A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection

Amandine Schreck, Gersende Fort, Sylvain Le Corff, Eric Moulines

A Shrinkage-Thresholding Metropolis adjusted Langevin algorithm for Bayesian Variable Selection

Amandine Schreck†

LTCI, Télécom ParisTech and CNRS, Paris, France.

Gersende Fort

LTCI, CNRS and Télécom ParisTech, Paris, France.

Sylvain Le Corff

Department of Statistics, University of Warwick, Coventry, UK.

Éric Moulines

LTCI, Télécom ParisTech and CNRS, Paris, France.

Summary. This paper introduces a new Markov Chain Monte Carlo method to perform Bayesian variable selection in high dimensional settings. The algorithm is a Hastings-Metropolis sampler with a proposal mechanism which combines *(i)* a Metropolis adjusted Langevin step to propose local moves associated with the differentiable part of the target density with *(ii)* a shrinkage-thresholding step based on the non-differentiable part of the target density which provides sparse solutions such that small components are shrunk toward zero. This allows to sample from distributions on spaces with different dimensions by actually setting some components to zero. The performances of this new procedure are illustrated with both simulated and real data sets. The geometric ergodicity of this new transdimensional Markov Chain Monte Carlo sampler is also established.

Keywords: Markov Chain Monte Carlo, Proximal operators, MALA, Bayesian variable selection, Sparsity

1. Introduction

This paper considers the long-standing problem of Bayesian variable selection in a linear regression model. Variable selection is a complicated task in high dimensional settings

†*adress for correspondance:* Amandine Schreck, Télécom ParisTech, 37-39 rue Dareau, 75014 Paris. Email: amandine.schreck@telecom-paristech.fr

where the number of regression parameters P is much larger than the number of observations N. In this context, it is crucial to introduce sparsity assumptions based on the prior knowledge that only a few number of regression parameters are significant. Using a sequence of observations from a linear regression model, the aims are *(i)* to determine which components of the regression vector are active and explain the observations and *(ii)* to estimate the regression vector.

Many methods have been proposed to perform variable selection, see O'Hara & Sillanpää (2009) for a review of Bayesian methods. Among the most popular are the penalized least squares estimators with a L_1-norm penalization introduced by Tibshirani (1996), also known as the Least Absolute Shrinkage and Selection Operator (LASSO); see also *e.g.* Bickel et al. (2009), Van de Geer (2009), Bunea et al. (2007) and the references therein.

In a Bayesian framework, many approaches introduce prior distributions both on the regression vector and on some hyper-parameters. These methods consist then in using a Gibbs sampler to draw alternatively the regression vector and each hyper-parameter in order to explore the joint posterior distribution and to find regression vectors with high posterior probability, see West (2003), Tan et al. (2010), Griffin & Brown (2011). These algorithms yield approximately sparse estimations of the regression vector which means that many components are close but not equal to zero. For example, Casella & Park (2008) proposed a Bayesian LASSO by interpreting the L_1 penalization as a double exponential prior which shrinks regression parameters toward zero.

The Bayesian spike and slab models have been first proposed by Beauchamp & Mitchell (1988). The model used in Beauchamp & Mitchell (1988) introduces independent priors for the regression parameters which are mixtures between a uniform flat distribution (the slab) and a Dirac distribution at zero (the spike), yielding exactly sparse estimators. This model is extended in George & McCulloch (1993) which uses a binary latent variable to locate the regression parameters that explain the observations. A normal distribution with high variance (the slab) is then associated with these parameters, and a normal distribution with very small variance (the spike) is associated with the other regression parameters. In Ishwaran & Rao (2005), the scale of the mixture components is set through a prior distribution on the hyper-variance. See Malsiner-Walli & Wagner (2011) for a comparison of the different spike and slab priors which can be used, and particularly for a study of the differences between the priors with a Dirac spike and those with a Gaussian spike. These spike and slab models based on a (non-degenerated) Gaussian spike provide better results for high dimensional regression settings. However, they do not allow to actually set to zero

some regression parameters.

In other Bayesian variable selection approaches, the prior distribution of the regression vector makes it difficult to sample from the corresponding posterior distribution. Therefore, Markov Chain Monte Carlo (MCMC) methods such as random walk Hastings Metropolis algorithms, block-Gibbs samplers which update a subset of regression parameters at each iteration, or Metropolis adjusted Langevin algorithm (MALA), have been widely used. In some cases, the prior distribution of the regression vector uses a penalization function yielding to approximately sparse samples (see Wipf et al. (2011) for a theoretical and empirical comparison of different penalization functions). Lucka (2012) uses a L_1-penalization and compares results obtained by different MCMC samplers. Dalalyan & Tsybakov (2012) introduces a smooth penalization function to obtain a differentiable posterior distribution which allows to use MALA. In the case of a non-differentiable penalized posterior distribution, Pereyra (2013) combines MALA with a proximal operator.

Other MCMC approaches for Bayesian variable selection define a posterior distribution on the model space, where a model is a binary vector locating the active (non null) components of the regression vector. The objective is then to explore this posterior distribution, which is equivalent to estimate probabilities of activation for each regression parameter. In Brown et al. (2001) for example, this exploration is performed with a Gibbs sampler. Variants and adaptive versions of the Gibbs sampler for this problem have been proposed in Nott & Kohn (2005), Lamnisos et al. (2013). Samples from the posterior distribution of the models are obtained in Shi & Dunson (2011) and in Schäfer & Chopin (2013) with particle filters. These methods are extended in Rigollet & Tsybakov (2012) to obtain estimators of the regression vector using the mean square estimators associated with each model.

The last class of MCMC methods for Bayesian variable selection is designed to obtain exactly sparse samples (with some components actually set to zero). These methods jointly sample a model and the regression parameters active in this model, see *e.g.* Dellaportas et al. (2002) and the references therein. The Reversible Jump MCMC (RJMCMC) is a popular algorithm introduced in Green (1995) which produces a Markov chain evolving between spaces of different dimensions. The dimension of the sample varies at any iteration as active parameters are added or discarded from the model. Each new sample is accepted or rejected using a Metropolis-Hastings step where the acceptance probability is adjusted to the transdimensional moves. See also Brooks et al. (2003), Karagiannis & Andrieu (2013) for efficient ways to implement RJMCMC. Carlin & Chib (1995) consider another

setting that encompasses all the models jointly: at each iteration, pseudo-prior distributions are used to jointly sample regression parameters associated with all models. For high dimensional statistical problems, such a joint sampling is computationally too expensive. A more efficient algorithm is proposed in Dellaportas et al. (2002) which only involves the simulation of a new model and of the regression parameters corresponding to this newly sampled model. This method, called Metropolized Carlin and Chib (MCC), avoids the need to sample from all the pseudo-priors and can be implemented in practice, see also Petralias & Dellaportas (2013). Ji & Schmidler (2013) use an adaptive Metropolized algorithm to sample from a posterior distribution which is a mixture of a Dirac at zero and slab distributions. This algorithm samples independently each regression parameter according to an adaptive mixture of a Dirac at zero and a Gaussian distribution. It is therefore not appropriate for high dimensional settings as the proposal strategy does not take into account the target distribution.

In this paper, we introduce a new MCMC algorithm, called Shrinkage-Thresholding MALA (STMALA), designed to sample sparse regression vectors by jointly sampling a model and a regression vector in this model. This algorithm, which is a transdimensional MCMC method, relies on MALA (see Roberts & Tweedie (1996)). The proposal distribution of MALA is based on the computation of the gradient of the logarithm of the target distribution. In order to both deal with a non-differentiable target posterior distribution and to actually set some components to zero, we propose to combine MALA with a shrinkage-thresholding operator by:

- computing a noisy gradient step involving the term of the logarithm of the target distribution which is continuously differentiable;
- then applying a shrinkage-thresholding operator to ensure sparsity and shrink small values of the regression parameters toward zero.

Such an algorithm is motivated by Bayesian variable selection with non-smooth priors. This algorithm can perform global moves from one model to a rather distant other one, which allows to explore efficiently high dimensional spaces (in comparison to local move algorithms). The geometric ergodicity of this new algorithm is proved for a large class of target distributions. To our knowledge, it is the first result providing a rate of convergence for a transdimensional MCMC algorithm (like RJMCMC and MCC); usually, only Harris recurrence is proved, see Roberts & Rosenthal (2006).

This paper is organized as follows. STMALA and its application to Bayesian variable selection is described in Section 2. Different implementations are proposed in Section 3.

The geometric ergodicity of this new sampler is addressed in Section 4. Numerical experiments on simulated and real data sets to assess the performance of STMALA are given in Section 5. Finally, all the proofs are postponed to Section 6.

