter coordinates. A natural fix is to randomize the order of the coordinate-wise updates and use different initializations, choosing the local minimum which provides the smallest overall KL-divergence to the posterior. We show, however, that due to the large number of local minima and their substantially different behavior, this approach can also perform badly (although somewhat better than the lexicographic approach).

We instead propose a novel *prioritized* update scheme. In a first preprocessing step, we compute an initial estimator $\hat{\mu}^{(0)}$ of the mean vector $\boldsymbol{\mu}$ of the variational class. We then place the coefficients in decreasing order with respect to the absolute value of their estimate and update the parameters coordinate-wise in the corresponding order, that is, denoting by $\boldsymbol{a} = (a_1, \ldots, a_p)$ the permutation of the indices $(1, 2, \ldots, p)$ such that $|\hat{\mu}_{a_i}^{(0)}| \geq |\hat{\mu}_{a_j}^{(0)}|$ for every $1 \leq i < j \leq p$, we update the coordinates in the order $\mu_{a_i}, \sigma_{a_i}, \lambda_{a_i}$, $i = 1, \ldots, p$.

The intuition behind this method is that when CAVI begins by updating indices whose signal coefficients are small or zero in the target VB posterior, it may incorrectly assign signal strength to such indices to better fit the data (this is especially the case if the initialization value of the signal coefficient is far from its value in the target VB posterior). Consequently, the estimates of the significant nonzero signal components may be overly small since part of the signal strength has already been falsely assigned to signal coefficients that should in fact be small under the VB posterior. This can trap the algorithm near a poor local minimum from which it cannot escape, see the corresponding simulation study in Section 5.

To avoid this, we wish to first update those coefficients which are large in the target VB posterior. Since these are unknown, the idea here is to identify them using a preliminary estimator: if the target VB posterior does a good job of estimating the signal, these large coefficients should roughly match those that are large in the true underlying signal, which can be identified using a reasonable estimator. The algorithm is given in Algorithm 1, where the function *order*($|\boldsymbol{\mu}|$) returns the indices of $|\boldsymbol{\mu}|$ in descending order.

Algorithm 1 Variational Bayes for Laplace prior slabs

Initialize: $(\Delta_H, \boldsymbol{\sigma}, \boldsymbol{\gamma})$, $\boldsymbol{\mu} := \hat{\mu}^{(0)}$ (for a preliminary estimator $\hat{\mu}^{(0)}$), $\boldsymbol{a} := \text{order}(|\boldsymbol{\mu}|)$
while $\Delta_H \geq \varepsilon$ **do**
  for $j = 1$ to p **do**
    $i := a_j$
    $\mu_i := \text{argmax}_{\mu_i} f_i(\mu_i | \boldsymbol{\mu}_{-i}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, z_i = 1)$ // see (16)
    $\sigma_i := \text{argmax}_{\sigma_i} g_i(\sigma_i | \boldsymbol{\mu}, \boldsymbol{\sigma}_{-i}, \boldsymbol{\gamma}, z_i = 1)$ // see (16)
    $\gamma_{\text{old},i} = \gamma_i$, $\gamma_i = \text{logit}^{-1}\big(\Gamma_i(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\gamma}_{-i})\big)$ // see (17)
  $\Delta_H := \max_i\{|H(\gamma_i) - H(\gamma_{\text{old},i})|\}$

Instead of the prior (15), if one takes deterministic weights w_i, the above algorithm can be easily adapted by using the same update steps for μ_i and σ_i, while updating γ_i as the solution to

$$\log\frac{\gamma_i}{1-\gamma_i} = \log\frac{w_i}{1-w_i} + \log\frac{\sqrt{\pi}\sigma_i\lambda}{\sqrt{2}} + (Y^TX)_i\mu_i + \frac{1}{2} - \sum_{j\neq i}(X^TX)_{ij}\gamma_j\mu_j\mu_i - \frac{\mu_i^2+\sigma_i^2}{2}(X^TX)_{ii} - \lambda\sigma_i\sqrt{2/\pi}\,e^{-\mu_i^2/(2\sigma_i^2)} - \lambda\mu_i(1-2\Phi(-\mu_i/\sigma_i)).$$

The closely related algorithm for computing the VB posterior $\tilde{Q}$ based on the family $\mathcal{Q}_{\text{MF}}$ is given in Algorithm 4 in Section C.2 in the supplementary materials.

