A FORWARD-BACKWARD APPROXIMATION SCHEME FOR A CLASS OF HIGH-DIMENSIONAL POSTERIOR DISTRIBUTIONS

YVES F. ATCHADÉ

(Feb. 2017, first version May 2015)

Abstract. Exact-sparsity inducing prior distributions in Bayesian analysis typically lead to posterior distributions that are very challenging to handle by standard Markov Chain Monte Carlo (MCMC) methods, particular in high-dimensional models with large number of parameters. We propose an approximation scheme for such posterior distributions based on the forward-backward envelope of Patrinos et al. (2014). We illustrate the method with a high-dimensional linear regression model, where we that the derived approximation is within $O(\sqrt{\gamma})$ of the true posterior distribution in the β-metric, where $\gamma > 0$ is a user-controlled parameter that defines the approximation.

1. Introduction

Successful handling of statistical models with large number of parameters from limited data hinges on the ability to simultaneously solve two problems: (a) weeding out non-significant variables, and (b) estimating the effect of the significant variables. The concept of sparsity has come to play a fundamental role in this endeavor. In the Bayesian framework, sparsity is naturally built in the prior distribution using spike-and-slab priors (Mitchell and Beauchamp (1988); George and McCulloch (1997)), which are mixtures of a point mass at the origin (the spike) and a continuous density (the slab). We will refer to such priors as exact-sparsity inducing priors. A number of recent works have established that these priors, with carefully chosen slab densities, produce posterior distributions with optimal posterior contraction rates (Castillo et al. (2015); Atchade (2017)). However, the flip side of such stellar statistical properties is the fact that these posterior distributions are computationally difficult to handle, particularly in high-dimensional applications. Deriving tractable

2000 *Mathematics Subject Classification.* 60F15, 60G42.

Key words and phrases. Moreau-Yosida approximation, Markov Chain Monte Carlo, Spike-and-slab priors, Variable selection, High-dimensional models.

This work is partially supported by the NSF grant DMS 1228164 and 1513040.

Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* yvesa@umich.edu.

and scalable approximations for such distributions is therefore a problem of practical importance.

In practice, exact-sparsity prior distributions are commonly used with Gaussian linear regression models and Gaussian slab densities by taking advantage of conjugacy (George and McCulloch (1997); Bottolo and Richardson (2010); Yang et al. (2016)). For general non-Gaussian models or non-Gaussian slab densities, several specialized MCMC methods have been developed with mixed results. Reversible jump MCMC algorithms (Chen et al. (2011); Ge et al. (2011)) work well in low-dimensional problems, but tend to mix poorly for high-dimensional problems as shown numerically by Schreck et al. (2013). An alternative to reversible jump is the Metropoli-Hastings framework of Gottardo and Raftery (2008). However, whether their algorithms can successfully deal with high-dimensional problems remains to be explored. Another recent development is the shrinkage-thresholding Metropolis adjusted Langevin algorithm (STMaLa) of Schreck et al. (2013) – which can be seen as a special case of the framework of Gottardo and Raftery (2008) – which we show below suffers as well from poor mixing in large scale problems. Note also that the Laplace approximation, one of the most standard approximation tool in Bayesian computation, cannot be straightforwardly applied when the dimension of the space is as big as the sample size (Shun and McCullagh (1995)). Variational Bayes approximations recently explored by Ormerod et al. (2014) are promising alternatives, but remained to be fully explored in high-dimensional settings.

1.1. Main contribution.

We consider high-dimensional variable selection problems with exact-sparsity inducing prior distributions, and derive an approximation of the posterior distribution based on the forward-backward envelope of Patrinos et al. (2014). The method works as long as the log-likelihood function is concave and smooth (differentiable with Lipschitz derivative). The forward-backward envelope is closely related to the Moreau envelope, a well-established regularization method in optimization (Moreau (1965); Bauschke and Combettes (2011); Parikh and Boyd (2013)). For several important examples of non-Gaussian slab densities, the resulting approximation is easily explored by standard Markov Chain Monte Carlo (MCMC) algorithms. An important advantage of our approach is that approximation errors are easy to control mathematically, and can be used to carry out a detailed analysis of the method, as we did below (see also the recent follow-up work Atchade and Bhattacharyya (2018)). Several recent works have recognized the usefulness of the Moreau regularization for Bayesian computation. Pereyra (2013) noted that a log-concave density can be well approximated by its Moreau-Yosida approximation. However, the framework developed by Pereyra (2013) does not handle the class of posterior distributions considered here. Another related work is the STMaLa of Schreck et al. (2013) mentioned above, which implicitly uses the Moreau approximation to design Metropolis-Hastings proposals.

If $\check{\Pi}(\cdot|z)$ denotes the posterior distribution of interest on $\mathbb{R}^d \times \{0,1\}^d$ given data z, we write $\check{\Pi}_\gamma(\cdot|z)$ to denote the proposed forward-backward approximation, where $\gamma > 0$ is a user-controlled parameter that defines the quality of the approximation. We derive in Theorem 7 – under assumption H1 – an upper bound on the β-metric (see Section 1.2 for precise definition) between $\check{\Pi}(\cdot|z)$ and $\check{\Pi}_\gamma(\cdot|z)$. The main interest of the approximation is that it is much easier to sample from $\check{\Pi}_\gamma$ compared to $\check{\Pi}$. In a recent work (Atchade and Bhattacharyya (2018)) we provide an even stronger justification for the proposed method by showing that $\check{\Pi}_\gamma$ (viewed as a pseudo-posterior distribution) contracts towards $(\delta_\star, \theta_\star)$ at the same rate as the true posterior distribution $\check{\Pi}$, as $n, p \to \infty$.

We illustrate the method using a linear regression model with a spike-and-slab prior, where the slab is the elastic net density (Li and Lin (2010)) – which includes the double exponential (Laplace) distribution as a special case. In this linear regression example, our proposed methodology produces an approximation $\check{\Pi}_\gamma$ of this posterior distribution, and we develop an efficient Markov Chain Monte Carlo algorithm to sample from $\check{\Pi}_\gamma$. We show that the approximation $\check{\Pi}_\gamma(\cdot|z)$ is always a well-defined probability measure provided that γ is chosen as in (22). Furthermore, we show in Corollary 8 that the β-metric between $\check{\Pi}(\cdot|z)$ and $\check{\Pi}_\gamma(\cdot|z)$ satisfies

$$\mathsf{d}_\beta\left(\check{\Pi}(\cdot|z), \check{\Pi}_\gamma(\cdot|z)\right) = O(\sqrt{\gamma}). \tag{1}$$

We illustrate these results in a simulation study which shows that the method performs well, and outperforms STMaLa for high-dimensional problems. A Matlab implementation can be obtained from
`http://dept.stat.lsa.umich.edu/∼ yvesa/Research.html`.

The remainder of the paper is organized as follows. We close the introduction with some notation that will be used throughout the paper. In Section 2, we first introduce the class of posterior distributions of interest, followed in Section 3 by the basic idea of the forward-backward approximation. In Section 4, we develop how the idea can be applied to approximate the posterior distributions of interest. Section 5 details an application to linear regression models. We close the paper with further discussion in Section 6. All the proofs are gathered in Section 7.

1.2. **Notation.** Throughout the paper, $d \geq 1$ is a given integer and $\mathbb{R}^d$ denotes the d-dimensional Euclidean space equipped with its Borel sigma-algebra, its Euclidean norm $\|\cdot\|$, and inner product $\langle \cdot, \cdot \rangle$. We also use the norms $\|\theta\|_1 \overset{\text{def}}{=} \sum_{j=1}^d |\theta_j|$, and

$\|\theta\|_0$ defined as the number of non-zero components of θ. The Lebesgue measure on $\mathbb{R}^d$ is written as $\mathrm{d}x$ when there is no confusion.

We set $\Delta \overset{\text{def}}{=} \{0,1\}^d$. For $\delta \in \Delta$, μ_δ denote the product measure on $\mathbb{R}^d$ defined as $\mu_\delta(\mathrm{d}\theta) \overset{\text{def}}{=} \prod_{j=1}^d \nu_{\delta_j}(\mathrm{d}\theta_j)$, where $\nu_0(\mathrm{d}z)$ is the Dirac mass at 0, and $\nu_1(\mathrm{d}z)$ is the Lebesgue measure on $\mathbb{R}$. Hence integration with respect to μ_δ sets to zero all the components for which $\delta_j = 0$, and integrates the remaining components using the standard Lebesgue measure.

For $\theta, \vartheta \in \mathbb{R}^d$, $\theta \cdot \vartheta$ denotes the component-wise product: $(\theta \cdot \vartheta)_j = \theta_j \vartheta_j$, $1 \leq j \leq d$. For $\delta \in \Delta$, we shall write θ_δ to denote $\theta \cdot \delta$, and we set

$$\mathbb{R}^d_\delta \overset{\text{def}}{=} \{\theta_\delta,\ \theta \in \mathbb{R}^d\} = \{\theta \in \mathbb{R}^d:\ \theta_j = 0 \text{ for } \delta_j = 0,\ j = 1, \ldots, d\}.$$

We will need ways to evaluate the distance between two probability measures. Let $(\mathsf{X}, \mathsf{d}_{\mathsf{X}})$ be some arbitrary separable complete metric space equipped with its Borel sigma-algebra. For any two probability measures μ_1, μ_2 on X, the β-distance between μ_1, μ_2 is defined as

$$\mathsf{d}_\beta(\mu_1, \mu_2) \overset{\text{def}}{=} \sup_{\|f\|_{\mathsf{BL}} \leq 1} \left| \int_{\mathsf{X}} f(x)\mu_1(\mathrm{d}x) - \int_{\mathsf{X}} f(x)\mu_2(\mathrm{d}x) \right|, \tag{2}$$

where the supremum is taken over all measurable functions $f:\ \mathsf{X} \to \mathbb{R}$ such that $\|f\|_{\mathsf{BL}} \overset{\text{def}}{=} \|f\|_\infty + \|f\|_{\mathsf{L}} \leq 1$, where

$$\|f\|_\infty \overset{\text{def}}{=} \sup_{x \in \mathsf{X}} |f(x)|, \quad \text{and} \quad \|f\|_{\mathsf{L}} \overset{\text{def}}{=} \sup \left\{ \frac{|f(x_1) - f(x_2)|}{\mathsf{d}_{\mathsf{X}}(x_1, x_2)},\ x_1, x_2 \in \mathsf{X},\ x_1 \neq x_2 \right\}.$$

It is well-known that this metric metricizes weak convergence (see e.g. Dudley (2002) Theorem 11.3.3). If the supremum in (2) is replaced by the supremum over all measurable functions $f:\ \mathsf{X} \to \mathbb{R}$ such that $\|f\|_\infty \leq 1$ (resp. $\|f\|_{\mathsf{L}} \leq 1$) one obtains the total variation metric d_{tv} (resp. the Wasserstein metric d_{w} with respect to the metric d_{X}).

On a referee's suggestion, we list here all the variable selection posterior distributions (and approximations thereof) that appear in the paper for easy reference.