2. The Shrinkage-Thresholding MALA algorithm

This section introduces the Shrinkage-Thresholding MALA (STMALA) algorithm which is designed to sample from a target distribution defined on $\mathbb{R}^{P\times T}$, where $P, T \in \mathbb{N}^*$, under the sparsity assumption that a large number of rows of each sample should be null.

Let $\mathcal{M} \stackrel{\text{def}}{=} \{0,1\}^P$ be the set of binary vectors locating the non-zero rows of elements of $\mathbb{R}^{P\times T}$. For any $m = (m_1, \ldots, m_P) \in \mathcal{M}$, set

$$I_m \stackrel{\text{def}}{=} \{i \in \{1, \cdots, P\};\ m_i = 1\}\ . \tag{1}$$

We consider target distributions on $\mathbb{R}^{P\times T}$ absolutely continuous with respect to the positive measure $\mathrm{d}\nu(x)$ given by

$$\mathrm{d}\nu(x) \stackrel{\text{def}}{=} \sum_{m\in\mathcal{M}} \left(\prod_{i\notin I_m} \delta_0(\mathrm{d}x_{i\cdot})\right) \left(\prod_{i\in I_m} \mathrm{d}x_{i\cdot}\right)\ , \tag{2}$$

where, for $x \in \mathbb{R}^{P\times T}$ and $1 \le i \le P$, $x_{i\cdot}$ is the i-th row of x. The STMALA algorithm is based on the MALA algorithm which proposes local moves using information about the gradient of the logarithm of the target density (when it is differentiable). Nevertheless, MALA is not designed to produce sparse samples. Therefore, we propose to combine a gradient step as in MALA with a shrinkage-thresholding step, which produces sparse matrices of $\mathbb{R}^{P\times T}$. This mechanism is followed by an accept-reject step to guarantee the convergence to the right target distribution. Before describing the algorithm, we introduce some notations;

Notations For any matrix $A \in \mathbb{R}^{\ell\times\ell'}$, A_{ij} denotes the entry (i,j) of the matrix A and $A_{i\cdot}$ is the i-th row of A. For any $m = (m_1, \cdots, m_P) \in \{0,1\}^\ell$, let $|m| \stackrel{\text{def}}{=} \sum_{i=1}^P m_i$ denotes the number of positive entries. $A_{m\cdot}$ denotes the $|m| \times \ell'$ matrix obtained by extracting from A the active components in m. Similarly, $A_{\cdot m}$ for $m \in \{0,1\}^{\ell'}$ collects the columns of A indexed by the active components in m. By convention, if m in $\{0,1\}^l$ (resp. $\{0,1\}^{l'}$) is such that $|m| = 0$, then $A_{m\cdot} = 0$ (resp. $A_{\cdot m} = 0$). $A_{-m\cdot}$ denotes the $(P - |m|) \times T$ matrix obtained by extracting from A the rows indexed by $i \notin I_m$. Define the Frobenius norm

$\|\cdot\|_2$, the $L_{2,1}$-norm $\|\cdot\|_{2,1}$ and the 1-norm $\|\cdot\|_1$ of a $\ell' \times \ell$-matrix A as

$$\|A\|_2 \stackrel{\text{def}}{=} \left(\sum_{i=1}^{\ell}\sum_{i=1}^{\ell'} A_{i,j}^2\right)^{1/2} , \quad \|A\|_{2,1} \stackrel{\text{def}}{=} \sum_{i=1}^{\ell}\left(\sum_{j=1}^{\ell'} A_{i,j}^2\right)^{1/2} \quad \text{and} \quad \|A\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{\ell}\sum_{j=1}^{\ell'} |A_{i,j}| .$$

2.1. The STMALA algorithm

It is assumed that

A1 $\pi \, \mathrm{d}\nu$ is the target distribution: $\sup_{\mathbb{R}^{P\times T}} \pi < \infty$ and there exists a continuously differentiable function $g : \mathbb{R}^{P\times T} \to \mathbb{R}$ and a measurable function $\bar{g} : \mathbb{R}^{P\times T} \to \mathbb{R}$ such that

$$\pi(x) \propto \exp\left\{-g(x) - \bar{g}(x)\right\} .$$

Let $\sigma > 0$ be a fixed stepsize and $\Psi : \mathbb{R}^{P\times T} \to \mathbb{R}^{P\times T}$ be a shrinkage-thresholding operator. In this paper, three different operators are considered to sample sparse matrices (see Section 3 for further comments on these operators): for any $\gamma > 0$, any $1 \le i \le P$ and any $1 \le j \le T$,

$$\begin{aligned}
(\Psi_1(u))_{i,j} &= u_{i,j}\left(1 - \frac{\gamma}{\|u_{i\cdot}\|_2}\right)_+ , \\
(\Psi_2(u))_{i,j} &= u_{i,j}\mathbf{1}_{\|u_{i\cdot}\|_2 > \gamma} , \\
(\Psi_3(u))_{i,j} &= u_{i,j}\left(1 - \frac{\gamma^2}{\|u_{i\cdot}\|_2^2}\right)_+ ,
\end{aligned}$$

where for $a \in \mathbb{R}$, a_+ denotes the positive part of a: $a_+ \stackrel{\text{def}}{=} \max(a, 0)$.

From a current state X^n the algorithm proposes a new point Z according to a proposal distribution $q_\Psi(X^n, \cdot)$ which can be seen as a noisy proximal gradient step: given the current value of the chain X^n, the candidate Z is defined by

$$Z = \Psi\left(X^n - \frac{\sigma^2}{2}\nabla g(X^n) + \sigma\Xi^{n+1}\right) , \tag{3}$$

where Ξ^{n+1} is a $\mathbb{R}^{P\times T}$ random matrix with independent and identically distributed (i.i.d.) standard Gaussian entries. This candidate is then accepted or not with an accept-reject step. If Ψ is the identity operator, then the candidate becomes

$$Z = X^n - \frac{\sigma^2}{2}\nabla g(X^n) + \sigma\Xi^{n+1} ,$$

which is the proposal mechanism of MALA.

STMALA is outlined in Algorithm 1. It produces a sequence $(X^n)_{n\in\mathbb{N}}$ which is a Hastings-Metropolis Markov chain with proposal distribution q_Ψ and target distribution

$\pi \, \mathrm{d}\nu$. The expression of the transition density q_Ψ is established in Section 3 for different shrinkage-thresholding operators Ψ. We will also interpret these proposal distributions as mechanisms to sample both a binary vector $m \in \{0,1\}^P$ and a $|m| \times T$ matrix. To motivate

Algorithm 1 One iteration of the STMALA algorithm given X^n

1: Draw a $P \times T$ matrix Ξ^{n+1} with i.i.d. entries sampled from $\mathcal{N}(0,1)$.
2: Set $Z = \Psi\left(X^n - \frac{\sigma^2}{2}\nabla g(X^n) + \sigma\Xi^{n+1}\right)$.
3: Set
$$\alpha(X^n, Z) = 1 \wedge \frac{\pi(Z)\, q_\Psi(Z, X^n)}{\pi(X^n)\, q_\Psi(X^n, Z)} \, .$$
4: Draw $U \sim U(0,1)$.
5: **if** $U \leq \alpha(X^n, Z)$ **then**
6: $\quad X^{n+1} = Z$.
7: **else**
8: $\quad X^{n+1} = X^n$.
9: **end if**

this framework, let us consider the Bayesian variable selection problem.

2.2. Application to Bayesian variable selection

Let $Y \in \mathbb{R}^{N \times T}$ be the observations modeled as

$$Y = GX + \sqrt{\tau}E \, , \tag{4}$$

where $G \in \mathbb{R}^{N \times P}$ is a known gain matrix, $X \in \mathbb{R}^{P \times T}$ is the unknown regression matrix, and $E \in \mathbb{R}^{N \times T}$ is a noise matrix. It is assumed that the entries $E_{i,j}$ for $1 \leq i \leq N$ and $1 \leq j \leq T$ are i.i.d. according to $\mathcal{N}(0,1)$. This Gaussian linear regression model appears in many different situations in modern applied statistics such as genomics or Magnetic Resonance Imaging (MRI) studies. For example, in the case of MRI, a stimulus is delivered to a patient and his brain activity is measured by magnetoencephalography and electroencephalography. The objective is then to retrieve the original signal X (source amplitudes) using the measured signal Y. By Maxwell's equations, the signal measured by the sensors is a linear combination of the electromagnetic fields produced by all the sources. In this case, N is the number of sensors, T is the number of measurement times, P is the number of sources and G is the gain matrix modeling the electromagnetic properties of the brain.

In high dimensional variable selection problems, the regression vector X has to be recovered under sparsity constraints. A sparse signal X can equivalently be defined by *(i)* a binary vector $m = (m_1, \cdots, m_P) \in \{0,1\}^P$ with the convention that $m_k = 1$ if and only

if $X_{k\cdot}$ is active i.e. is non null; and *(ii)* the matrix $X_{m\cdot}$ which collects the $|m|$ active rows of X. Hence, $m \in \mathcal{M}$ is a model, $|m|$ is the number of active rows, and I_m given by (1) is the set of indices corresponding to active rows.