5. Numerical Study

In this section, we empirically compare the performance of our VB method using Laplace prior slabs, implemented in the `sparsevb` package (Clara, Szabo, and Ray 2020), with various state-of-the-art Bayesian model selection methods on simulated data. We also demonstrate the importance of the prioritized updating scheme compared with standard CAVI implementations.

Additional numerical results are provided in the supplementary materials as follows:

- Section A.1: we apply our method and other Bayesian model selection methods to real world data.
- Section A.2: we show that Laplace prior slabs provide better estimation and model selection than Gaussian prior slabs. We also show that the optimization problem for finding the KL-optimizer for the class $\mathcal{Q}_{\text{MF}}$ is substantially harder than for the class $\mathcal{P}_{\text{MF}}$, with the former typically ending up at a poor local minimum.
- Section A.3: we show that although the theory indicates that the VB approach is (asymptotically) robust to the choice of the hyperparameter λ, in finite-sample cases it can still have an effect and it may be helpful to use a data-driven choice in practice (e.g., cross-validation).
- Section A.4: we show that several Bayesian model selection methods are robust to noise misspecification
- Section A.5: we compare different Bayesian model selection methods when the inputs are correlated.

We ran each experiment multiple times and report the average ℓ_2-distance between the posterior mean (or maximum a posteriori (MAP) estimate for the SSLASSO) and the true parameter θ_0, the false discovery rate (FDR), the true positive rate (TPR), and the computational time in seconds. We also report the standard deviations of these indicators to quantify their spread. For our computations, we used a MacBook Pro laptop with 2.9 GHz Intel Core i5 processor and 8 GB memory. Throughout the numerical study, we use the hyperparameter choices $a_0 = 1$, $b_0 = p$, $\lambda = 1$ (except in Section A.3 in the supplementary materials) and set the stopping threshold for the entropy change to $\Delta_H = 10^{-5}$, see Algorithm 1. In each experiment and for every method, we take the ridge regression estimator $\hat{\mu}^{(0)} = (X^TX + I)^{-1}X^TY$ as initialization. Given the sparsity framework, it may be tempting to take the LASSO as initialization, however, this is not recommended. The LASSO shrinks some coordinates to exactly zero and so is not suitable for μ, which represents the estimated coefficients *given* that they are included in the model, that is, nonzero [the LASSO solution should be compared to $(\gamma_1\mu_1, \ldots, \gamma_p\mu_p)$ rather than $(\mu_1, \ldots, \mu_p)$].

5.1. Prioritized Updates

We demonstrate here the relevance of our prioritized updating scheme for CAVI by comparing its performance with lexicographic and randomized updating orders, which are standard implementations for CAVI. We take $n = 100$, $p = 200$, $s = 20$, $\theta_i = 10$ for the nonzero coefficients, $\varsigma = 1$ assumed to be known, $X_{ij} \sim^{\text{iid}} N(0,1)$, and consider four scenarios for the locations of the nonzero signal components. We place all nonzero coordinates (i) at the beginning of the signal, (ii) at the end of the signal, (iii) in the middle of the signal, and (iv) uniformly at random. We ran the experiments 200 times and report the results in Table 1 (for the FDR and TPR, the ith coefficient is selected if $\gamma_i > 0.5$). We also plot the posterior means resulting from a typical run in Figure 1.