Name	Distribution
Exact spike-and-slab	$\check{\Pi}(\delta, \mathrm{d}\theta \mid z) \propto \pi_\delta e^{-h(\theta \mid \delta)} \mu_\delta(\mathrm{d}\theta)$
FB approximation	$\check{\Pi}_\gamma(\delta, \mathrm{d}\theta \mid z) \propto \pi_\delta (2\pi\gamma)^{\frac{\Vert\delta\Vert_0}{2}} e^{-h_\gamma(\theta \mid \delta)} \mathrm{d}\theta$
Weak spike-and-slab-1	$\bar{\Pi}_\gamma(\delta, \mathrm{d}\theta \mid z) \propto \pi_\delta (2\pi\gamma)^{\frac{\Vert\delta\Vert_0}{2}} \left\{ \prod_{j:\ \delta_j = 1} p(\theta_j) \right\} e^{-\frac{1}{2\gamma}\Vert\theta - \theta_\delta\Vert_2^2} e^{-\ell(\theta)} \mathrm{d}\theta$
Weak spike-and-slab-2	$\tilde{\Pi}_\gamma(\delta, \mathrm{d}\theta \mid z) \propto \pi_\delta (2\pi\gamma)^{\frac{\Vert\delta\Vert_0}{2}} \left\{ \prod_{j:\ \delta_j = 1} p(\theta_j) \right\} e^{-\frac{1}{2\gamma}\Vert\theta - \theta_\delta\Vert_2^2} e^{-\ell(\theta_\delta)} \mathrm{d}\theta$

2. High-dimensional posterior distributions with sparse priors

Let z be a realization of some random variable Z with conditional distribution f_θ, given a parameter $\theta \in \mathbb{R}^d$. With a prior distribution Π on θ, the posterior distribution for learning θ is

$$\check{\Pi}(\mathrm{d}\theta|z) = \frac{f_\theta(z)\Pi(\mathrm{d}\theta)}{\int_{\mathbb{R}^d} f_\theta(z)\Pi(\mathrm{d}\theta)}.$$

We consider a prior distribution Π on $\Delta \times \mathbb{R}^d$ of the form

$$\Pi(\delta, \mathrm{d}\theta) = \pi_\delta \Pi(\mathrm{d}\theta|\delta),$$

for a discrete distribution $\{\pi_\delta, \delta \in \Delta\}$ on Δ, and a prior $\Pi(\cdot|\delta)$ that is built as follows. Given δ, the components of θ are independent, and for $1 \leq j \leq d$,

$$\theta_j|\delta \sim \begin{cases} \mathsf{Dirac}(0) & \text{if } \delta_j = 0 \\ p(\cdot) & \text{if } \delta_j = 1 \end{cases}, \tag{3}$$

where $\mathsf{Dirac}(0)$ is the Dirac measure on $\mathbb{R}$ with full mass at 0, and $p(\cdot)$ is a positive density on $\mathbb{R}$. By the standard data-augmentation trick, we will take the variable δ as part of the posterior distribution. As defined, the support of $\Pi(\cdot|\delta)$ is $\mathbb{R}^d_\delta = \{\theta \in \mathbb{R}^d :\ \theta_j = 0 \text{ for } \delta_j = 0,\ 1 \leq j \leq d\}$, and $\Pi(\cdot|\delta)$ has a density with respect to the measure μ_δ defined in Section 1.2:

$$\Pi(\mathrm{d}\theta|\delta) = e^{-P(\theta|\delta)}\mu_\delta(\mathrm{d}\theta), \quad \text{where}$$

$$P(\theta|\delta) \overset{\text{def}}{=} \begin{cases} -\sum_{j:\ \delta_j=1} \log p(\theta_j) & \text{if } \theta \in \mathbb{R}^d_\delta \\ +\infty & \text{otherwise}. \end{cases}$$

In the above formula, and throughout the paper, we convene that $e^{-\infty} = 0$, and $0 \times \infty = 0$. We also define

$$\ell(\theta) \overset{\text{def}}{=} -\log f_\theta(z), \quad \text{and} \quad h(\theta|\delta) \overset{\text{def}}{=} \ell(\theta) + P(\theta|\delta),\ \theta \in \mathbb{R}^d,$$

so that the posterior distribution writes

$$\check{\Pi}(\delta, \mathrm{d}\theta|z) \propto \pi_\delta e^{-h(\theta|\delta)}\mu_\delta(\mathrm{d}\theta). \tag{4}$$

Monte Carlo simulation from this posterior distribution can be challenging. The issue is related to the discrete-continuous mixture form of the spike-and-slab prior on θ, which has the effect that any two distributions $\check{\Pi}(\delta, \cdot|z)$ and $\check{\Pi}(\delta', \cdot|z)$ are mutually singular for $\delta \neq \delta'$. As a result, if direct sampling from the conditional distribution of $\theta|\delta, z$ is not possible[1], then sampling from (4) requires the use of specialized MCMC methods (Green (1995); Gottardo and Raftery (2008); Chen et al. (2011); Schreck

[1] which is typically the case if the model or the slab prior is not Gaussian

et al. (2013)). However these algorithms are typically difficult to design and tune, particularly in high-dimensional settings.

3. The Moreau and the forward-backward envelopes

Our goal in this work is to develop a more tractable approximation to the posterior distribution $\check{\Pi}$ in (4). However to make the ideas easy to follow, we start with some general discussion of the Moreau and the forward-backward envelopes. Let $h : \mathbb{R}^d \to (-\infty, +\infty]$ be a convex, lower semi-continuous function that is not identically $+\infty$, and let μ be a sigma-finite measure on $\mathbb{R}^d$. In the applications, μ will naturally be taken as the Lebesgue measure on the domain of h (the domain of h is the set of points $x \in \mathbb{R}^d$ such that $h(x) < \infty$). Assuming that $Z \overset{\text{def}}{=} \int_{\mathbb{R}^d} e^{-h(x)}\mu(\mathrm{d}x) < \infty$, we consider the probability measure

$$\nu(\mathrm{d}x) = \frac{1}{Z} e^{-h(x)}\mu(\mathrm{d}x). \tag{5}$$

To fix the ideas, the reader may think of the case where h is finite everywhere and μ is the Lebesgue measure on $\mathbb{R}^d$. In that case ν is the probability distribution on $\mathbb{R}^d$ with density $(1/Z)e^{-h(x)}$. However our main interest is in the posterior distribution (4) for which the slightly more general setting is needed.

Suppose that we are interested in drawing samples from ν. The lack of smoothness of h, and the possibly complicated geometry of the support of ν can create difficulties for standard MCMC algorithms. An approximation of ν can be formed using the Moreau envelope of h defined for $\gamma > 0$ as

$$\tilde{h}_\gamma(x) = \min_{u\in\mathbb{R}^d}\left[h(u) + \frac{1}{2\gamma}\|u - x\|^2\right],\ \ x \in \mathbb{R}^d.$$

Under the assumptions imposed on h above, the function $\tilde{h}_\gamma$ is known to be well-defined and finite everywhere. It is also convex, continuously differentiable with a Lipschitz gradient, and $\tilde{h}_\gamma(x) \uparrow h(x)$, as $\gamma \downarrow 0$, for all $x \in \mathbb{R}^d$. All these properties are well-known and can be found in Bauschke and Combettes (2011) (Chapter 12). Assuming that $\tilde{Z}_\gamma \overset{\text{def}}{=} \int_{\mathbb{R}^d} e^{-\tilde{h}_\gamma(x)}\mathrm{d}x < \infty$, it seems natural to consider the probability measure

$$\tilde{\nu}_\gamma(\mathrm{d}x) = \frac{1}{\tilde{Z}_\gamma} e^{-\tilde{h}_\gamma(x)}\mathrm{d}x,$$

as an approximation of ν. To the best of our knowledge the idea of approximating the probability measure ν by $\tilde{\nu}_\gamma$ was first considered by Pereyra (2013), in the case where the function h is finite everywhere and μ is the Lebesgue measure on $\mathbb{R}^d$. We refer the reader to that paper for a good discussion of the basic properties of $\tilde{\nu}_\gamma$, and how well it approximates ν. In particular Pereyra (2013) showed that the smoothness of $\tilde{h}_\gamma$ can be exploited to derive efficient gradient-based MCMC samplers for ν. An important limitation of the Moreau envelop approximation is that it is typically not available in closed form, and its computation leads to a d-dimensional, possibly complicated optimization problem.

In many problems the function h takes the particular form

$$h(x) = \ell(x) + P(x), \quad x \in \mathbb{R}^d$$

where ℓ is convex, finite everywhere and twice continuously differentiable, and P is convex, not identically $+\infty$ and lower semi-continuous. In such cases, one can approximate ℓ around a given point x by its linear approximation $u \mapsto \ell(x) + \langle \nabla \ell(x), u - x \rangle$, where $\nabla \ell(x)$ denote the gradient of ℓ at x. This approximation leads to the so-called forward-backward envelope of h, defined for $\gamma > 0$ as

$$\begin{aligned} h_\gamma(x) &\stackrel{\text{def}}{=} \min_{u \in \mathbb{R}^d} \left[\ell(x) + \langle \nabla \ell(x), u - x \rangle + P(u) + \frac{1}{2\gamma} \|u - x\|^2 \right], \quad x \in \mathbb{R}^d \\ &= \ell(x) + -\frac{\gamma}{2} \|\nabla \ell(x)\|^2 + \min_{u \in \mathbb{R}^d} \left[P(u) + \frac{1}{2\gamma} \|u - x + \gamma \nabla \ell(x)\|^2 \right]. \quad (6) \end{aligned}$$

Under the assumptions imposed on ℓ and P above, the function h_γ is finite everywhere, continuously differentiable, and $h_\gamma \leq h$. These properties can be found in Patrinos et al. (2014) Theorem 2.2, but are easy to derive. For instance, the differentiability follows from the expression (6), the twice differentiability of ℓ, and the differentiability of the Moreau envelop approximation of P. Notice however that h_γ is no longer convex in general. Assuming that $Z_\gamma \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} e^{-h_\gamma(x)} \mathrm{d}x < \infty$ it seems also natural to consider the resulting approximation of ν defined as

$$\nu_\gamma(\mathrm{d}x) = \frac{1}{Z_\gamma} e^{-h_\gamma(x)} \mathrm{d}x.$$

The main advantage of h_γ over $\tilde{h}_\gamma$ is that in many problems of interest h_γ is available in closed form, whereas $\tilde{h}_\gamma$ is not. Furthermore, if the function P is separable, then the computation of h_γ leads to d separate one-dimensional optimization problems the solution of which can be easily parallelized. However, the price to pay for the computational convenience is that h_γ may not be convex, and it is a less accurate approximation of h. Indeed, by the convexity of ℓ, we have $\ell(u) \geq \ell(x) + \langle \nabla \ell(x), u - x \rangle$ for all $u \in \mathbb{R}^d$. Hence $h_\gamma(x) \leq \tilde{h}_\gamma(x) \leq h(x)$ for all $x \in \mathbb{R}^d$. But as we will see, the pointwise convergence $h_\gamma \uparrow h$, as $\gamma \downarrow 0$ still holds. Figure 1 gives an illustrative example of the differences between h_γ and $\tilde{h}_\gamma$ and how both functions approximate h. For this example, h_γ is available explicitly (using (6) and soft-thresholding), and $\tilde{h}_\gamma$ is obtained by numerical optimization for each value of x.

Since h_γ converges pointwise to h as $\gamma \downarrow 0$, it seems natural to expect that ν_γ approaches ν for small γ. If the function h is finite everywhere, one can easily show

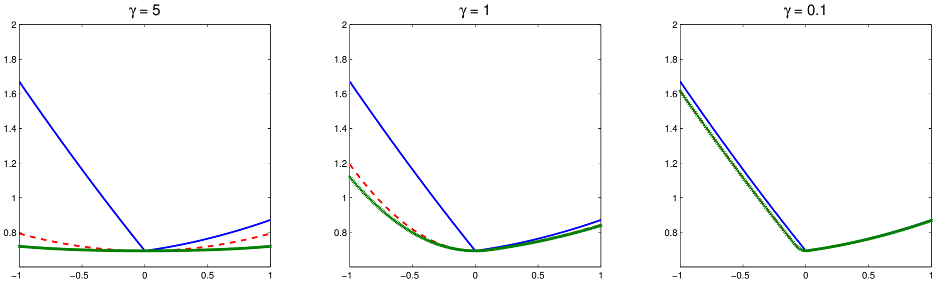

FIGURE 1. Figure showing the function $h(x) = -ax + \log(1 + e^{ax}) + b|x|$ for $a = 0.8$, $b = 0.5$ (blue/solid line), and the approximations h_γ and $\tilde{h}_\gamma$ ($h_\gamma \leq \tilde{h}_\gamma$), for $\gamma \in \{5, 1, 0.1\}$. For $\gamma = 0.1$, the curves of h_γ and $\tilde{h}_\gamma$ are almost undistinguishable on the figure.

(see Proposition 1 below) that indeed, ν_γ converges to ν in the total variation metric, as $\gamma \downarrow 0$. However this result is no longer true when the domain of h has zero $\mathbb{R}^d$-Lebesgue measure. In this latter case, we will show that the convergence of ν_γ occurs only weakly, or in the Wasserstein metric.

Proposition 1. *Suppose μ in (5) is the Lebesgue measure on $\mathbb{R}^d$, $h = \ell + P$ is convex, finite everywhere, and $h_\gamma(x) \uparrow h(x)$ for all $x \in \mathbb{R}^d$. Suppose also that there exists $\gamma_0 > 0$ such that $Z_{\gamma_0} < \infty$. Then for all $\gamma \in (0, \gamma_0]$, ν_γ is well-defined, and*

$$\mathsf{d}_{\mathrm{tv}}(\nu_\gamma, \nu) \leq 2\left(1 - \frac{Z}{Z_\gamma}\right) \downarrow 0, \quad \textit{as} \quad \gamma \downarrow 0.$$

Proof. See Section 7.1. □

Remark 2. (1) Notice that Proposition 1 can also be applied to $\tilde{\nu}_\gamma$ by taking $\ell \equiv 0$.
(2) We show in Lemma 2 in the Appendix that if ℓ is finite everywhere and differentiable, and P is finite everywhere, convex with a nonempty subdifferential at x for all $x \in \mathbb{R}^d$, then $h_\gamma \uparrow h$, as required in the proposition.