An algorithm to sample sparse matrices can be described as a sampler for the exploration of a posterior distribution absolutely continuous with respect to $\mathrm{d}\nu$ defined in (2). For any model $m \in \mathcal{M}$, denote by S_m the subset of $\mathbb{R}^{P\times T}$ associated with m, defined by

$$S_m \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{P\times T}, z_{i\cdot} \neq 0 \ \forall i \in I_m \text{ and } z_{-m\cdot} = 0\} \ . \tag{5}$$

Then $(S_m)_{m\in\mathcal{M}}$ is a partition of $\mathbb{R}^{P\times T}$ with non-null ν-measure. Sampling a distribution absolutely continuous with respect to $\mathrm{d}\nu$ on $\mathbb{R}^{P\times T}$ is equivalent to sampling a pair $(m, X_{m\cdot})$ where m is the set of the active rows and given m, $X_{m\cdot} \in \mathbb{R}^{|m|\times T}$ collects the value of these rows.

Under the statistical model (4), for $m \in \mathcal{M}$ and $X \in S_m$, the likelihood of the observation Y given X is

$$\pi(Y|X) \stackrel{\text{def}}{=} (2\pi\tau)^{-NT/2} \exp\left(-\frac{1}{\tau}\|Y - G_{\cdot m}X_{m\cdot}\|_2^2\right) = (2\pi\tau)^{-NT/2} \exp\left(-\frac{1}{\tau}\|Y - GX\|_2^2\right) \ .$$

The sparsity constraint is expressed by a joint prior distribution: $\pi(X_{m\cdot}|m)$ is a prior distribution on $\mathbb{R}^{|m|\times T}$ conditionally to the model m, and $(\omega_m)_{m\in\mathcal{M}}$ is a prior distribution on $\mathcal{M}$ that is $(\omega_m)_{m\in\mathcal{M}}$ is a non-negative sequence satisfying $\sum_{m\in\mathcal{M}} \omega_m = 1$. An example of prior distribution is a $L_{2,1}$-penalty on the regression vector:

$$\pi(X_{m\cdot}|m) \stackrel{\text{def}}{=} \exp(-\lambda\|X_{m\cdot}\|_{2,1} - |m| \ln c_\lambda) \ ,$$

with $\lambda \geq 0$ and (see Lemma 6.1)

$$c_\lambda \stackrel{\text{def}}{=} \begin{cases} 2\pi^{T/2}(T-1)!\lambda^{-T} \left(\Gamma\left(T/2\right)\right)^{-1} & \text{if } \lambda > 0 \ , \\ 1 & \text{if } \lambda = 0 \ , \end{cases} \tag{6}$$

where Γ is the standard Gamma function defined on $(0, +\infty)$ by $\Gamma : x \mapsto \int_0^{+\infty} t^{x-1}\mathrm{e}^{-t}\mathrm{d}t$. Therefore, the posterior density $\pi(X|Y)$ on $\mathbb{R}^{P\times T}$ is given by, for $m \in \mathcal{M}$ and $X \in S_m$

$$\pi(X|Y) \propto \omega_m \, c_\lambda^{-|m|} \exp\left(-\frac{1}{2\tau}\|Y - Gx\|_2^2 - \lambda\|x\|_{2,1}\right) \ . \tag{7}$$

In this application, the target density is $x \mapsto \pi(x|Y)$ and it is proportional to $\exp\{-g(x) - \bar{g}(x)\}$, with for any $m \in \mathcal{M}$ and $x \in S_m$,

$$g(x) = \frac{1}{2\tau}\|Y - Gx\|_2^2 \quad \text{and} \quad \bar{g}(x) = \lambda\|x\|_{2,1} - \log\left(w_m c_\lambda^{-|m|}\right) \ .$$

2.3. Partial updating

In high dimensional settings, STMALA may encounter some difficulties to accept the proposed moves. Following the idea introduced in Neal & Roberts (2006), we introduce in this section a variant of the algorithm in which only a fixed proportion of components of X^n are updated at each iteration n. This is achieved by combining STMALA and a Gibbs sampler in a STMALA-within-Gibbs algorithm, called block-STMALA.

This algorithm depends on a new parameter $\eta \in \{1, \cdots, P\}$ which specifies the number of rows to be updated at each iteration of the algorithm. Let η be fixed. Denote by $\mathcal{B}_\eta$ the set of subsets of $\{1, \ldots, P\}$ with exactly η elements. The first step consists in choosing at random a subset $b \in \mathcal{B}_\eta$. Then, given b, a STMALA algorithm is run with the conditional distribution of $x_{b\cdot}$ given the other components $x_{-b\cdot}$ under π, denoted by $\pi(x_{b\cdot}|x_{-b\cdot})$, as target distribution.

For $b \in \mathcal{B}_\eta$, denote by q_b the proposal transition density of this block-STMALA step, and by $\nabla_b g(x)$ the gradient of the function $x \mapsto g(x)$ with respect to $x_{b\cdot}$, *i.e.* $\nabla_b g(x) = (\nabla g(x))_{b\cdot}$. The block-STMALA algorithm is summarized in Algorithm 2 with a block size set to η. The transition kernel associated with the block-STMALA algorithm is given by

Algorithm 2 One iteration of the block-STMALA algorithm given $\boldsymbol{X}^n$

1: Select uniformly $b \in \mathcal{B}_\eta$.
2: Draw a $\eta \times T$ matrix Ξ^{n+1} with i.i.d. entries sampled from $\mathcal{N}(0,1)$.
3: Define Z: set $Z_{-b\cdot} = X^n_{-b\cdot}$ and $Z_{b\cdot} = \Psi\left(X^n_{b\cdot} - \frac{\sigma^2}{2}\nabla_b g(X^n) + \sigma\Xi^{n+1}\right)$.
4: Set

$$\alpha_b(X^n, Z) = 1 \wedge \frac{\pi(Z)\, q_b(Z_{b\cdot}, X^n_{b\cdot})}{\pi(X^n)\, q_b(X^n_{b\cdot}, Z_{b\cdot})} \,,$$

5: Draw $U \sim U(0,1)$.
6: **if** $U \leq \alpha_b(X^n, Z)$ **then**
7: $\quad X^{n+1} = Z$.
8: **else**
9: $\quad X^{n+1} = X^n$.
10: **end if**

$$P_{\text{block}} \stackrel{\text{def}}{=} \binom{P}{\eta}^{-1} \sum_{b \in \mathcal{B}_\eta} P_b \,,$$

where, for any $b \in \mathcal{B}_\eta$, P_b is given by

$$P_b(x, \mathrm{d}z) \stackrel{\text{def}}{=} \left(\prod_{i \notin b} \delta_{x_{i\cdot}}(\mathrm{d}z_{i\cdot})\right) \left(\alpha_b(x,z) q_b(x_{b\cdot}, \mathrm{d}z_{b\cdot}) + \delta_{x_{b\cdot}}(\mathrm{d}z_{b\cdot}) \int (1 - \alpha_b(x, \tilde{z})) q_b(x, \mathrm{d}\tilde{z})\right) \,.$$

Note that for any $b \in \mathcal{B}_\eta$, the target density π is invariant with respect to P_b, so that π is also invariant with respect to P_{block}.

3. Shrinkage-Thresholding operators for STMALA

We consider in turn three different shrinkage-thresholding operators Ψ. For each of them, we provide an explicit expression of the proposal distribution q_Ψ and show that this distribution is equivalent to *(i)* first sampling the indices of the active rows of the candidate matrix by sampling a binary vector $m \in \mathcal{M}$; and *(ii)* sampling a matrix in $\mathbb{R}^{|m|\times T}$. Figure 1 displays these operators in the case $P = T = 1$.

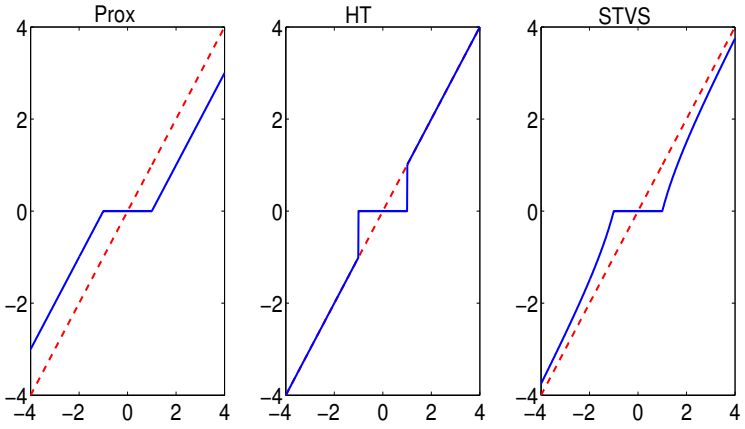

Fig. 1: Shrinkage-Thresholding functions associated with the $L_{2,1}$ proximal operator (Prox - left), the hard thresholding operator (HT - center) and the soft thresholding operator with vanishing shrinkage (STVS - right) in one dimension.