Apart from the first scenario, where the significant signal coefficients are all located at the beginning of the signal, the prioritized method substantially outperforms both the randomized and lexicographic updating schemes for parameter estimation and model selection (recall that all three methods are trying to compute the *same* VB estimate). The random updating order also slightly improves upon the lexicographic order, except for the first scenario, where the lexicographic order naturally updates the largest coefficients first. As well as being sensitive to initialization (Blei, Kucukelbir, and McAuliffe 2017), it seems CAVI can also be very sensitive to the updating order of the parameters. Indeed, we see here that without prioritized ordering, the algorithm often terminates at poor local minima of the VB objective function. Since the VB objective is nonconvex, naive (or random) update orderings may cause CAVI to return a solution that is far from the true minimizer of the KL divergence that it is trying to compute. Performing updates in a prioritized order can add some robustness against this, see Section 4 for some heuristics behind this idea. We also note that the runtime is comparable for the three updating orders.

Table 1. We compare the prioritized, lexicographic and random updating schemes in the CAVI algorithm. We take $X_{ij} \overset{\text{iid}}{\sim} N(0,1)$, $n = 100$, $p = 200$, $s = 20$, $\theta_i = 10$ for the nonzero coefficients, which are located at the (i) beginning, (ii) middle, (iii) end, (iv) (uniformly) random locations of the signal. We report the means and standard deviations over 200 runs.

Metric	Method	(i)	(ii)	(iii)	(iv)
ℓ_2-error	prioritized	1.03 ± 3.39	$\mathbf{1.18 \pm 3.86}$	$\mathbf{1.06 \pm 3.48}$	$\mathbf{0.61 \pm 1.65}$
	lexicographic	$\mathbf{0.71 \pm 2.14}$	26.61 ± 15.04	45.72 ± 5.45	37.91 ± 5.63
	randomized	27.81 ± 13.30	27.26 ± 13.78	25.14 ± 14.70	35.08 ± 8.28
FDR	prioritized	0.02 ± 0.12	$\mathbf{0.02 \pm 0.13}$	$\mathbf{0.02 \pm 0.12}$	$\mathbf{0.05 \pm 0.18}$
	lexicographic	$\mathbf{0.01 \pm 0.08}$	0.63 ± 0.35	0.87 ± 0.03	0.54 ± 0.38
	randomized	0.68 ± 0.31	0.66 ± 0.32	0.62 ± 0.352	0.69 ± 0.30
TPR	prioritized	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.01}$
	lexicographic	$\mathbf{1.00 \pm 0.00}$	0.93 ± 0.06	0.75 ± 0.11	0.95 ± 0.05
	randomized	0.93 ± 0.07	0.92 ± 0.06	0.93 ± 0.07	0.91 ± 0.07
Runtime (sec)	prioritized	0.28 ± 0.09	0.24 ± 0.06	0.26 ± 0.06	0.24 ± 0.08
	lexicographic	$\mathbf{0.22 \pm 0.06}$	$\mathbf{0.21 \pm 0.05}$	$\mathbf{0.21 \pm 0.04}$	$\mathbf{0.23 \pm 0.06}$
	randomized	0.24 ± 0.08	0.22 ± 0.05	0.23 ± 0.05	0.25 ± 0.06