If the domain of h has $\mathbb{R}^d$-Lebesgue measure 0, then ν and ν_γ are then automatically mutually singular and Proposition 1 cannot hold. The following toy example illustrates this case.

Example 3. Suppose that we take $\mathbb{R}^d = \mathbb{R}$, $\ell \equiv 0$, and we take $P(x) = 0$ if $x = 0$, and $P(x) = +\infty$ if $x \neq 0$. In that case $e^{-h(x)} = 1$ if $x = 0$, and $e^{-h(x)} = 0$ if $x \neq 0$. Let $\mu = \delta_0$ be the point mass probability measure at 0. Hence $\nu = \delta_0$. For $\gamma > 0$,

$h_\gamma(x) = \tilde{h}_\gamma(x) = x^2/(2\gamma)$, $x \in \mathbb{R}$. Hence ν_γ is the normal distribution $\mathbf{N}(0,\gamma)$. It follows that $\mathsf{d}_{\mathrm{tv}}(\nu_\gamma,\nu) = 2$, for all $\gamma > 0$. But for any Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant 1,

$$|\nu_\gamma(f) - \nu(f)| = |\nu_\gamma(f) - f(0)| \leq \mathbb{E}(|Z_\gamma|) = \sqrt{\frac{2\gamma}{\pi}},$$

where $Z_\gamma \sim \mathbf{N}(0,\gamma)$. By taking $f = |\cdot|$, it can be easily seen that $\mathsf{d}_{\mathsf{w}}(\nu_\gamma,\nu) = \sqrt{\frac{2\gamma}{\pi}}$. Hence ν_γ converges in the Wasserstein metric to ν, but not in total variation. And the convergence rate is $O(\sqrt{\gamma})$.

Remark 4. The fact that we only have convergence in the Wasserstein metric implies that one needs to be cautious about the fact that not all probabilities $\nu(A)$ are well approximated by $\nu_\gamma(A)$. For instance, in Example 3, if $A = [0,a)$ for some $a > 0$, then $\nu(A) = 1$, whereas $\lim_{\gamma\downarrow 0} \nu_\gamma(A) = 0$.

In the next section we will use the approximating measure ν_γ introduced above to approximate the posterior distribution (4). We will see that the situation is similar to the one in Example 3, and as in that example we will show that the approximation converges weakly to the posterior distribution $\check{\Pi}$, and the convergence rate of of order $O(\sqrt{\gamma})$.

4. The forward-backward approximation of the posterior distribution (4)

In this section, we return to the posterior distribution (4) defined in Section 2. And we make the following assumptions on the functions ℓ and P.

H1. *(1) The function $\theta \mapsto \ell(\theta)$ is finite everywhere, convex, and twice continuously differentiable.*

(2) For all $\delta \in \Delta$, the function $\theta \mapsto P(\theta|\delta)$ is convex, lower semi-continuous, not identically $+\infty$, and admits a sub-gradient $g(\theta|\delta)$ at θ, for all $\theta \in \mathbb{R}^d_\delta$.

Remark 5. (1) The convexity assumption on ℓ is fundamental and delineates the type of problems to which the proposed approximation could be easily applied. Extension beyond this set up is possible, but will require fundamentally different techniques.

(2) The convexity of $P(\cdot|\delta)$ boils down to the log-concavity of the density p in the prior (3). Most of the sparsity promoting prior densities used in practice are log-concave.

Given $\delta \in \Delta$, we consider the forward-backward envelope of $h(\cdot|\delta)$ defined as

$$h_\gamma(\theta|\delta) \stackrel{\text{def}}{=} \min_{u\in\mathbb{R}^d} \left[\ell(\theta) + \langle \nabla\ell(\theta), u-\theta\rangle + P(u|\delta) + \frac{1}{2\gamma}\|u-\theta\|^2\right],\ \theta \in \mathbb{R}^d, \tag{7}$$

for some parameter $\gamma > 0$. Using h_γ, we propose to approximate the posterior distribution $\check{\Pi}$ in (4) by

$$\check{\Pi}_\gamma(\delta, \mathrm{d}\theta|z) \propto \pi_\delta \,(2\pi\gamma)^{\frac{\|\delta\|_0}{2}}\, e^{-h_\gamma(\theta|\delta)}\mathrm{d}\theta, \tag{8}$$

that we call the forward-backward approximation of $\check{\Pi}$. In the expression (8), π denotes the irrational number. The function $h_\gamma(\cdot|\delta)$ is available in closed form whenever the Moreau envelope of $P(\cdot|\delta)$ has a closed form expression. More specifically, for $\delta \in \Delta$, and for $\gamma > 0$, we define the Moreau envelope of P as

$$P_\gamma(\theta|\delta) \stackrel{\text{def}}{=} \min_{u\in\mathbb{R}^d} \left[P(u|\delta) + \frac{1}{2\gamma}\|u-\theta\|^2\right],\ \ \theta \in \mathbb{R}^d, \tag{9}$$

and its associated proximal map as

$$\mathrm{Prox}_\gamma(\theta|\delta) \stackrel{\text{def}}{=} \mathsf{Argmin}_{\ u\in\mathbb{R}^d} \left[P(u|\delta) + \frac{1}{2\gamma}\|u-\theta\|^2\right],\ \ \theta \in \mathbb{R}^d.$$

From the definition of P_γ and Prox_γ, we see that h_γ can be alternatively written as

$$\begin{aligned} h_\gamma(\theta|\delta) &= \ell(\theta) - \frac{\gamma}{2}\|\nabla\ell(\theta)\|^2 + P_\gamma\left(\theta - \gamma\nabla\ell(\theta)|\delta\right) && (10)\\ &= \ell(\theta) + \langle\nabla\ell(\theta), J_\gamma(\theta|\delta) - \theta\rangle + P(J_\gamma(\theta|\delta)|\delta) \\ &\quad + \frac{1}{2\gamma}\|J_\gamma(\theta|\delta) - \theta\|^2, && (11)\end{aligned}$$

where

$$J_\gamma(\theta|\ \delta) \stackrel{\text{def}}{=} \mathrm{Prox}_\gamma\left(\theta - \gamma\nabla\ell(\theta)|\delta\right).$$

For $\gamma > 0$, $\theta \in \mathbb{R}^d$, let $\mathsf{s}_\gamma(\theta) \in \mathbb{R}^d$ be such that

$$(\mathsf{s}_\gamma(\theta))_j \stackrel{\text{def}}{=} \mathsf{Argmin}_{\ u\in\mathbb{R}} \left[-\log p(u) + \frac{1}{2\gamma}(u-\theta_j)^2\right],\ \ 1 \leq j \leq d.$$

Then it is easy to check that $\mathrm{Prox}_\gamma(\theta|\delta) = \delta\cdot\mathsf{s}_\gamma(\theta)$. Hence by Equation (11), we see that $h_\gamma(\cdot|\delta)$ is computable is closed form if the map s_γ (the proximal map of the negative log-prior) is easy to compute. Although this limits the applicability of the method, there are several priors commonly used for which this holds, including the Gaussian prior, the Laplace prior and more generally the elastic-net prior given by

$$p(u) \propto \exp\left(-\alpha\lambda_1|u| - (1-\alpha)\lambda_2\frac{u^2}{2}\right),$$

as well as the generalized double Pareto of Armagan et al. (2013), and the (improper) prior distribution that arises from the MCP of Zhang (2010), given respectively by

$$p(u) = \frac{1}{2\lambda}\left(1+\frac{|u|}{\alpha\lambda}\right)^{-(\alpha+1)}, \quad \text{and} \quad p(u) = \exp\left(-\lambda\int_0^{|u|}\left(1-\frac{t}{\alpha\lambda}\right)_+ \mathrm{d}t\right).$$

For more general prior distributions for which the proximal map is intractable, numerical solvers may be considered, particularly since the components of $\mathsf{s}_\gamma(\theta)$ can be computed in parallel.

Remark 6. Since $\mathrm{Prox}_\gamma(\theta|\delta) = \delta\cdot\mathsf{s}_\gamma(\theta)$, and given (7), it is easily seen that in $\check{\Pi}_\gamma$, the δ_j's are conditionally independent Bernoulli random variables given θ. And given δ, one can use various MCMC algorithms, including gradient-based MCMC algorithms to update θ. Hence the proposed approximation produces a distribution that is easy to explore by MCMC compared to $\check{\Pi}$. A detailed discussion of MCMC implementation in the linear regression setting is deferred to Section 5.2.

4.1. Connection with spike-and-slab priors.

Another widely used approximation to the exact-sparsity prior is the prior obtained by replacing the point-mass at 0 by a Gaussian distribution with mean 0 and a small variance γ (George and McCulloch (1997); Ishwaran and Rao (2005); Rockova and George (2014); Narisetty and He (2014)). This leads to the following model.

$$\delta \sim \{\pi_\delta\}, \quad \theta_j|\delta \sim \left\{\begin{array}{cc} \mathbf{N}(0,\gamma) & \text{if } \delta_j = 0 \\ p(\cdot) & \text{if } \delta_j = 1 \end{array}\right., 1 \leq j \leq d, \;\; \text{and} \;\; Z|\delta,\theta \sim f_\theta, \tag{12}$$

for some constant $\gamma > 0$. Notice that given (δ,θ), we draw Z from f_θ. The resulting posterior distribution is

$$\bar{\Pi}_\gamma(\delta, \mathrm{d}\theta|z) \propto \pi_\delta \left\{\prod_{j:\,\delta_j=1} p(\theta_j)\right\}\left\{\prod_{j:\,\delta_j=0}\frac{1}{\sqrt{2\pi\gamma}}e^{-\frac{\theta_j^2}{2\gamma}}\right\} e^{-\ell(\theta)}\mathrm{d}\theta. \tag{13}$$

Just like $\check{\Pi}_\gamma$ (see Remark 6), one can easily construct MCMC algorithms to sample from $\bar{\Pi}_\gamma$. Indeed, given θ, the components of δ are independent Bernoulli; and given δ, we can update θ relatively easily by MCMC, depending on the choice of p. However, since (12) does not actively explore sparse models, one expects $\bar{\Pi}_\gamma$ to under-perform the exact-sparsity posterior $\check{\Pi}$.

Another possible approximation of the sparsity-inducing spike-and-slab prior is obtained by enforcing sparsity in model (12):

$$\delta \sim \{\pi_\delta\}, \quad \theta_j|\delta \sim \left\{\begin{array}{cc} \mathbf{N}(0,\gamma) & \text{if } \delta_j = 0 \\ p(\cdot) & \text{if } \delta_j = 1 \end{array}\right., 1 \leq j \leq d, \;\; \text{and} \;\; Z|\delta,\theta \sim f_{\theta_\delta}. \tag{14}$$

Notice that in (14) given (δ, θ), we draw Z from f_{θ_δ}, with a sparse parameter θ_δ. The posterior distribution thus defined is

$$\tilde{\Pi}_\gamma(\delta, \mathrm{d}\theta|z) \propto \pi_\delta \left\{ \prod_{j:\, \delta_j=1} p(\theta_j) \right\} \left\{ \prod_{j:\, \delta_j=0} \frac{1}{\sqrt{2\pi\gamma}} e^{-\frac{\theta_j^2}{2\gamma}} \right\} e^{-\ell(\theta_\delta)} \mathrm{d}\theta. \tag{15}$$

The distribution $\tilde{\Pi}_\gamma$ clearly seems a much better approximation to $\check{\Pi}$ than $\bar{\Pi}_\gamma$. And we show in Corollary 8 below that in the linear regression problem that $\check{\Pi}_\gamma$ is very close to $\tilde{\Pi}_\gamma$, since

$$\mathsf{d}_{\mathsf{w}}(\tilde{\Pi}_\gamma, \check{\Pi}) \approx \sqrt{d\gamma}, \quad \text{and} \quad \mathsf{d}_{\mathrm{tv}}(\check{\Pi}_\gamma, \tilde{\Pi}_\gamma) = O(d\gamma),$$

where d_{w} (resp. d_{tv}) denotes the Wasserstein metric (resp. the total variation metric). Hence for the purpose of approximating $\check{\Pi}$, $\tilde{\Pi}_\gamma$ and $\check{\Pi}_\gamma$ are roughly equivalent when γ is small, since the two are within $O(\gamma)$ of each other and within $O(\sqrt{\gamma})$ of $\check{\Pi}$. More broadly, the main feature of our proposed method is that its variational nature leads to a good mathematical control of its approximation errors, and this makes a detailed analysis of $\check{\Pi}_\gamma$ possible, as we do below (and also as in Atchade and Bhattacharyya (2018)).