3.1. The $L_{2,1}$ proximal operator

A first idea is to consider the $L_{2,1}$ proximal operator $\Psi_1 : \mathbb{R}^{P\times T} \to \mathbb{R}^{P\times T}$ defined componentwise by

$$(\Psi_1(u))_{i,j} = u_{i,j}\left(1 - \frac{\gamma}{\|u_{i.}\|_2}\right)_+ , \tag{8}$$

for some (fixed) positive parameter γ. The function $u \mapsto \Psi_1(u)$ is displayed on Figure 1[left] in the case $P = T = 1$.

When g is a continuously differentiable convex function such that ∇g is L_g-Lipschitz, it is known (see e.g. (Beck & Teboulle 2009, theorem 3.1), or Parikh & Boyd (2013)) that the deterministic sequence $(x^n)_n$ defined by $x^{n+1} = \Psi_1(x^n - \sigma^2\nabla g(x^n)/2)$ for some fixed

σ such that $\sigma^2/2 \in (0, L_g^{-1}]$ converges to a minimum of $x \mapsto g(x) + 2\gamma/\sigma^2\|x\|_{2,1}$. In the case when $\pi(x) \propto \exp(-(g(x)+2\gamma/\sigma^2\|x\|_{2,1}))$, this remark gives an insight on the proposal mechanism (3) of STMALA and shows that it can be read as an extension of MALA to non-differentiable target densities: the proposed sample $Z = \Psi_1(X^n - \sigma^2\nabla g(X^n)/2 + \sigma\Xi^{n+1})$ is obtained by moving from the current sample X^n to a point which is a sparse perturbation of the point $\Psi_1(X^n - \sigma^2\nabla g(X^n)/2)$ which has higher probability under π than X^n (as soon as $\sigma^2/2 \leq L_g^{-1}$). Therefore, our proposal mechanism can be seen as one iteration of a stochastic $L_{2,1}$-gradient proximal algorithm designed to converge to the minima of $x \mapsto g(x) + 2\gamma/\sigma^2\|x\|_{2,1}$.

We now provide an explicit expression of the proposal distribution q_{Ψ_1}, which is required in order to compute the acceptance probability in Algorithm 1. Lemma 3.1 applied with $\mu = X^n - \frac{\sigma^2}{2}\nabla g(X^n)$ answers the question.

LEMMA 3.1. *Let $\mu \in \mathbb{R}^{P\times T}$ and positive constants $\gamma, \sigma > 0$. Set $z \stackrel{\text{def}}{=} \Psi_1(\mu + \sigma\xi)$ where $\xi \in \mathbb{R}^{P\times T}$ is a matrix of independent standard Gaussian random variables. The distribution of $z \in \mathbb{R}^{P\times T}$ is given by*

$$\sum_{m\in\mathcal{M}} \left(\prod_{i\notin I_m} p(\mu_{i\cdot})\,\delta_0(\mathrm{d}z_{i\cdot})\right) \left(\prod_{i\in I_m} f(\mu_{i\cdot}, z_{i\cdot})\mathrm{d}z_{i\cdot}\right) , \tag{9}$$

where for any $c \in \mathbb{R}^T$ and $z \in \mathbb{R}^T \setminus \{0\}$

$$p(c) \stackrel{\text{def}}{=} \mathbb{P}\left\{\|c + \xi\|_2 \leq \gamma\right\} \text{ , with } \xi \sim \mathcal{N}(0, \sigma^2 I_T) ,$$
$$f(c, z) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-T/2} \exp\left(-\frac{1}{2\sigma^2}\left\|\left(1 + \frac{\gamma}{\|z\|_2}\right) z - c\right\|_2^2\right)\left(1 + \frac{\gamma}{\|z\|_2}\right)^{T-1} .$$

Lemma 3.1 is proved in Section 6. It implies that the proposal distribution $q_{\Psi_1}(x, z)$ is the mixture (9) when $\mu = x - \frac{\sigma^2}{2}\nabla g(x)$. This proposal distribution is equivalent to sampling a new binary vector $m' = (m'_1, \cdots, m'_P) \in \mathcal{M}$ conditionally to x; and then sampling a new matrix with non null rows in $\mathbb{R}^{|m'|\times T}$ conditionally to (m', x) as follows:

(i) sample independently the components $(m'_i, i \in \{1, \cdots, P\})$ such that m'_i is a Bernoulli random variable with success parameter

$$1 - \mathbb{P}\left(\left\|\left(x - \frac{\sigma^2}{2}\nabla g(x)\right)_{i\cdot} + \xi\right\|_2 \leq \gamma\right) \quad \text{where } \xi \sim \mathcal{N}(0, \sigma^2 I_T) ;$$

(ii) for $i \notin I_{m'}$, set $z_{i\cdot} = 0$; conditionally to (m', x), sample independent random rows such that for any $i \in I_{m'}$, the distribution of $z_{i\cdot}$ is proportional to

$$\exp\left(-\frac{1}{2\sigma^2}\left\|\left(1 + \frac{\gamma}{\|z_{i\cdot}\|_2}\right) z_{i\cdot} - \left(x - \frac{\sigma^2}{2}\nabla g(x)\right)_{i\cdot}\right\|_2^2\right)\left(1 + \frac{\gamma}{\|z_{i\cdot}\|_2}\right)^{T-1} .$$

Other gradient-proximal operators could be considered to define the Shrinkage-Thresholding operator: for any $\gamma > 0$ and any convex function $h : \mathbb{R}^{P\times T} \to \mathbb{R}$ set

$$\Psi(u) \stackrel{\text{def}}{=} \text{argmin}_{x\in\mathsf{X}} \left(h(x) + \frac{1}{\sigma^2}\|x-u\|_2^2 \right) . \tag{10}$$

This operator is such that the deterministic sequence given by $x^{n+1} = \Psi(x^n - \sigma^2\nabla g(x^n)/2)$ converges to a minimum of $x \mapsto g + h$ (see e.g. Beck & Teboulle (2009), or Parikh & Boyd (2013)). Note that the definition (8) corresponds to the case $h(x) = 2\gamma/\sigma^2\|x\|_{2,1}$. When $\pi(x) \propto (-(g(x)+\bar{g}(x)))$ with $\bar{g}$ convex, it is natural to choose $h = \bar{g}$ as soon as the proposal distribution q_Ψ has an explicit expression.

3.2. The hard thresholding operator

Another suggestion of operator for Algorithm 1 is the hard thresholding operator $\Psi_2 : \mathbb{R}^{P\times T} \to \mathbb{R}^{P\times T}$ defined componentwise by

$$(\Psi_2(u))_{i,j} \stackrel{\text{def}}{=} u_{i,j}\mathbf{1}_{\|u_{i\cdot}\|_2>\gamma} . \tag{11}$$

The function $u \mapsto \Psi_2(u)$ is displayed in Figure 1[center] in the case $P = T = 1$. Compared to the $L_{2,1}$ proximal operator (8), this operator avoids shrinkage of the active rows caused by the proximal operator. Lemma 3.2 applied with $\mu = X^n - \frac{\sigma^2}{2}\nabla g(X^n)$ gives the expression of the transition density q_{Ψ_2} in this case.

LEMMA 3.2. *Let $\mu \in \mathbb{R}^{P\times T}$ and positive constants $\gamma, \sigma > 0$. Set $z \stackrel{\text{def}}{=} \Psi_2(\mu + \sigma\xi)$ where $\xi \in \mathbb{R}^{P\times T}$ is a matrix of independent standard Gaussian random variables. The distribution of $z \in \mathbb{R}^{P\times T}$ is given by*

$$\sum_{m\in\mathcal{M}} \left(\prod_{i\notin I_m} p(\mu_{i\cdot})\,\delta_0(\mathrm{d}z_{i\cdot}) \right) \left(\prod_{i\in I_m} f_{ht}(\mu_{i\cdot}, z_{i\cdot})\mathrm{d}z_{i\cdot} \right) ,$$

where for any $c \in \mathbb{R}^T$ and $z \in \mathbb{R}^T$

$$f_{ht}(c,z) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-T/2} \exp\left(-\frac{1}{2\sigma^2}\|z-c\|_2^2 \right) \mathbf{1}_{\|z\|_2>\gamma} ,$$

and $c \mapsto p(c)$ is defined in Lemma 3.1.