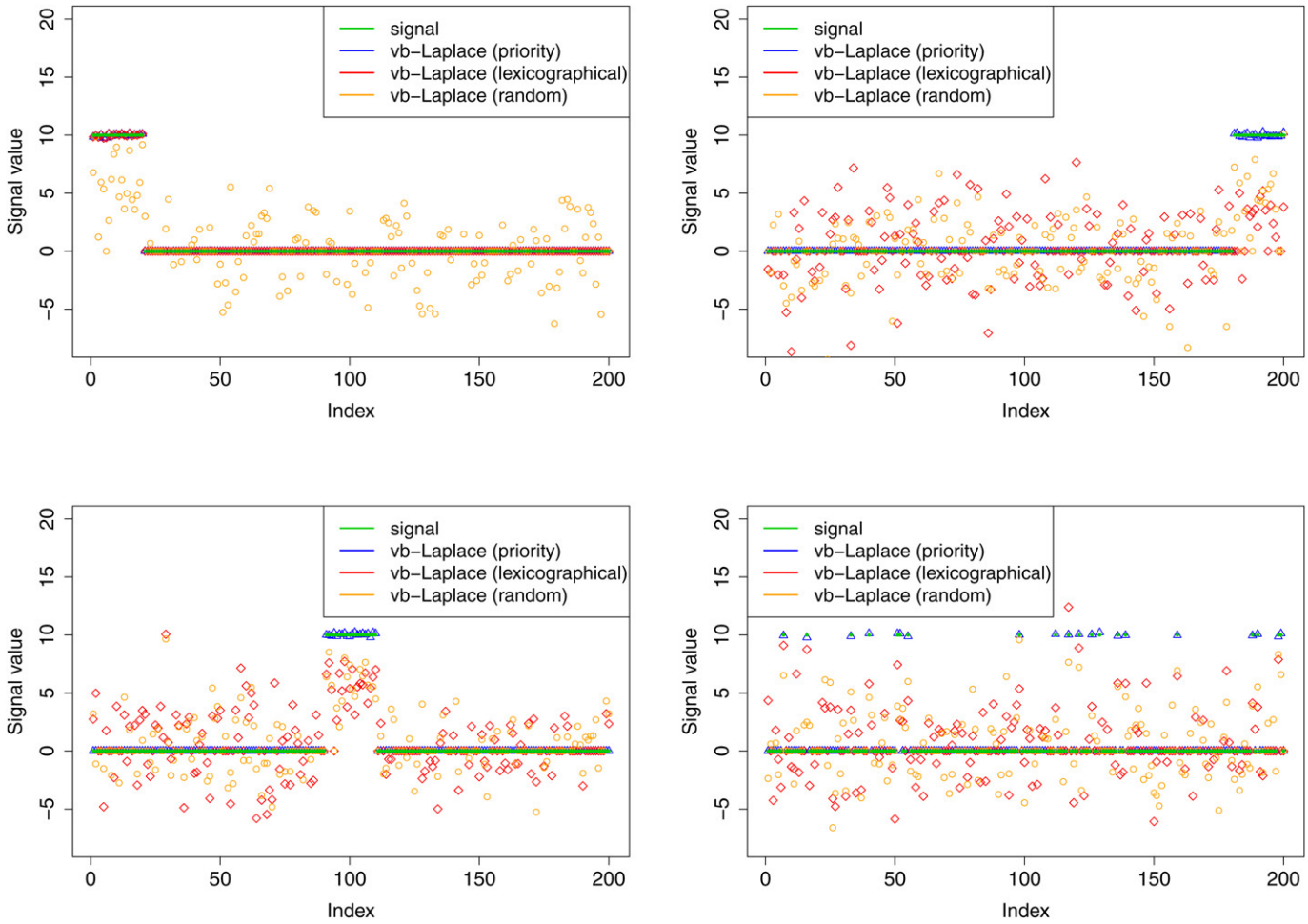


Figure 1. Linear regression with Gaussian design $X_{ij} \overset{\text{iid}}{\sim} N(0,1)$. We plot the underlying signal (green) and posterior means of the VB method with Laplace prior slabs computed using CAVI with parameter updates ordered in a prioritized way (blue), lexicographically (red) and randomly (light blue). We took $n = 100$, $p = 200$, $s = 20$, $\theta_i = 10$. From left to right and top to bottom we have: the nonzero coordinates are at the beginning, end, middle, and at random locations of the signal.