4.2. **Approximation bounds.** We will now derive a result that bounds the β-distance between $\check{\Pi}_\gamma$ and $\check{\Pi}$. We recall that $\tilde{\Pi}_\gamma$ denotes the posterior distribution defined in (15). We define

$$\varrho_\gamma(z) \overset{\text{def}}{=} \log \int e^{r_\gamma(\delta,\theta)} \tilde{\Pi}_\gamma(\mathrm{d}\delta, \mathrm{d}\theta|z), \tag{16}$$

where

$$r_\gamma(\delta, \theta) \overset{\text{def}}{=} \langle \nabla\ell(\theta) - \nabla\ell(\theta_\delta), \theta - \theta_\delta) \rangle + \frac{\gamma}{2} \|\delta \cdot \nabla\ell(\theta) + \delta \cdot g\left(\theta_\delta|\delta\right)\|^2.$$

For simplicity, we shall omit the dependence of $r_\gamma(\delta, \theta)$ on z (same with $\ell(\theta)$ and $\nabla\ell(\theta)$). We note that by the convexity of ℓ, $r_\gamma(\delta, \theta) \geq 0$. Hence $\varrho_\gamma(z) \geq 0$.

Theorem 7. *Assume H1, for some fixed data z.*

(1) For any $\gamma > 0$, we have

$$\mathsf{d}_{\mathsf{w}}\left(\tilde{\Pi}_\gamma(\cdot|z), \check{\Pi}(\cdot|z)\right) \leq \sqrt{\gamma d}.$$

(2) Suppose that there exists $\gamma_0 > 0$ such that $\check{\Pi}_{\gamma_0}(\cdot|z)$ is well-defined. Then for all $\gamma \in (0, \gamma_0]$, $\check{\Pi}_\gamma(\cdot|z)$ is well-defined and

$$\mathsf{d}_{\mathrm{tv}}\left(\check{\Pi}_\gamma(\cdot|z), \tilde{\Pi}_\gamma(\cdot|z)\right) \leq 2\left(1 - e^{-\varrho_\gamma(z)}\right).$$

Proof. See Section 7.2. □

Combining the two parts of the theorem yields for all $\gamma \in (0, \gamma_0]$

$$\mathsf{d}_\beta\left(\check{\Pi}_\gamma(\cdot|z), \check{\Pi}(\cdot|z)\right) \leq \sqrt{\gamma d} + 2\left(1 - e^{-\varrho_\gamma(z)}\right). \tag{17}$$

Notice that $1 - e^{-x} \leq x$ for all $x \geq 0$. Therefore, the convergence to zero of $\mathsf{d}_\beta\left(\check{\Pi}_\gamma(\cdot|z), \check{\Pi}(\cdot|z)\right)$ would follow if the term $\varrho_\gamma(z)$ converges to 0 as $\gamma \to 0$. We show how to obtain such result below. To save space we focus on the linear regression model, although more general result is possible along similar lines.

5. Application to Bayesian linear regression with sparse priors

As an application we consider a high-dimensional linear regression problem, with dependent variable $z \in \mathbb{R}^n$, and design matrix $X \in \mathbb{R}^{n \times d}$. The variance term σ^2 is assumed known. The negative-log-likelihood function ℓ for this problem can be taken as

$$\ell(\theta) = \frac{1}{2\sigma^2}\|z - X\theta\|^2, \;\; \theta \in \mathbb{R}^d.$$

We will set up the prior distribution of θ using $\delta \in \Delta$, and using an auxiliary variable $\phi \overset{\text{def}}{=} (\mathsf{q}, \lambda_1, \lambda_2)$, where $\mathsf{q} \in (0,1)$ is a sparsity parameter, and $\lambda_1 > 0$, $\lambda_2 > 0$ are regularization parameters. Given ϕ, we assume that the components of δ are independent and identically distributed, with distribution $\mathbf{Ber}(\mathsf{q})$. Hence $\pi_\delta = \mathsf{q}^{\|\delta\|_0}(1-\mathsf{q})^{d-\|\delta\|_0}$. Given ϕ and δ, the components of θ are independent, and for $1 \leq j \leq d$,

$$\theta_j|\delta, \phi \sim \left\{ \begin{array}{cl} \mathsf{Dirac}(0) & \text{if } \delta_j = 0 \\ \mathsf{EN}\left(\frac{\lambda_1}{\sigma^2}, \frac{\lambda_2}{\sigma^2}\right) & \text{if } \delta_j = 1 \end{array} \right.,$$

where $\mathsf{Dirac}(0)$ is the Dirac measure on $\mathbb{R}$ with full mass at 0, and $\mathsf{EN}(\lambda_1/\sigma^2, \lambda_2/\sigma^2)$ is the (elastic-net) distribution with density given by

$$\frac{1}{Z(\phi)} \exp\left(-\alpha\frac{\lambda_1}{\sigma^2}|x| - (1-\alpha)\frac{\lambda_2}{2\sigma^2}x^2\right), \;\; x \in \mathbb{R}, \tag{18}$$

for a parameter $\alpha \in [0,1]$, assumed known. We recover the Gaussian prior $\mathsf{N}(0, \frac{\sigma^2}{\lambda_2})$ by setting $\alpha = 0$, and we recover the Laplace (double-exponential) prior $\mathsf{Laplace}(\frac{\lambda_1}{\sigma^2})$ by setting $\alpha = 1$. The normalizing constant $Z(\phi)$ can be written as

$$Z(\phi) = \left\{ \begin{array}{ll} \sigma\sqrt{\frac{2\pi}{(1-\alpha)\lambda_2}}\mathsf{erfcx}\left(\frac{\alpha\lambda_1}{\sigma\sqrt{2(1-\alpha)\lambda_2}}\right) & \text{if } \alpha \in [0,1) \\ \frac{2\sigma^2}{\lambda_1} & \text{if } \alpha = 1 \end{array} \right.,$$

where $\mathsf{erfcx}(x)$ is the scaled complementary error function, which can be written as $\mathsf{erfcx}(x) = 2e^{x^2}\Phi(-\sqrt{2}x)$, where Φ is the cdf of standard normal distribution. The prior density (18) is a reparametrization of the elastic-net (Zou and Hastie (2005)) prior used by Li and Lin (2010). Notice that $\alpha = 1$ makes λ_2 inactive, and setting

$\alpha = 0$ makes λ_1 inactive. With this parametrization of elastic net prior (18), the proximal function $\mathrm{Prox}_\gamma(\theta|\delta)$ can be computed as

$$\mathrm{Prox}_\gamma(\theta|\delta) = \delta \cdot \mathsf{s}_\gamma(\theta),$$

where

$$(\mathsf{s}_\gamma(\theta))_j = \frac{\mathsf{sign}(\theta_j)\left(|\theta_j| - \alpha\gamma\frac{\lambda_1}{\sigma^2}\right)_+}{1 + \gamma\frac{\lambda_2}{\sigma^2}(1-\alpha)}, \tag{19}$$

and for $x \in \mathbb{R}$, $x_+ \overset{\text{def}}{=} \max(x, 0)$, $\mathsf{sign}(x)$ is the sign of x. In the next result, we derive an approximation bound. For a matrix A, let $\lambda_{\mathsf{max}}(A)$ denote its largest eigenvalue.

Corollary 8. *Suppose that $(1-\alpha)\lambda_2 \leq \lambda_{\mathsf{max}}(X'X)$, and suppose that $\gamma > 0$ satisfies*

$$\frac{4\gamma}{\sigma^2}\lambda_{\mathsf{max}}(X'X) \leq 1. \tag{20}$$

Then for all $z \in \mathbb{R}^n$, $\check{\Pi}_\gamma(\cdot|z)$ is a well-defined probability measure on $\Delta \times \mathbb{R}^d$, and

$$\mathsf{d}_\beta\left(\check{\Pi}_\gamma(\cdot|z), \check{\Pi}(\cdot|z)\right) \leq \sqrt{\gamma d} + 2\left(1 - e^{-\varrho_\gamma(z)}\right),$$

where $\varrho_\gamma(z)$ satisfies

$$\varrho_\gamma(z) \leq \frac{3\gamma}{2}\left(\frac{\alpha\lambda_1}{\sigma^2}\right)^2 d + \frac{3\gamma}{\sigma^2}\lambda_{\mathsf{max}}(X'X)\left[d\left(3 + \log Z(\phi)\right) + \frac{\|z\|^2}{2\sigma^2}\right]. \tag{21}$$

Proof. See Section 7.3. □

The bound in (20) provides some guidelines for choosing γ, as it suggests that one can choose γ as

$$\gamma = \min\left(\frac{1}{d}, \frac{\gamma_0\sigma^2}{\lambda_{\mathsf{max}}(X'X)}\right), \quad \gamma_0 \in (0, 1/4]. \tag{22}$$

As we show in the simulations setting $\gamma_0 \in (0.1, 1/4]$ works well. We cautious against setting γ_0 overly small, since in that case the mixing time of the MCMC sampler proposed below to sample from $\check{\Pi}_\gamma$ increases.

5.1. **Dealing with the hyper-parameter ϕ.** We use a fully Bayesian approach for selecting the hyper-parameter $\phi = (\mathsf{q}, \lambda_1, \lambda_2)$. We assume independent priors such that $\mathsf{q} \sim \mathbf{Beta}(1, d^u)$ for some constant $u > 1$, $\lambda_1 \sim \mathbf{U}(\mathsf{a}, M)$, and $\lambda_2 \sim \mathbf{U}(\mathsf{a}, M)$ for some small positive constant a (we use $\mathsf{a} = 10^{-5}$ in the simulations), and for a large positive constant M such that $(1-\alpha)M \leq \lambda_{\mathsf{max}}(X'X)$.

5.2. **Markov Chain Monte Carlo.** We propose a Metropolized-Gibbs strategy in order to draw samples from $\check{\Pi}_\gamma$.

5.2.1. *Updating δ.* Given θ and ϕ, it is easy to see that $h_\gamma(\theta|\delta)$ depends on δ_j only through the expression

$$\delta_j\left[(\nabla\ell(\theta))_j d_j + \log Z(\phi) + \frac{\alpha\lambda_1|d_j| + 0.5(1-\alpha)\lambda_2 d_j^2}{\sigma^2} + \frac{d_j^2 - 2\theta_j d_j}{2\gamma}\right],$$

where d_j is the j-th component of $\mathsf{s}_\gamma(\theta - \gamma\nabla\ell(\theta); \lambda_1/\sigma^2, \lambda_2/\sigma^2)$. Hence, we update jointly and independently the δ_j by setting $\delta_j = 1$ with probability $e^r/(1+e^r)$, where

$$r = \log\frac{\mathsf{q}}{1-\mathsf{q}} + \frac{1}{2}\log(2\pi\gamma) \\ - \left[(\nabla\ell(\theta))_j d_j + \log Z(\phi) + \frac{\alpha\lambda_1|d_j| + 0.5(1-\alpha)\lambda_2 d_j^2}{\sigma^2} + \frac{d_j^2 - 2\theta_j d_j}{2\gamma}\right].$$

5.2.2. *Updating θ.* Given δ and ϕ, we update the components of θ using a mix of an independence Metropolis sampler, and a gradient-based Metropolis-Hastings algorithm. The function $\theta \mapsto h_\gamma(\theta|\delta)$ is differentiable and its gradient is given by

$$\nabla_\theta h_\gamma(\theta|\delta) = \frac{1}{\gamma}\left(I_d - \gamma\nabla^{(2)}\ell(\theta)\right)(\theta - J_\gamma(\theta|\delta,\phi)).$$

To avoid dealing with second order derivatives, and since γ is typically small, we approximate $I_d - \gamma\nabla^{(2)}\ell(\theta)$ by I_d. This implies that we can approximate $\nabla_\theta h_\gamma(\theta|\delta)$ by

$$G_\gamma(\theta|\delta) \stackrel{\text{def}}{=} \frac{1}{\gamma}(\theta - J_\gamma(\theta|\delta)), \quad \text{and} \quad \bar{G}_\gamma(\theta|\delta) \stackrel{\text{def}}{=} \frac{\mathsf{c}}{\mathsf{c} \vee \|G_\gamma(\theta|\delta)\|} G_\gamma(\theta|\delta), \tag{23}$$

for a positive constant c. The function $\bar{G}_\gamma$ is introduced for further stability, in the spirit of the truncated Metropolis adjusted Langevin algorithm (see e.g. Atchadé (2006)). Hence, given δ and ϕ, and given the non-selected components of θ we update each selected components of θ (one component at the time) using a gradient-based Metropolis-Hastings algorithm where the drift function is given by the corresponding components of $\bar{G}_\gamma$. As in Atchadé (2006) we adaptively tune the variance of the proposal distribution (we use the same value for all components) to automatically yield a 60% acceptance probability. This update is similar to the proximal MaLa of Pereyra (2013).