The proof of Lemma 3.2 follows the same lines as the proof of Lemma 3.1 and is omitted. Here again, the proposal distribution can be read as sampling first the indices of the active rows in the candidate matrix by sampling a binary vector $m' \in \mathcal{M}$ (with the same sampling mechanism as with the $L_{2,1}$ gradient-proximal operator (8) - see Lemma 3.1); and then, conditionally to (m', x), sampling independently the active rows of the candidate matrix with a truncated Gaussian distribution.

3.3. A soft thresholding function with vanishing shrinkage

The hard thresholding operator Ψ_2 proposed in Section 3.2 avoids the shrinkage of the proposed active rows but also prevents these rows from having a L_2-norm lower than a given threshold. The efficiency of STMALA with Ψ_2 as shrinkage-thresholding operator highly depends on the choice of the threshold, as illustrated in Section 5. To overcome this difficulty, a soft thresholding operator with a vanishing shrinkage can be used. An example of such an operator, known as the empirical Wiener operator (see Siedenburg (2012)), is defined componentwise as follows: for some $\gamma > 0$, $\Psi_3 : \mathbb{R}^{P\times T} \to \mathbb{R}^{P\times T}$ is given by

$$(\Psi_3(u))_{i,j} \stackrel{\text{def}}{=} u_{i,j}\left(1 - \frac{\gamma^2}{\|u_{i\cdot}\|_2^2}\right)_+ . \tag{12}$$

Figure 1[right] displays $u \mapsto \Psi_3(u)$ when $P = T = 1$.

Lemma 3.3, applied with $\mu = X^n - \frac{\sigma^2}{2}\nabla g(X^n)$, gives the expression of the transition density q_{Ψ_3}.

LEMMA 3.3. *Let $\mu \in \mathbb{R}^{P\times T}$ and positive constants $\gamma, \sigma > 0$. Set $z \stackrel{\text{def}}{=} \Psi_3(\mu + \sigma\xi)$ where $\xi \in \mathbb{R}^{P\times T}$ is a matrix of independent standard Gaussian random variables. The distribution of $z \in \mathbb{R}^{P\times T}$ is given by*

$$\sum_{m\in\mathcal{M}} \left(\prod_{i\notin I_m} p(\mu_{i\cdot})\,\delta_0(\mathrm{d}z_{i\cdot})\right)\left(\prod_{i\in I_m} f_{st}(\mu_{i\cdot}, z_{i\cdot})\mathrm{d}z_{i\cdot}\right) ,$$

where for any $c \in \mathbb{R}^T$, $z \in \mathbb{R}^T \setminus \{0\}$, $u > 0$

$$f_{st}(c,z) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-T/2}\left(g\left(\frac{\gamma^2}{\|z\|_2^2}\right)\right)^T \tilde{g}\left(\frac{\gamma^2}{\|z\|_2^2}\right)\exp\left(-\frac{1}{2\sigma^2}\left\|g\left(\frac{\gamma^2}{\|z\|_2^2}\right)\,z - c\right\|_2^2\right) ,$$

$$g(u) \stackrel{\text{def}}{=} 1 + \frac{2u}{1+\sqrt{1+4u}} , \qquad \tilde{g}(u) \stackrel{\text{def}}{=} \frac{1}{\sqrt{1+4u}} ,$$

and $c \mapsto p(c)$ is given by Lemma 3.1.

Lemma 3.3 is proved in section 6. Here again, the proposal distribution can be read as sampling first the indices of the active rows in the candidate matrix by sampling a binary vector $m' \in \mathcal{M}$ (with the same sampling mechanism as with the $L_{2,1}$ gradient-proximal operator (8) - see Lemma 3.1); and then, conditionally to (m', x), sampling independently the active rows of the candidate matrix under the distribution f_{st}.

Lemma 3.4 shows that Ψ_3 compromises between minimizing a (non-convex) function h and being near to u.

LEMMA 3.4. *For any $\gamma > 0$ and $u \in \mathbb{R}^\ell$,*

$$\Psi_3(u) = \mathrm{argmin}_{x\in\mathbb{R}^\ell}\left(h(x) + \frac{1}{2}\|x-u\|^2\right) ,$$

where the function $h : \mathbb{R}^\ell \to \mathbb{R}$ *is given by*

$$h(x) = \gamma^2 \left[asinh\left(\frac{\|x\|}{2\gamma}\right) - \frac{1}{2}\exp\left(-2asinh\left(\frac{\|x\|}{2\gamma}\right)\right)\right] .$$

PROOF. The proof of Lemma 3.4 is in Section 6.2.

4. V-Geometric ergodicity of the $L_{2,1}$ proximal STMALA

In this section, we address the V-geometric ergodicity of the STMALA chain $(X^n)_{n\geq 0}$ where, at iteration n, the candidate Z is given by

$$Z = \Psi_1\left(X^n - \frac{\sigma^2}{2}\frac{D\nabla g(x)}{\max(D, \|\nabla g(x)\|_2)} + \sigma\Xi^{n+1}\right) , \tag{13}$$

where $(\Xi^n, n \geq 1)$ is a sequence of $P \times T$ random matrices with i.i.d. $\mathcal{N}(0,1)$ entries and Ψ_1 is given by (8). Hereafter, γ (in the definition of Ψ_1) and D are fixed positive constants.

This update differs from the update proposed in (3) by truncating the gradient, as it was already suggested in Roberts & Tweedie (1996) and used in Atchadé (2006). This truncation makes the algorithm more stable in practice. In most of the examples presented in Section 5, we observed that truncating the gradient has only a minor impact on the results, which are therefore presented with no truncation. For the real data set presented in Section 5.3, the truncation prevents the algorithm from moving too far from the current state and therefore avoids too low acceptance rates.

To make the remainder of the paper less technical, the proof is given in the case $T = 1$ and $\Psi = \Psi_1$. Extensions to $T > 1$ and other shrinkage-thresholding operators are not addressed in this paper.

The sets $\{S_m, m \in \mathcal{M}\}$, where S_m is given by (5), are a partition of $\mathbb{R}^P$ and we denote by π_m the restriction of π to S_m:

$$\pi(x) = \sum_{m\in\mathcal{M}} \pi_m(x)\mathbf{1}_{S_m}(x) .$$

The convergence of this STMALA is addressed under the following assumptions on the target density π.

A2 (i) For any $m \in \mathcal{M}$, π_m is continuous on S_m.

(ii) $\pi(x) \to 0$ when $\|x\|_2 \to \infty$.

Assumption A3 below ensures that π is super-exponential, *i.e.* that the target density π decreases fast enough when $\|x\|_2$ is large.

A3 For any $s > 0$, $m \in \mathcal{M}$,

$$\lim_{r\to\infty} \sup_{x\in S_m, \|x\|\geq r} \frac{\pi_m(x + s\, n(x))}{\pi_m(x)} = 0 \,, \qquad \text{where} \qquad n(x) \stackrel{\text{def}}{=} \frac{x}{\|x\|} \,.$$

When for any $m \in \mathcal{M}$, the restriction π_m of π to the subset S_m is differentiable, A3 is satisfied if (see (Jarner & Hansen 2000, Section 4) for details)

$$\forall m \in \mathcal{M}, \quad \lim_{x\in S_m, \|x\|_2\to\infty} \left\langle \frac{x}{\|x\|_2}, \nabla \log(\pi_m(x)) \right\rangle = -\infty \,. \tag{14}$$

Let $u, b, \epsilon > 0$ such that $u \in (0, b)$. For any $m \in \mathcal{M}$ and $x \in S_m$, define

$$W_m(x) \stackrel{\text{def}}{=} \{x_m - u\; n(x_m) - s\xi : s \in (0, b-u), \|\xi\|_2 = 1, \|\xi - n(x_m)\|_2 \leq \epsilon\} \,. \tag{15}$$

$W_m(x)$ is the cone of $\mathbb{R}^{|m|}$ with apex $x_m - u\; n(x_m)$ and aperture 2ϵ. We will prove (see Lemma 6.7) that A4 guarantees that, for $\|x\|_2$ large enough, the probability to accept a move from x to any point of $W_m(x)$ equals one.

A4 There exist $b, R, \epsilon > 0$ and $u \in (0, b)$ such that for any $m \in \mathcal{M}$, for any $x \in S_m \cap \{\|x\|_2 \geq R\}$,

$$\forall\; y \in S_m \text{ such that } y_m \in W_m(x), \quad \pi_m(x - u\; n(x)) \leq \pi_m(y) \,,$$

Note that, when π_m is differentiable for any $m \in \mathcal{M}$, A4 is implied by

$$\forall m \in \mathcal{M}, \quad \limsup_{x\in S_m, \|x\|_2\to\infty} \left\langle \frac{x}{\|x\|_2}, \frac{\nabla\pi_m(x)}{\|\nabla\pi_m(x)\|_2} \right\rangle < 0 \,, \tag{16}$$

which is similar to the conditions used in (Jarner & Hansen 2000, condition (32)) for instance (see (Jarner & Hansen 2000, proof of Theorem 4.3) for details).