5.2. Comparing Bayesian Variable Selection Methods

We consider here the realistic situation of unknown noise variance ς^2, that is the model $Y = X\theta + \varsigma Z$. As mentioned in Section 2 (see (6)), dividing both sides of this model by an empirical estimator $\hat{\varsigma}$ for the noise standard deviation ς gives $\tilde{Y} = \tilde{X}\theta + \tilde{Z}$, where $\tilde{Y} = Y/\hat{\varsigma}$, $\tilde{X} = X/\hat{\varsigma}$ and $\tilde{Z} = (\varsigma/\hat{\varsigma})Z$, $Z \sim N(0, I_n)$. Endowing θ with the spike-and-slab prior and if the estimator $\hat{\varsigma}$ is close to ς, we should approximately recover the $\varsigma = 1$ case studied above. We thus compute our VB estimator as described above based on the design matrix $\tilde{X}$ and data $\tilde{Y}$. For estimating ς, we have used the R package `selectiveInference`, see (Reid, Tibshirani, and Friedman 2016).

Table 2. Linear regression with Gaussian design $X_{ij} \overset{\text{iid}}{\sim} N(0,1)$, unknown noise variance ς^2 and nonzero signal coefficients $\theta_i = A$, with parameter (n, p, s, A, ς) choices (i) $(100, 400, 20, \log n, 5)$ (nonzero coefficients at the beginning); (ii) $\big(100, 1000, 3, (1, 2, 3), 1\big)$ (at the end); (iii) $(200, 800, 5, \overset{\text{iid}}{\sim} U(-5, 5), 0.2)$ (in the middle); (iv) $(100, 400, 20, 2\log n, 5)$ (at the end).

Metric	Method	(i)	(ii)	(iii)	(iv)
ℓ_2-error	sparsevb	10.48 ± 6.84	0.21 ± 0.14	$\mathbf{0.03 \pm 0.01}$	$\mathbf{6.55 \pm 7.80}$
	varbvs	14.23 ± 6.51	0.18 ± 0.07	$\mathbf{0.03 \pm 0.01}$	20.43 ± 17.15
	EMVS	14.02 ± 2.46	3.57 ± 0.03	5.04 ± 0.33	21.52 ± 11.29
	SSLASSO	20.62 ± 0.17	$\mathbf{0.16 \pm 0.11}$	0.09 ± 0.12	37.92 ± 9.84
	ebreg	$\mathbf{9.38 \pm 6.05}$	0.18 ± 0.07	0.17 ± 0.04	7.39 ± 7.42
FDR	sparsevb	0.12 ± 0.17	0.06 ± 0.16	$\mathbf{0.00 \pm 0.00}$	0.02 ± 0.07
	varbvs	0.06 ± 0.11	0.01 ± 0.04	$\mathbf{0.00 \pm 0.00}$	0.07 ± 0.15
	EMVS	0.24 ± 0.13	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	0.43 ± 0.25
	SSLASSO	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$
	ebreg	0.38 ± 0.20	0.01 ± 0.02	$\mathbf{0.00 \pm 0.00}$	0.28 ± 0.16
TPR	sparsevb	0.70 ± 0.31	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.13	0.94 ± 0.18
	varbvs	0.340 ± 0.37	$\mathbf{1.00 \pm 0.00}$	0.57 ± 0.43	0.53 ± 0.44
	EMVS	0.59 ± 0.14	0.00 ± 0.00	0.86 ± 0.09	0.88 ± 0.10
	SSLASSO	0.01 ± 0.01	0.94 ± 0.13	0.10 ± 0.29	0.09 ± 0.28
	ebreg	$\mathbf{0.88 \pm 0.18}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.98 \pm 0.07}$	$\mathbf{1.00 \pm 0.04}$
Runtime (sec)	sparsevb	0.43 ± 0.27	0.71 ± 0.21	0.35 ± 0.25	0.65 ± 0.53
	varbvs	0.60 ± 0.28	2.02 ± 0.50	0.51 ± 0.38	0.56 ± 0.23
	EMVS	0.20 ± 0.07	1.72 ± 0.44	0.21 ± 0.05	0.19 ± 0.09
	SSLASSO	$\mathbf{0.06 \pm 0.03}$	$\mathbf{0.37 \pm 0.11}$	$\mathbf{0.06 \pm 0.01}$	$\mathbf{0.07 \pm 0.03}$
	ebreg	35.05 ± 7.03	21.20 ± 6.05	31.42 ± 4.01	36.33 ± 9.13