However, when $\delta_j = 0$, the corresponding component of $G_\gamma(\theta|\delta)$ is θ_j/γ and is typically very large and not very informative (particularly for γ small). To deal with this, we use the following strategy. We update the components θ_j for which $\delta_j = 1$ – one component at the time – using the gradient-based algorithm outlined above. Then, we group together all the components for which $\delta_j = 0$ and we update them jointly using an independence Metropolis sampler. The proposal density of the Independence Metropolis sampler is built by approximating $J_\gamma(\theta|\delta)$ by $\mathrm{Prox}_\gamma(\theta -$

$\gamma\nabla\ell(\theta\cdot\delta)|\delta)$ in h_γ. This approximation comes from the fact that for $\gamma \approx 0$, $J_\gamma(\theta|\delta) = \mathrm{Prox}_\gamma(\theta - \gamma\nabla\ell(\theta)|\delta) \approx \mathrm{Prox}_\gamma(\theta - \gamma\nabla\ell(\theta\cdot\delta)|\delta)$. The resulting proposal density is the density of the Gaussian distribution

$$\mathbf{N}\left(\frac{\gamma}{\sigma^2}\Sigma X'_{\delta^c}X\left[\mathrm{Prox}_\gamma\left(\theta - \gamma\nabla\ell(\theta\cdot\delta)|\delta\right) - \delta\cdot\theta\right], \gamma\Sigma\right), \qquad \text{where} \quad \Sigma \overset{\mathrm{def}}{=} \left(I_{\|\delta^c\|} - \frac{\gamma}{\sigma^2}X'_{\delta^c}X_{\delta^c}\right)^{-1},$$

where $\delta^c \overset{\mathrm{def}}{=} 1 - \delta$, and for any $\delta \in \Delta$, $X_\delta \in \mathbb{R}^{n\times\|\delta\|}$ denotes the sub-matrix of X obtained by selecting the columns for which $\delta_j = 1$. Notice that under the assumption $\gamma \le \frac{\sigma^2}{4\lambda_{\mathsf{max}}(X'X)}$, the matrix Σ is always positive definite. We found this independence sampler to be extremely efficient, with an acceptance probability typically above 90%.

5.2.3. *Updating* $\phi = (q, \lambda_1, \lambda_2)$. We update $\mathsf{q} \sim \mathbf{Beta}(\|\delta\|_1 + 1, d + d^u - \|\delta\|_1)$, and we update (λ_1, λ_2) jointly using a Random Walk Metropolis algorithm with Gaussian proposal. For improved mixing, we adaptively tune the scale parameter of the proposal density to give an acceptance probability of 30% (for more details on adaptive MCMC, see for instance Atchadé et al. (2011) and the reference therein).

5.3. **Simulation results and comparison with STMaLa.** We illustrate the method with a simulated data example. All the computations in this example were done using `Matlab 7.14` on a 2.8 GHz Quad-Core `Mac Pro` with 24 GB of 1066 DDR3 Ram.

We set $n = 200$, $p = 500$ and we generate the design matrix X by simulating the rows of X independently from a Gaussian distribution with correlation $\rho^{|j-i|}$ between components i and j. We set $\rho = 0.9$. Using X, we general the outcome $z = X\theta_\star + \sigma\epsilon$, with $\sigma = 1$ that we assume known. We build $\theta_\star$ by randomly selecting 10 components that we fill with draws from the uniform distribution $\epsilon\mathbf{U}(\mathsf{v}/2, 3\mathsf{v}/2)$, where $\epsilon = \pm 1$ with probability 1/2, all other components being set to zero. We consider two cases for v: $\mathsf{v} = 1$ (SCENARIO 1), and $\mathsf{v} = \sqrt{\log(d)/n} \approx 0.18$ (SCENARIO 2).

We set $\gamma = \gamma_0\sigma^2/\lambda_{\mathsf{max}}(X'X)$ as prescribed by (22) with two choices of γ_0: $\gamma_0 = 0.25$, and $\gamma_0 = 0.01$.

We compare these two samplers to the STMaLa sampler of Schreck et al. (2013). In our notations, the target posterior distribution of STMaLa is

$$\pi_\delta \exp\left(-\frac{1}{2\sigma^2}\|z - X\theta\|_2^2\right)\prod_{j=1}^{d}\left(1 + \frac{\theta_j^2}{2aK}\right)^{-a-\frac{1}{2}}\mu_\delta(\mathrm{d}\theta),$$

for positive hyper-parameters a, K, where $\pi_\delta = \mathsf{q}^{\|\delta\|_0}(1-\mathsf{q})^{d-\|\delta\|_0}$ The comparison is slightly tricky because STMaLa uses a different prior, namely a scale-mixture of Gaussian as slab density. However, we expect both posterior distribution on (δ, θ) to be close, and we expect the true value $(\delta_\star, \theta_\star)$ to be close to the center of both distributions. For the STMaLa, we use the Matlab code provided online by the authors, with the default setting. Unlike our approach, this sampler requires the true value of the sparsity parameter q, which we provide. We also edit their code to return the summary statistics presented below.

We evaluate the mixing of these samplers by computing the following two metrics along the MCMC iterations: the relative error and the F-score (to evaluate structure recovery), defined respectively as

$$\mathcal{E}^{(k)} = \frac{\|\theta^{(k)} - \theta_\star\|}{\|\theta_\star\|}, \quad \text{and} \quad \mathcal{F}^{(k)} = \frac{2 \times \mathsf{SEN}^{(k)}\mathsf{PREC}(k)}{\mathsf{SEN}^{(k)} + \mathsf{PREC}(k)},$$

where

$$\mathsf{SEN}^{(k)} = \frac{\sum_{j=1}^d \mathbf{1}_{\{|\delta_j^{(k)}|>0\}}\mathbf{1}_{\{|\delta_{\star,j}|>0\}}}{\sum_{j=1}^d \mathbf{1}_{\{|\delta_{\star,j}|>0\}}}, \ \mathsf{PREC}(k) = \frac{\sum_{j=1}^d \mathbf{1}_{\{|\delta_j^{(k)}|>0\}}\mathbf{1}_{\{|\delta_{\star,j}|>0\}}}{\sum_{j=1}^d \mathbf{1}_{\{|\delta_j^{(k)}|>0\}}}. \tag{24}$$

In stationarity we expect values of $\mathcal{E}^{(k)}$ (resp. $\mathcal{F}^{(k)}$) to be close to zero (resp. one). In the absence of a better metric, we will graphically access the mixing time of the samplers by looking at how quickly the sequence $\mathcal{E}^{(k)}$ (resp. $\mathcal{F}^{(k)}$) converges towards zero (resp. one). In order to account for the computing time, and for better comparison, we plot these metrics, not as function of the iterations k, but as function of the computing time needed to reach iteration k. For further stability in the comparison, we repeat all the samplers 30 times and average the two metrics and the computing times over these 30 replications.

All the chains are initialized by setting all components of $\theta^{(0)}$ (and $\delta^{(0)}$) to zero. We run the samplers for a number of iterations that depends on $\theta_\star$. In SCENARIO 1, we run the newly proposed sampler for $10,000$, and we run STMaLa for $120,000$ iterations. In SCENARIO 2, we run our proposed sampler for $40,000$, and we run STMaLa for $250,000$ iterations.

Figure 2 and 3 present the results. First, we observe that that $\gamma_0 = 0.25$ mixes significantly better than $\gamma_0 = 0.01$. We notice also that $\check{\Pi}_\gamma$ approximates $(\theta_\star, \delta_\star)$ only slightly better when $\gamma_0 = 0.01$ compared to $\gamma_0 = 0.25$. Overall, we found that $\gamma_0 \in (0.1, 0.25)$ produces a very good approximation.

We also look at the usual sample path mixing of the proposed sampler by plotting the trace plot, histogram, and the autocorrelation plot from a single run of the sampler (Figure 4). Here, we consider only SCENARIO 1, and we modify the true parameter $\theta_\star$ to have one significant but small component. All other parameters are as above. We look at the MCMC output $\{\theta_j^{(k)},\ k \geq 0\}$, for one component j for which $\delta_j = 0$, for

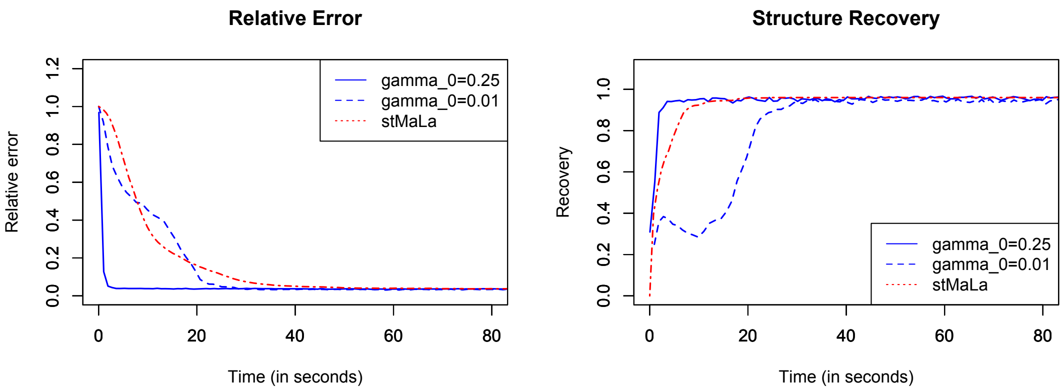

FIGURE 2. Relative error and structure recovery as function of time in SCENARIO 1. Based on 30 MCMC replications. The curves are sub-sampled to improve the readability of the figure.

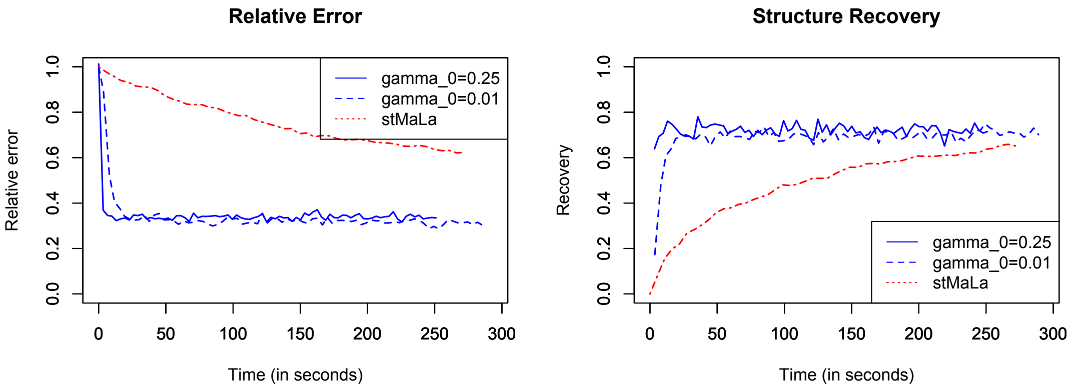

FIGURE 3. Relative error and structure recovery as function of time in SCENARIO 2. Based on 30 MCMC replications. The curves are sub-sampled to improve the readability of the figure.

the weakly significant component, and for one significantly large component. From this sample path perspective, the plots suggest that the proposed MCMC sampler has a good mixing, except in the second case where the marginal distribution of the parameter is a bi-modal distribution and the sampler needs to switch between the two modes.