An example of target density satisfying A1 to A4 and related to the density defined in Section 2.2 is presented in Appendix A.

Theorem 4.1 establishes the V-geometric ergodicity of the STMALA algorithm with truncated gradient; denote by P_{trunc} the transition kernel associated with the Hastings-Metropolis chain $(X^n)_n$ with proposal distribution given by (13). The dependence upon the constants D, γ, σ (which are fixed by the user prior any run of the algorithm) is omitted to simplify the notations.

THEOREM 4.1. *Assume A1 to A4 hold. Then, for any $\beta \in (0, 1)$, there exist $C > 0$ and $\rho \in (0, 1)$ such that for any $n \geq 0$ and any $x \in \mathbb{R}^P$,*

$$\|P^n_{trunc}(x, .) - \pi\|_V \leq C\, V(x)\, \rho^n \,, \tag{17}$$

where $V(x) \propto \pi(x)^{-\beta}$ and for any signed measure η, $\|\eta\|_V = \sup_{f, |f|\leq V} |\int f \mathrm{d}\eta|$.

By definition of the acceptance probability in Algorithm 1, π is invariant with respect to P_{trunc}. The rate of convergence is then a consequence of Proposition 6.4 and Proposition 6.5 given in Section 6.3: Proposition 6.4 establishes that the chain is psi-irreducible and aperiodic and shows that any Borel set $C \in \mathbb{R}^P$ such that $C \cap S_m$ is a compact subset of S_m is a small set for P_{trunc}; Proposition 6.5 shows that there exists a small set $C \in \mathbb{R}^P$ and constants $c_1 \in (0,1)$ and $c_2 < \infty$ such that for any $x \in \mathbb{R}^P$,

$$P_{trunc}V(x) \leq c_1 V(x) + c_2 \mathbf{1}_C(x) \; .$$

The proof is then concluded by (Meyn & Tweedie 1993, Theorem 15.0.2).

5. Numerical illustrations

In this section, STMALA‡ is compared to the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. We detail in Appendix B.1 how RJMCMC is implemented. We also discuss in Appendix B.2 how to approximate the probability $p(c)$ given by Lemma 3.1, a quantity which appears in the implementation of STMALA whatever the shrinkage-thresholding operator $\Psi \in \{\Psi_1, \Psi_2, \Psi_3\}$.

5.1. STMALA on a toy example

5.1.1. Presentation of the data set

The data Y are sampled from the model (4). The design matrix G is obtained by sampling i.i.d. $\mathcal{N}(0,1)$ entries. We consider the target distribution described in Section 2.2, with the prior distribution on $\mathcal{M}$ defined by $w_m = 0.1^{|m|}0.9^{P-|m|}$ and $\lambda = 1$. Moreover, $N = 100$, $P = 16$, $T = 1$, $\tau = 1$ and $X_j = \mathbf{1}_{j \leq S}$ with $S = 8$. Since $T = 1$, note that $\|.\|_{2,1} = \|.\|_1$.

In this section, P is chosen small enough ($P = 16$), so that the posterior distribution of the models $\pi(m|Y)$ can be explicitly computed (see below). This allows to compare the algorithms using the error when estimating the activation probabilities, defined by

$$\mathcal{E} \stackrel{\text{def}}{=} \sum_{i=1}^{P} \left| \mathbb{P}(X_i \neq 0) - \frac{1}{N_{it}} \sum_{n=B}^{N_{it}+B} \mathbf{1}_{X_i^n \neq 0} \right| \; , \tag{18}$$

where N_{it} is the number of iterations used to compute the approximations, B denotes the number of iterations discarded as a burn-in period, and $\mathbb{P}(X_i \neq 0)$ is the posterior probability of activation of the ith component of X, defined for any $1 \leq i \leq P$, by

$$\mathbb{P}(X_i \neq 0) = \sum_{m \in \mathcal{M}} \pi(m|Y) \; m_i \; . \tag{19}$$

‡MATLAB codes for STMALA are available at the address http://perso.telecom-paristech.fr/∼schreck/recherche.html

Let us derive the expression of $\pi(m|Y)$. By (7), for any $m \in \mathcal{M}$,

$$\pi(m|Y) =\propto w_m \; c_\lambda^{-|m|} \int_{\mathbb{R}^{|m|}} \exp\left(-\frac{1}{2\tau}\|Y - G_{\cdot m}x\|_2^2 - \lambda\|x\|_1\right) \mathrm{d}x \; .$$

Then, $\pi(m|Y) = 0$ when the matrix $G'_{\cdot m}G_{\cdot m}$ is not invertible.; otherwise,

$$\pi(m|Y) \propto w_m \; c_\lambda^{-|m|} \exp\left(\frac{1}{2\tau}Y' A_m Y\right) (2\pi\tau)^{|m|/2} \left(\det\left(G'_{\cdot m}G_{\cdot m}\right)\right)^{-1/2} \cdots$$
$$\times \int_{\mathbb{R}^{|m|}} \phi_m(x) \exp\left(-\lambda\|x\|_1\right) \mathrm{d}x \; ,$$

where $\overline{X}(m) = (G'_{\cdot m}G_{\cdot m})^{-1}G'_{\cdot m}Y$, $A_m \stackrel{\text{def}}{=} G_{\cdot m}(G'_{\cdot m}G_{\cdot m})^{-1}(G_{\cdot m})'$ and ϕ_m denotes the probability density function of a Gaussian vector with mean $\overline{X}(m)$ and covariance matrix $\tau\ (G'_{\cdot m}G_{\cdot m})^{-1}$. The last integral is not explicit when $\lambda \neq 0$, but it can be estimated by a standard Monte Carlo method.

5.1.2. Discussion on the implementation parameters

Comparison of the shrinkage-thresholding operators Figure 2 provides a comparison of the three shrinkage-thresholding operators proposed in Section 3 for the model described in Section 5.1.1. Figure 2 (left) displays the evolution of the mean error given by block-STMALA, with $L_{2,1}$ proximal operator (Prox), hard thresholding operator (HT) and soft thresholding with vanishing shrinkage operator (STVS), over 100 independent trajectories as a function of the number of iterations. Let L_g be the lipschitz norm of the gradient of g, $L_g \stackrel{\text{def}}{=} \|GG^t\|/\tau$. All the algorithms are run with $\sigma = \sqrt{2/L_g}$. The number of components to be updated at each iteration is $\eta = 4$ (the role of the block size is discussed below) and the threshold is set to $\gamma = 0.1$ (see details on the choice of the threshold below). All the algorithms start from the null regressor.

As HT (solid blue line) does not shrink the nonzero regression parameters (see figure 1), it cannot produce any values lower than γ. Therefore, STMALA with hard thresholding is not robust to the choice of the threshold and did not reach convergence in the situation presented in Figure 2 (left). On the other hand, the estimation errors of block-STMALA with STVS (dash-dot red line) and with Prox (dashed green line) decrease at a similar rate and variability in this case.

Figure 2 (right) displays the mean acceptance rate as a function of the number of samples N_{it} for the three algorithms. At each iteration, if the proposed point is sampled into a region of high probability under the target distribution, the shrinking step of block-STMALA with proximal shrinkage may drive this sample toward regions of lower probability, and therefore decrease the mean acceptance-rejection ratio. Figure 2 (right) shows that block-STMALA,

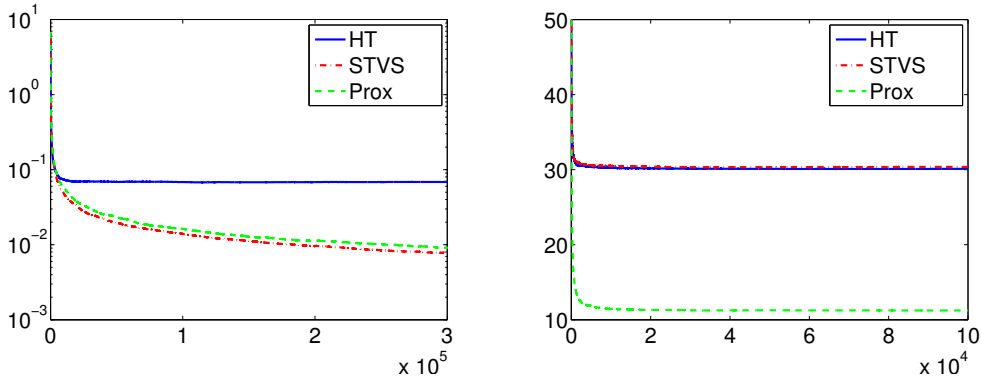

Fig. 2: (left) Evolution of the mean estimation error (18) of the activation probabilities for block-STMALA as a function of the number of iterations, when $\Psi = \Psi_1$ (Prox), $\Psi = \Psi_2$ (HT) and $\Psi = \Psi_3$ (STVS) as shrinkage-thresholding operator. (right) Evolution of the mean acceptance rate.

with no shrinkage (HT), and block-STMALA with a vanishing shrinkage (STVS), accept twice as many proposed samples as block-STMALA with Prox. For more complex models, the mean acceptance-rejection ratio of block-STMALA with Prox can even decrease to zero so that the algorithm remains quickly trapped in one point.