We compare the performance of our VB method with various Bayesian (based) variable selection algorithms for sparse linear regression using simulated data. We consider the `varbvs` R-package (VB for spike-and-slab priors with Gaussian prior slabs using an importance sampling outer circle for estimating the posterior inclusion probabilities and noise variance, Carbonetto and Stephens 2012), `EMVS` R-package (an expectation-maximization algorithm for spike-and-slab, Ročková and George 2014), `SSLASSO` R-package (spike-and-slab LASSO, Ročková and George 2018), and "ebreg.R" R-function (a fractional likelihood empirical Bayes approach using MCMC for recentered Gaussian slab priors, (Martin, Mess, and Walker 2017)—the function is available on the first author's website).

For `varbvs`, we set $tol = 10^{-4}$ and $maxiter = 10^4$. For `EMVS` we took $v_0 \in \{0.1, 0.2, \ldots, 2\}$, $v_1 = 1000$ (these quantities were used in one of the examples provided in the package), $a = 1$, $b = p$, and $\epsilon = 10^{-5}$ and report the posterior mean corresponding to the $v_0 = 0.1$ case. For `SSLASSO`, we took $\lambda_1 = 0.01$, λ_0 an arithmetic series between λ_1 and p with 200 elements, set the variance "unknown," $a = 1$, $b = p$, and penalty="adaptive," and report the results corresponding to the stabilized λ_0 value as recommended by the authors (Ročková and George 2018). In the ebreg algorithm, we took the default parameters $M = 5000$, $\alpha = 0.99$, $\gamma = 0.001$ and used the `selectiveInference` R-package for the estimation of ς. We note that for most of these methods, additional careful hyperparameter tuning beyond the default settings can often lead to improved performance, see Section A.3 in the supplementary materials for our VB method or Section 5 of George and Ročková (2020) for discussion concerning the SSLASSO.

We first consider (i) $n = 100$, $p = 400$, $s = 20$, $\varsigma = 5$ with the nonzero signal coefficients set to $\theta_i = A$, with $A = \log n$, and located at the end of the signal. The entries of the design matrix are taken to be iid normal random variables $X_{ij} \overset{\text{iid}}{\sim} N(0,1)$. In the other experiments, we take (n, p, s, ς) equal to (ii) $(100, 1000, 40, 1)$ (with nonzero coefficients at the beginning of the signal) and set the nonzero parameters to be 1, 2, 3; (iii) $(200, 800, 5, 0.2)$ (in the middle) and take $\theta_i \overset{\text{iid}}{\sim} U(-5, 5)$; (iv) $(100, 400, 20, 5)$ (at the end) and take $\theta_i = 2\log n$. We ran each algorithm 100 times and report the results in Table 2. Our method performs well compared to the other methods, in some cases providing substantially better estimation and model selection.

6. Conclusion

We studied theoretical oracle contraction rates of a natural sparsity-inducing mean-field VB approximation to posteriors arising from widely used, but computationally challenging, model selection priors in high-dimensional sparse linear regression. We showed that under compatibility conditions on the design matrix, such an approximation converges to a sparse truth at an oracle rate in ℓ_1, ℓ_2, and prediction loss, implying optimal (minimax) recovery, and also performs suitable dimension selection. This provides a theoretical justification for this approximation algorithm in a sparsity context. Minimax guarantees for this VB method extend to high-dimensional logistic regression, as we show in the follow up work (Ray, Szabo, and Clara 2020).