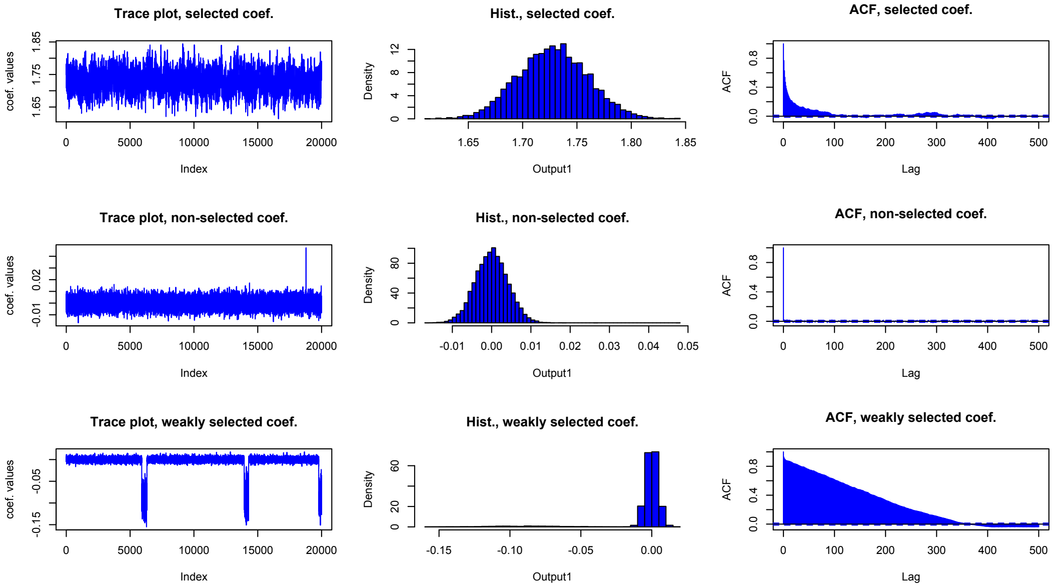

FIGURE 4. Trace plot, histogram, and autocorrelation plot, from one MCMC run, using $\gamma_0 = 0.25$. Top row: a strongly significant component j; middle row: non-significant component; bottom row: a weakly significant component.

5.4. **Empirical Bayes implementation and further experimentation.** A limitation of the methodology is that σ^2 is assumed known, which is rarely the case in practice. We explore by simulation an empirical Bayes solution whereby σ^2 is estimated from data. Following Reid et al. (2013) we estimate σ^2 by

$$\hat{\sigma}_n^2 = \frac{1}{n - \hat{s}_{\lambda_n}} \sum_{i=1}^{n} \left(y_i - x_i \hat{\beta}_{\lambda_n} \right)^2,$$

where $\hat{\beta}_\lambda$ is the lasso estimate at regularization level λ, and λ_n is selected by 10-fold cross-validation, and where $\hat{s}_{\lambda_n}$ is the number of non-zeros components of $\hat{\beta}_{\lambda_n}$. In the cross-validation, we choose λ_n as the value of λ that minimizes the MSE. This leads to the empirical Bayes Moreau envelop posterior approximation that we denote $\check{\Pi}_\gamma(\cdot|z, \hat{\sigma}_n^2)$. We do a simulation study using a semi-real dataset to compare the distributions $\check{\Pi}_\gamma(\cdot|z, \hat{\sigma}_n^2)$ and $\check{\Pi}_\gamma(\cdot|z)$ (with the true value of σ^2 set to one). We use the colon dataset (Buhlmann and Mandozzi (2014)) downloaded from `http://stat.ethz.ch/~dettling/bagboost.html`. The data gives microarray gene expression levels for $2,000$ genes for $n = 62$ patients in a colon cancer study. We randomly select a subset of $p = 1,000$ variables to form a design matrix $X \in \mathbb{R}^{62\times 1,000}$. Following Buhlmann and Mandozzi (2014), we normalize each column of X to have mean zero and variance unity. We simulate a sparse signal vector $\theta_\star \in \mathbb{R}^p$ with $s = 5$ non-zeros components, and where the non-zeros components are drawn from $\mathbf{U}(-\mathsf{v}-1, -\mathsf{v}) \cup (\mathsf{v}, \mathsf{v}+1)$. We consider two scenarios: $\mathsf{v} = 1$ and $\mathsf{v} = 3$. Using X and $\theta_\star$, we generate $z = X\theta_\star + \sigma\epsilon$, with $\sigma = 1$, and $\epsilon \sim \mathbf{N}(0, I_n)$.

We set γ as in (22) with $\gamma_0 = 0.25$. We evaluate the samplers along the same metrics $\mathcal{E}$ and $\mathcal{F}$. We average the results over 30 replications[2] of the samplers, where each sampler is run for $50,000$ iterations. The result is presented on Table 1. We notice that the recoery of $\theta_\star$ is poor in both cases when $\mathsf{v} = 1$. When the signal is strong ($\mathsf{v} = 3$), the empirical Bayes posterior distribution performs well, but as expected, under-performs the posterior distribution with known variance.

	Weak signal ($\mathsf{v} = 1$)		Strong signal ($\mathsf{v} = 3$)	
	EB	True σ	EB	True σ
Relative error (in %)	97.3	91.7	12.4	9.4
F-score (in %)	14.5	25.1	79.6	88.5

TABLE 1. Table showing the posterior estimates $(N-B)^{-1}\sum_{k=B+1}^{N}\mathcal{E}^{(k)}$, and $(N-B)^{-1}\sum_{k=B+1}^{N}\mathcal{F}^{(k)}$, averaged over 30 MCMC replications, each MCMC run is 5×10^4 iterations.

6. Further Discussion

We have proposed a forward-backward approximation for spike-and-slab posterior distributions. The methodology can be applied more broadly to statistical models with smooth and concave log-likelihood functions, and for several classes of slab densities. Several theoretical issues remain. One interesting problem that we did not directly address concerns the mixing properties of the proposed MCMC algorithms, and the trade-off inherent to the methodology between good approximation properties of $\check{\Pi}_\gamma$, and good mixing of gradient-based MCMC simulation from $\check{\Pi}_\gamma$. Another potentially interesting direction of research is the idea of treating $\check{\Pi}_\gamma$ itself as a quasi-posterior distribution, and investigating directly its posterior contraction properties.

7. Proof of the main results

For convenience, we introduce the product space $\bar{\Theta} \stackrel{\text{def}}{=} \Delta \times \mathbb{R}^d$ that we implicitly equip with the metric $\mathsf{d}_{\bar{\Theta}}(\bar{\theta}_1, \bar{\theta}_2) \stackrel{\text{def}}{=} \sqrt{\|\delta_1 - \delta_2\|_0^2 + \|\theta_1 - \theta_2\|^2}$, $\bar{\theta}_j = (\delta_j, \theta_j)$, $j = 1, 2$.

[2] here only X and $\theta_\star$ are kept fixed. For each replication, the dataset z is re-simulated, and σ_n^2 is re-estimated.

7.1. **Proof of Proposition 1.** For all $x \in \mathbb{R}^d$, and $\gamma \in (0, \gamma_0]$, $e^{-h(x)} \leq e^{-h_\gamma(x)} \leq e^{-h_{\gamma_0}(x)}$. Hence $Z \leq Z_\gamma \leq Z_{\gamma_0} < \infty$. Since μ is the Lebesgue measure on $\mathbb{R}^d$, we shall write it as $\mathrm{d}x$. For any bounded measurable function $f : \mathbb{R}^d \to \mathbb{R}$, we have

$$\begin{aligned} |\Pi_\gamma(f) - \Pi(f)| &\leq \frac{1}{Z_\gamma}\left|\int_{\mathbb{R}^d} f(x)\left(e^{-h_\gamma(x)} - e^{-h(x)}\right)\mathrm{d}x\right| \\ &\quad + \frac{(Z_\gamma - Z)}{Z_\gamma Z}\int_{\mathbb{R}^d} |f(x)| e^{-h(x)}\mathrm{d}x \\ &\leq \frac{2\|f\|_\infty}{Z_\gamma}\int_{\mathbb{R}^d}\left(e^{-h_\gamma(x)} - e^{-h(x)}\right)\mathrm{d}x \\ &= 2\|f\|_\infty\left(1 - \frac{Z}{Z_\gamma}\right). \end{aligned}$$

The fact that $Z_\gamma \to Z$ as $\gamma \downarrow 0$, follows from Lebesgue's monotone convergence applied to $e^{-h_{\gamma_0}} - e^{-h_\gamma}$.

7.2. **Proof of Theorem 7.** We work on the product space $\bar{\Theta} = \Delta \times \mathbb{R}^d$ introduced above. Throughout the proof, we assume that z is fixed, and at times we write $\check{\Pi}(\cdot|z)$ simply as $\check{\Pi}$. Same for $\tilde{\Pi}_\gamma(\cdot|z)$ and $\check{\Pi}_\gamma(\cdot|z)$.

We prove the theorem in two steps. First in Lemma 9, we bound the Wasserstein distance between the distributions $\tilde{\Pi}_\gamma$ and $\check{\Pi}$ by showing that for all $\gamma > 0$,

$$\mathsf{d}_{\mathsf{w}}(\tilde{\Pi}_\gamma, \check{\Pi}) \leq \sqrt{\gamma d}. \tag{25}$$

Then in Lemma 11 we bound the total variation distance between $\check{\Pi}_\gamma$ and $\tilde{\Pi}_\gamma$ by showing that for all $\gamma \in (0, \gamma_0]$,

$$\mathsf{d}_{\mathsf{tv}}(\tilde{\Pi}_\gamma, \check{\Pi}_\gamma) \leq 2\left(1 - e^{-\varrho_\gamma(z)}\right). \tag{26}$$

It is clear from their definitions that both the Wasserstein metric and the total variation metric are upper bounds for the metrix β, and the Theorem 7 follows by combining (25) and (26). The proof of Lemma 11 relies on a comparison result between the functions h and h_γ established in Lemma 10 that is also of independent interest.

Lemma 9. *Let $\tilde{\Pi}_\gamma$ be the probability measure defined in (15). For all $\gamma > 0$,*

$$\sqrt{\frac{2}{\pi}}\sqrt{\gamma d}\left(1 - \frac{1}{d}\mathbb{E}(\|\eta\|_0)\right) \leq \mathsf{d}_{\mathsf{w}}(\tilde{\Pi}_\gamma, \check{\Pi}) \leq \sqrt{d\gamma}. \tag{27}$$

Proof. For all $\delta \in \Delta$, and $\gamma > 0$, by integrating out the non-selected components we have

$$\left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_1}{2}}\int_{\mathbb{R}^d} e^{-\frac{1}{2\gamma}\|\theta - \theta_\delta\|^2} e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta = \int_{\mathbb{R}^d} e^{-h(\theta|\delta)}\mu_\delta(\mathrm{d}\theta).$$

This implies that the distributions $\check{\Pi}(\cdot|z)$ and $\tilde{\Pi}_\gamma(\cdot|z)$ have the same normalizing constant given by

$$C = \sum_{\delta\in\Delta} \pi_\delta C(\delta), \quad \text{where} \quad C(\delta) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} e^{-h(\theta|\delta)}\mu_\delta(\mathrm{d}\theta). \tag{28}$$

Using this notation, we can write

$$\check{\Pi}(\delta, \mathrm{d}\theta|z) = \frac{\pi_\delta C(\delta)}{C}\check{\Pi}(\mathrm{d}\theta|\delta, z), \quad \text{where} \quad \check{\Pi}(\mathrm{d}\theta|\delta, z) \stackrel{\text{def}}{=} \frac{1}{C(\delta)}e^{-h(\theta|\delta)}\mu_\delta(\mathrm{d}\theta),$$

and and

$$\tilde{\Pi}_\gamma(\delta, \mathrm{d}\theta|z) = \frac{\pi_\delta C(\delta)}{C}\tilde{\Pi}_\gamma(\mathrm{d}\theta|\delta, z),$$
$$\text{where} \quad \tilde{\Pi}_\gamma(\mathrm{d}\theta|\delta, z) \stackrel{\text{def}}{=} \frac{1}{C(\delta)}e^{-h(\theta_\delta|\delta)}\left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_1}{2}} e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2}\mathrm{d}\theta.$$

We build the following coupling of $\check{\Pi}$ and $\tilde{\Pi}_\gamma$. First we generate $\eta \in \Delta$ from the distribution $\delta \mapsto \frac{\pi_\delta C(\delta)}{C}$, and we generate $\check{\vartheta}|\eta \sim \check{\Pi}(\mathrm{d}\theta|\eta, z)$. Hence clearly, $(\eta, \check{\vartheta}) \sim \check{\Pi}$. Given $(\eta, \check{\vartheta})$, we generate $\tilde{\vartheta}$ as follows. If $\eta_j = 1$, we set $\tilde{\vartheta}_j = \check{\vartheta}_j$. Otherwise we generate independently $Z_j \sim \mathbf{N}(0, 1)$, and set $\tilde{\vartheta}_j = \sqrt{\gamma}Z_j$. It is also easy to check that $(\eta, \tilde{\vartheta}) \sim \tilde{\Pi}_\gamma$.