As block-STMALA with STVS combines the best behavior both in terms of acceptance rate and of mean error, this variant is used for the rest of the numerical experiments. Therefore, from now on, the shrinkage-thresholding operator is chosen as $\Psi = \Psi_3$.

The block size η When all the components of a point are updated at each iteration (*i.e.* if $\eta = P$), the distance between the current point and the proposed one is such that it leads to a low acceptance-rejection ratio. Figure 3 (left) displays the mean acceptance-rejection ratio over 100 independent trajectories after $N_{it} = 10^5$ iterations of block-STMALA with no burn-in period ($B = 0$) as a function of η (expressed as a percentage of the total number of regressors) in the model described in Section 5.1.1.

Note that if the acceptance-rejection ratio is too low, the convergence is very slow as few proposed samples are accepted. However, a slow convergence can also be the consequence of small block sizes, since two consecutive samples have at least $P - \eta$ equal coefficients. We observed that, in general, choosing rather small block sizes yields good results.

The threshold γ The choice of the threshold is crucial. If γ is too large, few nonzero samples are proposed and the algorithm will converge slowly. If γ is too small, the algorithm

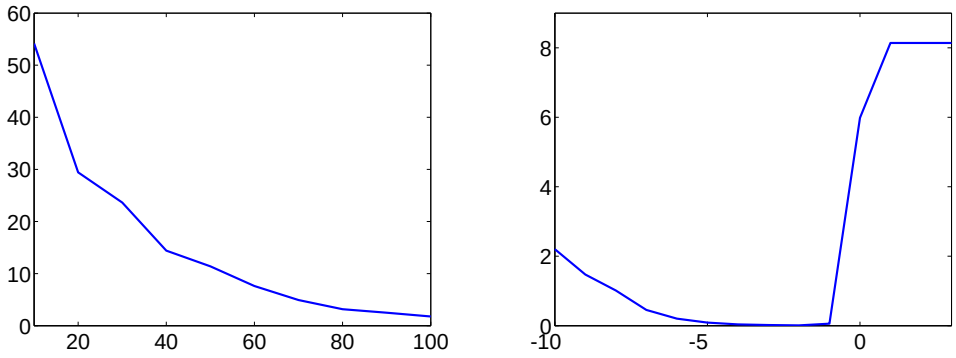

Fig. 3: (left) Mean acceptance-rejection ratio as a function of the block size (expressed as a percentage of P) and (right) mean error as a function of $\ln(\gamma)$.

proposes non-sparse solutions that are not likely to be accepted. This is illustrated in figure 3 (right) which displays the mean error made by block-STMALA when estimating the activation probabilities, computed over 100 independent trajectories of $N_{it} = 10^5$ iterations with no burn-in ($B = 0$), as a function of $\ln(\gamma)$. Here $\eta = 4$ and $\sigma = \sqrt{2/L_g}$, and the computations are made for the model described in Section 5.1.1.

The standard deviation σ Figure 4 displays the mean acceptance-rejection ratio and the mean error (18) over 100 independent trajectories, of block-STMALA as a function of σ (the scale is $\sigma\sqrt{L_g/2}$ on the x-axis) after 10^5 iterations. η is set to 4 and the threshold γ is chosen so that the mean number of thresholded coefficients in one iteration of STMALA starting from the empty model is about 55%. If σ is too large, the distance between the current point and the proposed point is high and leads to a low acceptance-rejection ratio and a slow convergence. If σ is too small, $q_\Psi(z, \cdot)$, where z is the proposed point, is a spike function centered at $\Psi_3\left(z - \frac{\tau^2}{2}\nabla g(z)\right)$, with $\gamma = \tau^2\lambda/2$, which could be quite far from the current point X^n, thus producing too small values of $q_\Psi(z, X^n)/q_\Psi(X^n, z)$ and leading therefore to a low acceptance-rejection ratio and a slow convergence.

5.1.3. Further illustrations

This section provides additional experiments using block-STMALA for the model described in Section 5.1.1. We set $\sigma = \sqrt{2/L_g}$, $\eta = 4$ and $\gamma = 0.28$. The algorithm is initialized to the null regression vector.

Figure 5 compares the mean error over 100 independent trajectories made by block-STMALA with STVS operator (solid blue) and RJMCMC (dash-dot red) when estimating

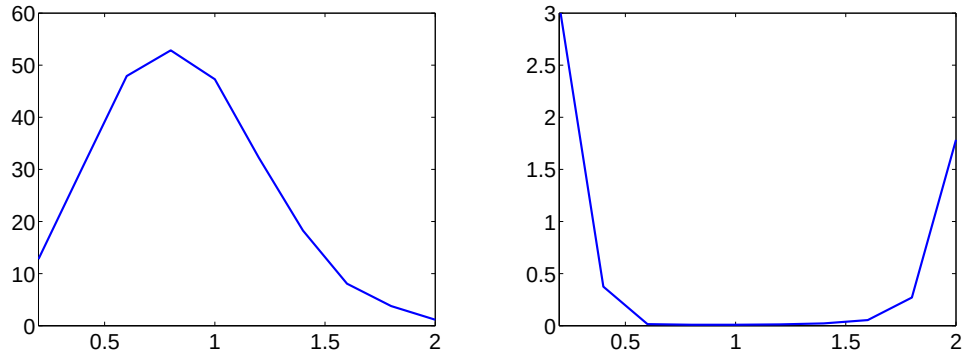

Fig. 4: (left) Mean acceptance rate. (right) Mean error as a function of $\sigma\sqrt{L_g/2}$.

the activation probabilities as a function of the number of iterations (right) and displays the associated boxplots (left). For block-STMALA, we set $\gamma = 0.07$, $\eta = 4$, $\sigma = \sqrt{2/L_g}$ and $B = 0$. RJMCMC is implemented as described in Section B.1 with $\sigma_{\mathrm{RJ}} = 0.02$. The parameters of block-STMALA and RJMCMC are chosen so that the two algorithms have similar acceptance rates (about 23 %). Note that both algorithms are compared for the same number of evaluations of the target density π, *i.e.* for the same number of iterations, as the computation time depends on the code efficiency. In Figure 5, block-STMALA clearly outperforms RJMCMC. For example, after 300.000 iterations, the error made by RJMCMC is twice as big as the error made by block-STMALA. This is due to the fact that RJMCMC only modifies one component at each iteration. Therefore, block-STMALA, which modifies 4 components at each iteration, moves faster.

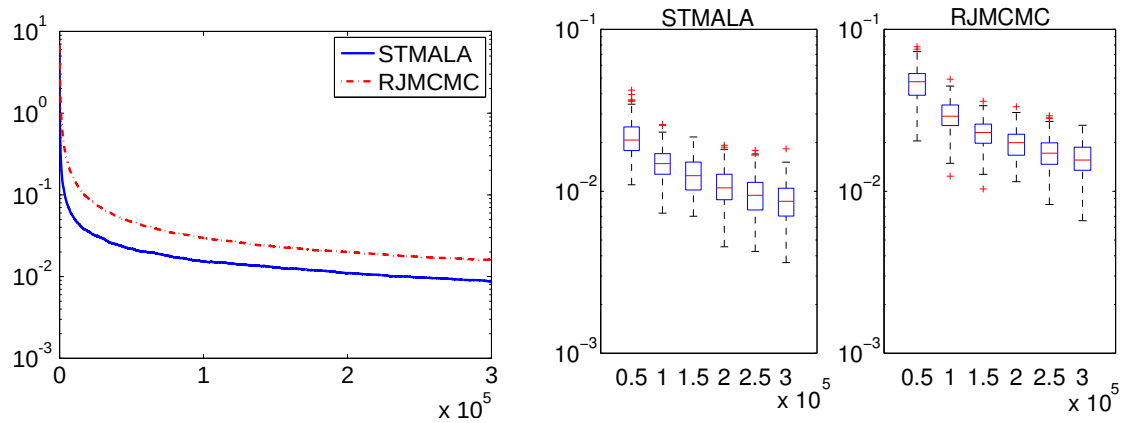

Fig. 5: Evolution of the mean error for block-STMALA and RJMCMC as a function of the number of iterations (left) and the associated boxplots (right).