We investigated the empirical performance of our algorithm via simulated and real world data and showed that it generally performs at least as well as other state-of-the-art Bayesian variable selection methods, including existing VB approaches. We also demonstrated how the widely used CAVI algorithm can be highly sensitive to the updating order of the parameters. We therefore proposed a novel prioritized updating scheme that uses a data-driven updating order and performs better in simulations. This idea may be applicable for CAVI approaches

in other settings. Our variational algorithm is implemented in the R-package `sparsevb` (Clara, Szabo, and Ray 2020).

Supplementary Materials

Supplementary materials In Section A, additional numerical results are given. First, we provide a real world data example, where we compare Bayesian model selection methods. We then consider various VB methods, demonstrating the advantages of using Laplace instead of Gaussian prior slabs, investigate the effect of the hyperparameter λ and further study Bayesian variable selection methods under noise misspecification and correlated inputs. Section B contains full oracle results and all proofs, Section C contains additional methodological details, and Section D contains further discussion of the design matrix assumption, including examples.

Acknowledgments

We would like to thank two referees for valuable comments that helped considerably improve this manuscript.

Funding

Botond Szabó received funding from the Netherlands Organization for Scientific Research (NWO) under project number 639.031.654.

References

Alquier, P., and Ridgway, J. (2020), "Concentration of Tempered Posteriors and of Their Variational Approximations," *The Annals of Statistics*, 48, 1475–1497. [1271]

Banerjee, S., Castillo, I., and Ghosal, S. (2020), "Survey Paper: Bayesian Inference in High-Dimensional Models." [1270]

Belitser, E., and Ghosal, S. (2020), "Empirical Bayes Oracle Uncertainty Quantification for Regression," *The Annals of Statistics* (to appear). [1271]

Belitser, E., and Nurushev, N. (2020), "Needles and Straw in a Haystack: Robust Confidence for Possibly Sparse Sequences," *Bernoulli*, 26, 191–225. [1271]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [1270,1271,1274,1275,1276,1277]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1270]

Boucheron, S., Lugosi, G., and Massart, P. (2013), *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford: Oxford University Press. [1275]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Series in Statistics, Heidelberg: Springer. [1272,1273]

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, 73–107. [1271,1273,1275,1279]

Castillo, I., and Roquain, E. (2020), "On Spike and Slab Empirical Bayes Multiple Testing," *The Annals of Statistics*, 48, 2548–2574. [1270]

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015), "Bayesian Linear Regression With Sparse Priors," *The Annals of Statistics*, 43, 1986–2018. [1270,1271,1272,1273,1275]

Castillo, I., and Szabó, B. (2020), "Spike and Slab Empirical Bayes Sparse Credible Sets," *Bernoulli*, 26, 127–158. [1270]

Castillo, I., and van der Vaart, A. (2012), "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *The Annals of Statistics*, 40, 2069–2101. [1270,1271,1272]

Chae, M., Lin, L., and Dunson, D. B. (2019), "Bayesian Sparse Linear Regression With Unknown Symmetric Error," *Information and Inference*, 8, 621–653. [1270]

Clara, G., Szabo, B., and Ray, K. (2020), "sparsevb: Spike and Slab Variational Bayes for Linear and Logistic Regression," R Package Version 1.0. [1271,1277,1280]

Efron, B. (2008), "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, 23, 1–22. [1270]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [1270]

George, E. I., and Ročková, V. (2020), "Comment: Regularization via Bayesian Penalty Mixing," *Technometrics*, 62, 438–442. [1279]

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000), "Convergence Rates of Posterior Distributions," *The Annals of Statistics*, 28, 500–531. [1271]

Griffin, J., Latuszynski, K., and Steel, M. (2017), "In Search of Lost (Mixing) Time: Adaptive Markov Chain Monte Carlo Schemes for Bayesian Variable Selection With Very Large *p*," arXiv no. 1708.05678. [1270]

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347. [1270]

Huang, X., Wang, J., and Liang, F. (2016), "A Variational Algorithm for Bayesian Variable Selection," arXiv no. 1602.07640. [1271,1273,1275,1276]