For any Lipschitz function on $\bar{\Theta}$ with Lipschitz constant less of equal to 1, we have

$$\left|\int f(\delta,\theta)\tilde{\Pi}_\gamma(\mathrm{d}\delta, \mathrm{d}\theta) - \int f(\delta,\theta)\check{\Pi}(\mathrm{d}\delta, \mathrm{d}\theta)\right| = \left|\mathbb{E}\left[f(\eta,\tilde{\vartheta}) - f(\eta,\check{\vartheta})\right]\right|$$
$$\leq \mathbb{E}\left[\|\tilde{\vartheta} - \check{\vartheta}\|\right] \leq \sqrt{d\gamma}\sqrt{1 - \frac{1}{d}\mathbb{E}(\|\eta\|_0)} \leq \sqrt{d\gamma},$$

and this proves the upper bound. For the lower bound, consider the function $f_0(\delta, \theta) = \frac{1}{\sqrt{d}}\sum_{j=1}^d |\theta_j|$. It is Lipschitz with Lipschitz constant 1. Hence

$$\mathsf{d}_\mathsf{w}(\tilde{\Pi}_\gamma, \check{\Pi}) \geq \left|\mathbb{E}\left[f_0(\eta,\tilde{\vartheta}) - f_0(\eta,\check{\vartheta})\right]\right| = \sqrt{\frac{\gamma}{d}}\mathbb{E}\left(\sum_{j:\ \eta_j=0}|Z_j|\right)$$
$$= \sqrt{\frac{2}{\pi}}\sqrt{\gamma d}\left(1 - \frac{1}{d}\mathbb{E}(\|\eta\|_0)\right),$$

and the result is proved. □

Lemma 10. *Assume H1 and fix* $\delta \in \Delta$. *For all* $\theta \in \mathbb{R}^d$,

$$h(\theta_\delta|\delta) + \frac{1}{2\gamma}\|\theta - \theta_\delta\|^2 \geq h_\gamma(\theta|\delta) \geq h(\theta_\delta|\delta) + \frac{1}{2\gamma}\|\theta - \theta_\delta\|^2 - r_\gamma(\theta, \delta), \tag{29}$$

with

$$r_\gamma(\delta, \theta) \stackrel{\text{def}}{=} \langle\nabla\ell(\theta) - \nabla\ell(\theta_\delta), \theta - \theta_\delta)\rangle + \frac{\gamma}{2}\|\delta \cdot \nabla\ell(\theta) + \delta \cdot g\left(\theta_\delta|\delta\right)\|^2,$$

and where $g(\theta_\delta|\delta)$ denotes a sub-gradient of $P(\cdot|\delta)$ at θ_δ. It follows in particular that for all $\theta \in \mathbb{R}^d$, $h_\gamma(\theta|\delta) \uparrow h(\theta|\delta)$, as $\gamma \downarrow 0$.

Proof. From the definition we have

$$\begin{aligned} h_\gamma(\theta|\delta) &= \min_{u\in\mathbb{R}^d}\left[\ell(\theta) + \langle \nabla\ell(\theta), u-\theta\rangle + P(u|\delta) + \frac{1}{2\gamma}\|u-\theta\|^2\right] \\ &\leq \ell(\theta) + \langle \nabla\ell(\theta), \theta_\delta-\theta\rangle + P(\theta_\delta|\delta) + \frac{1}{2\gamma}\|\theta-\theta_\delta\|^2. \end{aligned}$$

By convexity of ℓ, $\ell(\theta) + \langle \nabla\ell(\theta), \theta_\delta-\theta\rangle \leq \ell(\theta_\delta)$, which proves the first inequality in (29). To prove the second inequality, we start by using again the convexity of ℓ to write for all $\theta \in \mathbb{R}^d$,

$$\ell(\theta) \geq \ell(\theta_\delta) + \langle \nabla\ell(\theta_\delta), \theta-\theta_\delta\rangle .$$

Hence for all $\theta \in \mathbb{R}^d$, adding $\langle \nabla\ell(\theta), J_\gamma(\theta|\delta)-\theta\rangle$ on both sides and rearranging, we get

$$\begin{aligned} \ell(\theta) + \langle \nabla\ell(\theta), J_\gamma(\theta|\delta)-\theta\rangle \geq \ell(\theta_\delta) + \langle \nabla\ell(\theta_\delta) - \nabla\ell(\theta), \theta-\theta_\delta\rangle \\ + \langle \nabla\ell(\theta), J_\gamma(\theta|\delta)-\theta_\delta\rangle , \quad (30) \end{aligned}$$

where we recall that $J_\gamma(\theta|\delta) = \operatorname{Prox}_\gamma(\theta - \gamma\nabla\ell(\theta)|\delta)$. By H1, $P(\cdot|\delta)$ is convex, and if $g(\theta_\delta|\delta)$ denotes a sub-gradient of $P(\cdot|\delta)$ at θ_δ, we have

$$P(J_\gamma(\theta|\delta)|\delta) \geq P(\theta_\delta|\delta) + \langle g\left(\theta_\delta|\delta\right), J_\gamma(\theta|\delta)-\theta_\delta\rangle . \quad (31)$$

(30)-(31) together with the expression (11) of h_γ from the main paper imply that

$$\begin{aligned} h_\gamma(\theta|\delta) \geq h(\theta_\delta|\delta) - \langle \nabla\ell(\theta) - \nabla\ell(\theta_\delta), \theta-\theta_\delta\rangle \\ + \langle \nabla\ell(\theta) + g\left(\theta_\delta|\delta\right), J_\gamma(\theta|\delta)-\theta_\delta\rangle + \frac{1}{2\gamma}\|\theta - J_\gamma(\theta|\delta)\|^2. \end{aligned}$$

Since $J_\gamma(\theta|\delta) \in \mathbb{R}^d_\delta$, we can split $\|\theta - J_\gamma(\theta|\delta)\|^2$ as $\|\theta-\theta_\delta\|^2 + \|\theta_\delta - J_\gamma(\theta|\delta)\|^2$. We use this in the last inequality to conclude that

$$\begin{aligned} h_\gamma(\theta|\delta) &\geq h(\theta_\delta|\delta) + \frac{1}{2\gamma}\|\theta-\theta_\delta\|^2 - \langle \nabla\ell(\theta) - \nabla\ell(\theta_\delta), \theta-\theta_\delta\rangle \\ &\quad + \langle \nabla\ell(\theta) + g\left(\theta_\delta|\delta\right), J_\gamma(\theta|\delta)-\theta_\delta\rangle + \frac{1}{2\gamma}\|J_\gamma(\theta|\delta)-\theta_\delta\|^2 \\ &\geq h(\theta_\delta|\delta) + \frac{1}{2\gamma}\|\theta-\theta_\delta\|^2 - \langle \nabla\ell(\theta) - \nabla\ell(\theta_\delta), \theta-\theta_\delta\rangle \\ &\quad -\frac{\gamma}{2}\|\delta\cdot\nabla\ell(\theta) + \delta\cdot g\left(\theta_\delta|\delta\right)\|^2, \end{aligned}$$

as claimed. In the last inequality, the δ appearing in front of $\nabla\ell(\theta) + g(\theta|\delta)$ comes from the fact that $J_\gamma(\theta|\delta) - \theta_\delta \in \mathbb{R}^d_\delta$.

It is obvious from its definition that $h_\gamma(\theta|\delta)$ is non-decreasing as $\gamma \downarrow 0$. If $\theta \notin \mathbb{R}^d_\delta$, then $\|\theta - \theta \cdot \delta\| > 0$, and then both extreme sides of (29) converges to $+\infty = h(\theta|\delta)$ as $\gamma \downarrow 0$. If $\theta \in \mathbb{R}^d_\delta$, then $\|\theta - \theta \cdot \delta\| = 0$ and both extreme sides of (29) converges to $h(\theta \cdot \delta|\delta) = h(\theta|\delta)$ as $\gamma \downarrow 0$. □

Lemma 11. *Assume H1. Suppose that there exists $\gamma_0 > 0$ such that $\check{\Pi}_{\gamma_0}(\cdot|z)$ is well-defined. Then for all $\gamma \in (0, \gamma_0]$, $\check{\Pi}_\gamma(\cdot|z)$ is well-defined and*

$$\mathsf{d}_{\mathrm{tv}}(\check{\Pi}_\gamma, \tilde{\Pi}_\gamma) \leq 2\left(1 - e^{-\varrho_\gamma(z)}\right), \tag{32}$$

where $\varrho_\gamma(z)$ is as defined in (16).

Proof. For all $\gamma > 0$, we define

$$C_\gamma(\delta) \overset{\text{def}}{=} \int_{\mathbb{R}^d} e^{-h_\gamma(\theta|\delta)}\mathrm{d}\theta, \quad \text{and} \quad C_\gamma = \sum_\delta \pi_\delta (2\pi\gamma)^{\frac{\|\delta\|_0}{2}} C_\gamma(\delta).$$

The term C_γ is the normalizing constant of $\check{\Pi}_\gamma$. The function h_γ is nondecreasing as $\gamma \downarrow 0$. Hence, if $C_{\gamma_0} < \infty$, then $C_\gamma < \infty$ for all $\gamma \in (0, \gamma_0]$, which guarantees that $\check{\Pi}_\gamma$ is well-defined for all $\gamma \in (0, \gamma_0]$. For the remaining of the proof, we fix $\gamma \in (0, \gamma_0]$. To derive the total variation majoration, we start with a bound on C_γ. Using the second inequality of (29), we write

$$\begin{aligned}(2\pi\gamma)^{-d/2} C_\gamma &= \sum_\delta \pi_\delta \left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_0}{2}} \int_{\mathbb{R}^d} e^{-h_\gamma(\theta|\delta)}\mathrm{d}\theta \\ &\leq \sum_\delta \pi_\delta \left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_0}{2}} \int_{\mathbb{R}^d} e^{r_\gamma(\delta,\theta)} e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2} e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta.\end{aligned}$$

where

$$r_\gamma(\delta, \theta) = \langle \nabla\ell(\theta) - \nabla\ell(\theta_\delta), \theta - \theta_\delta)\rangle + \frac{\gamma}{2}\|\delta \cdot \nabla\ell(\theta) + \delta \cdot g\left(\theta_\delta|\delta\right)\|^2.$$

We recall from the proof of Lemma 9 that the normalizing constant of $\tilde{\Pi}_\gamma$ is given by

$$C = \sum_{\delta\in\Delta} \pi_\delta \left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_1}{2}} \int_{\mathbb{R}^d} e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2} e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta = \sum_{\delta\in\Delta} \pi_\delta \int_{\mathbb{R}^d} e^{-h(\theta|\delta)}\mu_\delta(\mathrm{d}\theta).$$

In view of the last inequality, and the definitions of $\tilde{\Pi}_\gamma$, C, and ϱ_γ, we get

$$\frac{(2\pi\gamma)^{-d/2} C_\gamma}{C} \leq e^{\varrho_\gamma(z)}. \tag{33}$$

The total variation bound between $\tilde{\Pi}_\gamma(\delta, \mathrm{d}\theta|z)$ and $\check{\Pi}_\gamma(\delta, \mathrm{d}\theta|z)$ now follows from a comparison of the two measures. Indeed, Using the first inequality of (29), and for

$\gamma \in (0, \gamma_0]$, we deduce that

$$\begin{aligned} \check{\Pi}_\gamma(\delta, \mathrm{d}\theta|z) &= \frac{1}{C_\gamma}\pi_\delta \left(\frac{1}{2\pi\gamma}\right)^{-\frac{\|\delta\|_0}{2}} e^{-h_\gamma(\theta|\delta)}\mathrm{d}\theta \\ &\geq \frac{1}{C_\gamma}\pi_\delta \left(\frac{1}{2\pi\gamma}\right)^{-\frac{\|\delta\|_0}{2}} e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2} e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta \\ &= \frac{C}{(2\pi\gamma)^{-\frac{d}{2}}\, C_\gamma}\tilde{\Pi}_\gamma(\delta, \mathrm{d}\theta|z) \\ &\geq e^{-\varrho_\gamma(z)}\tilde{\Pi}_\gamma(\delta, \mathrm{d}\theta|z), \end{aligned} \tag{34}$$

using (33). By a standard coupling argument (see e.g. Lindvall (1992) Equation 5.1), the minorization (34) implies (32). □