Johnstone, I. M., and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [1270]

Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. (2010), "A Variational Bayes Algorithm for Fast and Accurate Multiple Locus Genome-Wide Association Analysis," *BMC Bioinformatics*, 11, 58. [1271,1273,1275]

Lu, Y., Stuart, A., and Weber, H. (2017), "Gaussian Approximations for Probability Measures on $\mathbb{R}^d$," *SIAM/ASA Journal on Uncertainty Quantification*, 5, 1136–1165. [1271]

Martin, R., Mess, R., and Walker, S. G. (2017), "Empirical Bayes Posterior Concentration in Sparse High-Dimensional Linear Models," *Bernoulli*, 23, 1822–1847. [1271,1279]

Martin, R., and Tang, Y. (2019), "Empirical Priors for Prediction in Sparse High-Dimensional Linear Regression," arXiv no. 1903.00961. [1271]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036. [1270]

Nickl, R., and Ray, K. (2020), "Nonparametric Statistical Inference for Drift Vector Fields of Multi-Dimensional Diffusions," *The Annals of Statistics*, 48, 1383–1408. [1275]

Ormerod, J. T., You, C., and Müller, S. (2017), "A Variational Bayes Approach to Variable Selection," *Electronic Journal of Statistics*, 11, 3549–3594. [1271,1273]

Pati, D., Bhattacharya, A., and Yang, Y. (2018), "On Statistical Optimality of Variational Bayes," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (Vol. 84), pp. 1579–1588. [1271]

Ray, K. (2017), "Adaptive Bernstein–von Mises Theorems in Gaussian White Noise," *The Annals of Statistics*, 45, 2511–2536. [1270]

Ray, K., Szabo, B., and Clara, G. (2020), "Spike and Slab Variational Bayes for High Dimensional Logistic Regression," in *Advances in Neural Information Processing Systems* (Vol. 34). [1271,1279]

Reid, S., Tibshirani, R., and Friedman, J. (2016), "A Study of Error Variance Estimation in Lasso Regression," *Statistica Sinica*, 26, 35–67. [1278]

Ročková, V., and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [1279]

——— (2018), "The Spike-and-Slab LASSO," *Journal of the American Statistical Association*, 113, 431–444. [1279]

Tipping, M. E. (2001), "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, 1, 211–244. [1270]

Titsias, M. K., and Lázaro-Gredilla, M. (2011), "Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning," in *Advances in Neural Information Processing Systems*, pp. 2339–2347. [1271,1273]

van Erven, T., and Szabo, B. (2020), "Fast Exact Bayesian Inference for Sparse Signals in the Normal Sequence Model," *Bayesian Analysis* (to appear). [1270]

Wang, Y., and Blei, D. M. (2019), "Frequentist Consistency of Variational Bayes," *Journal of the American Statistical Association*, 114, 1147–1161. [1271]

West, M. (2003), "Bayesian Factor Regression Models in the 'Large *p*, Small *n*' Paradigm," in *Bayesian Statistics* (Vol. 7, Tenerife, 2002), eds. J. M.

Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, New York: Oxford University Press, pp. 733–742. [1270]

Yang, Y., and Martin, R. (2020), "Variational Approximations of Empirical Bayes Posteriors in High-Dimensional Linear Models," arXiv no. 2007.15930. [1271]

Yang, Y., Pati, D., and Bhattacharya, A. (2020), "α-Variational Inference With Statistical Guarantees," *The Annals of Statistics*, 48, 886–905. [1271]

Zhang, A. Y., and Zhou, H. H. (2020), "Theoretical and Computational Guarantees of Mean Field Variational Inference for Community Detection," *The Annals of Statistics*, 48, 2575–2598. [1271]

Zhang, F., and Gao, C. (2020), "Convergence Rates of Variational Posterior Distributions," *The Annals of Statistics*, 48, 2180–2207. [1271]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [1273]