7.3. **Proof of Corollary 8.** The function ℓ is clearly convex and $\nabla\ell(\theta) = -\frac{1}{\sigma^2}X'(z-X\theta)$. Hence H1(1) holds. The elastic-net density in (18) is log-concave and continuous, which implies that $P(\cdot|\delta)$ is convex and lower semi-continuous for any given δ. Furthermore, For $\theta \in \mathbb{R}^d_\delta$, $\mathsf{sign}(\theta)$ is a subgradient of $x \mapsto \|x\|_1$ at θ. Hence $g(\theta|\delta) \overset{\text{def}}{=} \frac{\alpha\lambda_1}{\sigma^2}\mathsf{sign}(\theta) + \frac{(1-\alpha)\lambda_2}{\sigma^2}\theta$ is a subgradient of $P(\cdot|\delta)$ at $\theta \in \mathbb{R}^d_\delta$. Hence H1 holds, and the conclusion of Theorem 7 applies. From its definition, we have

$$e^{\varrho_\gamma(z)} = \frac{\sum_{\delta\in\Delta}\pi_\delta\left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_0}{2}}\int_{\mathbb{R}^d} e^{r_\gamma(\delta,\theta)}e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2}e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta}{\sum_{\delta\in\Delta}\left(\frac{1}{2\pi\gamma}\right)^{\frac{d-\|\delta\|_0}{2}}\int_{\mathbb{R}^d} e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2}e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta}.$$

From the expression of $\nabla\ell$, we have

$$\|\nabla\ell(\theta_2) - \nabla\ell(\theta_1)\| \leq L_1\|\theta - \theta_2\|. \tag{35}$$

with $L_1 \overset{\text{def}}{=} \lambda_{\mathsf{max}}(X'X)/\sigma^2$. Furthermore, for all $\delta \in \Delta$ and $\theta \in \mathbb{R}^d_\delta$,

$$\|\delta\cdot\nabla\ell(\theta)\|^2 = \frac{1}{\sigma^4}(z-X\theta)'X_\delta X_\delta'(z-X\theta) \leq 2L_1\frac{1}{2\sigma^2}\|z-X\theta\|^2 = 2L_1\ell(\theta). \tag{36}$$

From the expression of $g(\cdot|\delta)$, we have

$$\begin{aligned} \|g(\theta|\delta)\|^2 &\leq \left(\frac{\alpha\lambda_1}{\sigma^2}\right)^2\|\delta\|_0 + \frac{2(1-\alpha)\lambda_2}{\sigma^2}\left[\alpha\frac{\lambda_1}{\sigma^2}\|\theta\|_1 + (1-\alpha)\frac{\lambda_2}{2\sigma^2}\|\theta\|^2\right] \\ &\leq c(\delta) + 2L_1P(\theta|\delta),\ \theta\in\mathbb{R}^d_\delta \end{aligned} \tag{37}$$

where $c(\delta) \overset{\text{def}}{=} \left(\frac{\alpha\lambda_1}{\sigma^2}\right)^2\|\delta\|_0$, and using the assumption $(1-\alpha)\lambda_2 \leq \lambda_{\mathsf{max}}(X'X)$. Using (35-37), we have

$$r_\gamma(\delta,\theta) \leq L_1\left(1+\frac{3\gamma}{2}L_1\right)\|\theta-\theta_\delta\|^2 + 3\gamma L_1\ell(\theta_\delta) + \frac{3\gamma}{2}c(\delta) + 3\gamma L_1P(\theta_\delta|\delta). \tag{38}$$

We set $h_\gamma \overset{\text{def}}{=} 1-2\gamma L_1\left(1+\frac{3\gamma}{2}L_1\right)$, and $a \overset{\text{def}}{=} 3L_1$. Then (38) gives

$$\int_{\mathbb{R}^d} e^{r_\delta(\delta,\theta)}e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2}e^{h(\theta_\delta|\delta)}\mathrm{d}\theta \leq e^{\frac{3\gamma}{2}c(\delta)} \times \int_{\mathbb{R}^d} e^{-\frac{h_\gamma}{2\gamma}\|\theta-\theta_\delta\|^2}e^{-(1-\gamma a)\ell(\theta_\delta)-(1-\gamma a)P(\theta_\delta|\delta)}\mathrm{d}\theta. \quad (39)$$

Notice that the integral on the right-side of (39) can be factorized as the product of two integrals, with one integral taken over the components for which $\delta_j = 0$, and the other taken over the components for which $\delta_j = 1$. We introduce some notation to do this rigorously. Fix $\delta \in \Delta$, and $s = \|\delta\|_0$. For a given function $f : \mathbb{R}^d \to \mathbb{R}$, we define $f^{[s]} : \mathbb{R}^s \to \mathbb{R}$ as $f^{[s]}(u) = f(u^\delta)$, where $u^\delta \in \mathbb{R}^d$, and $u_i^\delta = 0$ if $\delta_i = 0$, and $u_j^\delta = u_{\sum_{k=1}^j \delta_k}$ if $\delta_j = 1$. With this notation, and for $4\gamma L_1 \leq 1$ (which implies that $h_\gamma > 0$), the integral on the right-hand side of (39) is equal to

$$\left(\frac{2\pi\gamma}{h_\gamma}\right)^{\frac{d-s}{2}} \int_{\mathbb{R}^s} e^{-(1-\gamma a)\ell^{[s]}(u)-(1-\gamma a)P^{[s]}(u|\delta)}\mathrm{d}u.$$

A similar calculation on the denominator of $e^{\varrho_\gamma(z)}$ gives

$$\int_{\mathbb{R}^d} e^{-\frac{1}{2\gamma}\|\theta-\theta_\delta\|^2}e^{-h(\theta_\delta|\delta)}\mathrm{d}\theta = (2\pi\gamma)^{\frac{d-s}{2}}\int_{\mathbb{R}^s} e^{-\ell^{[s]}(u)-P^{[s]}(u|\delta)}\mathrm{d}u.$$

We conclude that

$$e^{\varrho_\gamma(z)} \leq \frac{\sum_{\delta\in\Delta}\pi_\delta e^{\frac{3\gamma}{2}c(\delta)}\left(\frac{1}{h_\gamma}\right)^{\frac{d-s}{2}}\int_{\mathbb{R}^s} e^{-(1-\gamma a)\ell^{[s]}(u)-(1-\gamma a)P^{[s]}(u|\delta)}\mathrm{d}u}{\sum_{\delta\in\Delta}\pi_\delta\int_{\mathbb{R}^s} e^{-\ell^{[s]}(u)-P^{[s]}(u|\delta)}\mathrm{d}u}, \quad (40)$$

For $4\gamma L_1 \leq 1$, and using the inequality $\log(1-2x-3x^2) \geq -6x$, valid for all $x \in [0, 1/4]$, we have

$$\left(\frac{1}{h_\gamma}\right)^{\frac{d-s}{2}} = \exp\left[-\frac{d-s}{2}\log\left(1-2\gamma L_1-3\gamma^2L_1^2\right)\right] \leq e^{3d\gamma L_1}. \quad (41)$$

Fix $u_0 \in \mathbb{R}^s$, arbitrary. Since $\gamma > 0$ is taken such that $4\gamma L_1 \leq 1$, we see that $\gamma a = 3\gamma L_1 \leq 3/4$. Then by the convexity of $\ell^{[s]}$ we have

$$\begin{aligned}(1-\gamma a)\ell^{[s]}(u) &= -\gamma a\ell^{[s]}(u_0) + (1-\gamma a)\ell^{[s]}(u) + \gamma a\ell^{[s]}(u_0)\\ &\geq -\gamma a\ell^{[s]}(u_0) + \ell^{[s]}\left(\gamma a u_0 + (1-\gamma a)u\right).\end{aligned}$$

Similarly, by the convexity of $P^{[s]}(\cdot|\delta)$,

$$(1-\gamma a)P^{[s]}(u|\delta) \geq -\gamma a P^{[s]}(u_0|\delta) + P^{[s]}\left(\gamma a u_0 + (1-\gamma a)u|\delta\right).$$

Using these last two inequalities, and the change of variable $(1-\gamma a)u + \gamma a u_0 = w$, we conclude that

$$\int_{\mathbb{R}^s} e^{-(1-\gamma a)\ell^{[s]}(u)-(1-\gamma a)P^{[s]}(u|\delta)}\mathrm{d}u \\ \le e^{\gamma a\left(\ell^{[s]}(u_0)+P^{[s]}(u_0|\delta)\right)}\,(1-\gamma a)^{-s}\int_{\mathbb{R}^s} e^{-\ell^{[s]}(u)-P^{[s]}(u|\delta)}\mathrm{d}u.$$

Setting $\mathcal{R}(z) \overset{\text{def}}{=} \max_{\delta\in\Delta}\inf_{u\in\mathbb{R}^s}\left[\ell^{[s]}(u)+P^{[s]}(u|\delta)\right]$, and using the inequality $\log(1-3x)\ge -6x$, $x\in[0,1/4]$ we obtain,

$$\int_{\mathbb{R}^s} e^{-(1-\gamma a_1)\bar{\ell}(u)-(1-\gamma a_2)\bar{P}(u)}\mathrm{d}u \le e^{\gamma a\mathcal{R}(z)}e^{6d\gamma L_2}\int_{\mathbb{R}^s} e^{-\bar{\ell}^{(s)}(u)-\bar{P}^{(s)}(u|\delta)}\mathrm{d}u.$$

It follows from this last inequality, (41) and (40) that

$$\varrho_\gamma(z) \le \frac{3\gamma}{2}\max_{\delta\in\Delta} c(\delta) + 3\gamma L_1\left(3d+\mathcal{R}(z)\right) \\ \le \frac{3\gamma}{2}\left(\frac{\alpha\lambda_1}{\sigma^2}\right)^2 d + \frac{3\gamma}{\sigma^2}\lambda_{\mathsf{max}}(X'X)\left[d\left(3+\log Z(\phi)\right)+\frac{\|z\|^2}{2\sigma^2}\right],$$

as claimed.

□

References

Armagan, A., Dunson, D. B. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23** 119–143.

Atchade, Y. and Bhattacharyya, A. (2018). Regularization and Computation with high-dimensional spike-and-slab posterior distributions. *ArXiv e-prints* .

Atchadé, Y., Fort, G., Moulines, E. and Priouret, P. (2011). Adaptive Markov chain Monte Carlo: theory and methods. In *Bayesian time series models*. Cambridge Univ. Press, Cambridge, 32–51.

Atchade, Y. A. (2017). On the contraction properties of some high-dimensional quasi-posterior distributions. *Ann. Statist.* **45** 2248–2273.

Atchadé, Y. F. (2006). An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodol Comput Appl Probab* **8** 235–254.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York. With a foreword by Hédy Attouch. URL `http://dx.doi.org/10.1007/978-1-4419-9467-7`

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

Buhlmann, P. and Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics* **29** 407–430.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018.

Chen, X., Wang, Z. J. and McKeown, M. J. (2011). A bayesian lasso via reversible-jump {MCMC}. *Signal Processing* **91** 1920 – 1932.

Dudley, R. (2002). *Real Analysis and Probability.* Cambridge Series in advanced mathematics, Cambridge University Press, NY.

Ge, D., Idier, J. and Carpentier, E. L. (2011). Enhanced sampling schemes for MCMC based blind bernoulli-gaussian deconvolution. *Signal Processing* **91** 759 – 772.

George, E. I. and McCulloch, R. E. (1997). Approaches to bayesian variable selection. *Statist. Sinica* **7** 339–373.

Gottardo, R. and Raftery, A. E. (2008). Markov chain monte carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics* **17** 949–975.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Statist.* **33** 730–773.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Anal.* **5** 151–170.

Lindvall, T. (1992). *Lectures on the coupling method.* John Wiley & Sons, Inc., New York.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *JASA* **83** 1023–1032.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93** 273–299.

Narisetty, N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817.

Ormerod, J. T., You, C. and Muller, S. (2014). A variational Bayes approach to variable selection. Tech. rep., Preprint.

Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.

Patrinos, P., Stella, L. and Bemporad, A. (2014). Forward-backward truncated Newton methods for convex composite optimization. *ArXiv e-prints* .

Pereyra, M. (2013). Proximal Markov chain Monte Carlo algorithms. *ArXiv e-prints* .

Reid, S., Tibshirani, R. and Friedman, J. (2013). A Study of Error Variance Estimation in Lasso Regression. *ArXiv e-prints* .

Rockova, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* **109** 828–846.

Schreck, A., Fort, G., Le Corff, S. and Moulines, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *ArXiv e-prints* .

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B* **57** 749–760.

Yang, Y., Wainwright, M. J. and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *Ann. Statist.* **44** 2497–2532.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320.