Figure 6 (left) shows the empirical autocorrelation function of X_1 and X_8 for the two

algorithms, with $\eta = 4$, $\gamma = 0.07$ and $\sigma = \sqrt{2/L_g}$ for block-STMALA and the standard deviation of the random walk in RJMCMC set to $\sigma_{\mathrm{RJ}} = 0.02$. The autocorrelation is computed along a single trajectory of 300.000 iterations (with 30% of these iterations discarded as a burn-in period). The mean regression vectors obtained by block-STMALA and RJMCMC are displayed in Figure 6 (right).

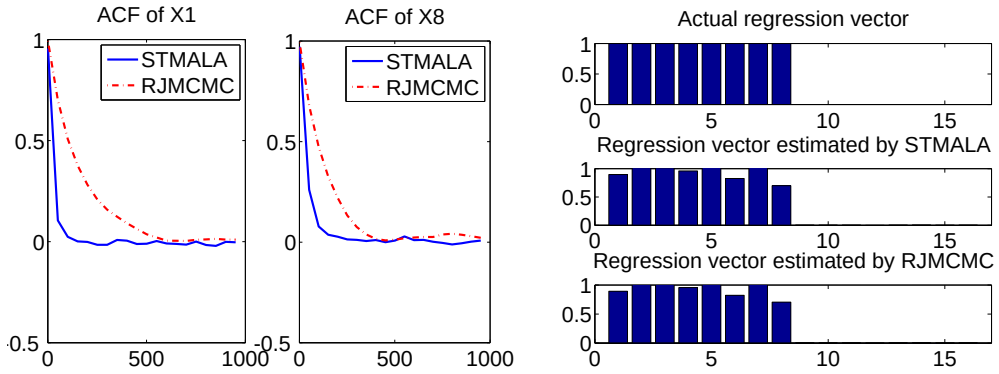

Fig. 6: (left) Empirical autocorrelation function of X_1 and X_8 of block-STMALA and RJMCMC. (right) Regression vectors estimated by block-STMALA and RJMCMC.

Finally, Figure 7 displays the evolution of the mean estimators of $\int x_i \, \pi(x|Y)\mathrm{d}\nu(x)$ for $i = 1$ and $i = 8$ computed by block-STMALA and RJMCMC as a function of the number of iterations and the associated boxplots. Once again the computations are made for the model described in Section 5.1.1 over 100 independent trajectories of 300.000 iterations. In figure 7 (left), block-STMALA converges faster than RJMCMC. This is confirmed by the boxplots shown in figure 7 (right).

5.2. A sparse spike and slab model

5.2.1. The model

The model for the observations $Y \in \mathbb{R}^N$ is assumed to be

$$Y = GX + \vartheta^{-1/2} E \, ,$$

where G is a $N \times P$ (known) design matrix, E is a Gaussian random vector with i.i.d. standard entries and ϑ is the (known) precision. We want to find the subset of nonzero covariate parameters from $X \in \mathbb{R}^N$. We consider the following sparse spike and slab hierarchical model.

- Given $m = (m_\ell)_{1 \le \ell \le P} \in \mathcal{M}$ and positive precisions $(\vartheta_\ell)_{1 \le \ell \le P}$, the entries of the covariate

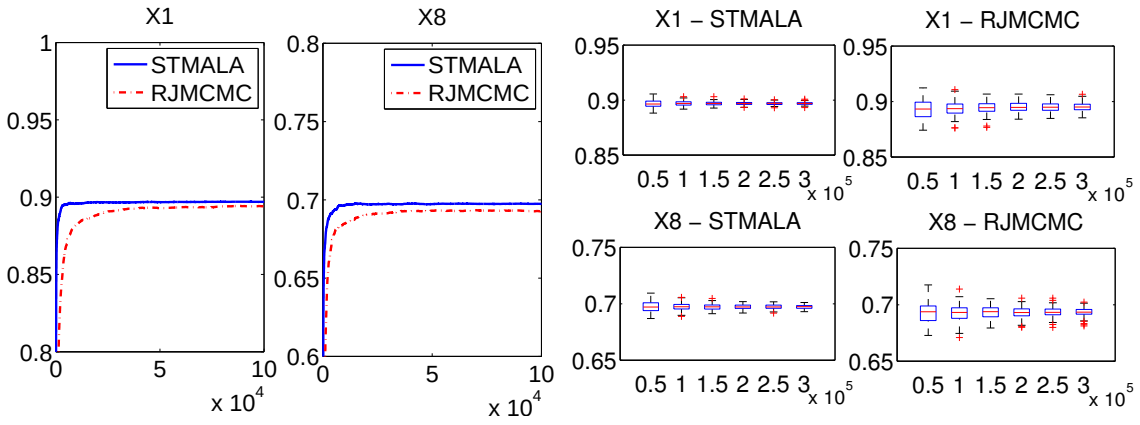

Fig. 7: (left) Evolution of the mean estimators (over 100 independent runs) of $\int x_i \ \pi(x|Y)\mathrm{d}\nu(x)$ for $i = 1$ and $i = 8$ computed by block-STMALA and RJMCMC as a function of the number of iterations. (right) Associated boxplots.

vector $X = (X_\ell)_{1\leq\ell\leq P}$ are independent with distribution

$$(X_k|m, \vartheta_1, \cdots, \vartheta_P) \sim \begin{cases} \delta_0(X_k) & \text{if } m_k = 0, \\ \mathcal{N}(0, 1/\vartheta_k) & \text{if } m_k = 1. \end{cases}$$

- The precision parameters $(\vartheta_\ell)_{1\leq\ell\leq P}$ are i.i.d. with Gamma distribution $\mathrm{Ga}\,(a, aK)$, where a, K are fixed.

- The components of $m \in \mathcal{M}$ are i.i.d. Bernoulli random variables with parameter $\omega_\star$.

Under this model the posterior density $\pi(X, m, \vartheta_1, \cdots, \vartheta_P|Y)$ can be derived analytically:

$$\pi(X, m, \vartheta_1, \cdots, \vartheta_P|Y) \propto \exp\left(-\frac{\vartheta}{2}\|Y - GX\|^2\right) \omega_\star^{|m|}(1-\omega_\star)^{P-|m|}$$
$$\times \prod_{\ell=1}^{P} \vartheta_\ell^{a-1} \exp(-aK\vartheta_\ell)\mathbf{1}_{\mathbb{R}^+}(\vartheta_\ell) \left\{\exp\left(-\frac{\vartheta_\ell}{2}X_\ell^2\right)\sqrt{\vartheta_\ell}\mathbf{1}_{m_\ell=1} + \delta_0(X_\ell)\ \mathbf{1}_{m_\ell=0}\right\}.$$

By integrating out, we obtain the posterior density of (X, m):

$$\pi\,(X, m|Y) \propto \exp\left(-\frac{\vartheta}{2}\|Y - GX\|^2\right) \omega_\star^{|m|}(1-\omega_\star)^{P-|m|}$$
$$\times \prod_{\ell=1}^{P} \left\{\left(1 + \frac{X_\ell^2}{2aK}\right)^{-(a+1/2)} \mathbf{1}_{m_\ell=1} + \delta_0(X_l)\ \mathbf{1}_{m_\ell=0}\right\}.$$

5.2.2. Numerical illustrations

The performance of block-STMALA with STVS operator is illustrated with the model introduced in Breiman (1992) and presented in (Ishwaran & Rao 2005, Section 8). It is

assumed that $N = 100$, $P = 200$, $\omega_\star = 0.1$ and that ϑ is known and fixed to $\vartheta = 1$. The covariates $(G_{:,i})_{1\leq i\leq P}$ are sampled from a Gaussian distribution with $\mathbb{E}[G_{\cdot i}] = 0$ and $\mathbb{E}[G_{ji}G_{ki}] = \rho^{|j-k|}$. In the example below, ρ is given by $\rho = 0.3$. The nonzero coefficients of X are in 4 clusters of 5 adjacent variables such that, for all $k \in \{1,2,3,4\}$ and all $j \in \{1,2,3,4,5\}$, $X_{50*(k-1)+j} = (-1)^{k+1}\, j^{1/k}$. $a = 2$ and $K = 0.08$ are chosen so that the Gamma distribution with parameters a and aK has a mode at $\vartheta_\star$ such that $\vartheta_\star^{-1/2} = \max(|X|)/2$. The other parameters are given by $\sigma = \sqrt{2/L_g}$, $\eta = 20$ and $\gamma = 0.35$. The standard deviation of the RJMCMC proposal is $\sigma_{\mathrm{RJ}} = 0.01$ and chosen so that block-STMALA and RJMCMC have similar acceptance rates (between 15% and 20%). The computations are made over 50 independent trajectories of 10^6 iterations. As the dimension is high and RJMCMC modifies only one component uniformly chosen at each iteration, its autocorrelation function is expected to converge more slowly than the autocorrelation function of block-STMALA. This is illustrated in figure 8 (left) which displays the two mean autocorrelation functions estimated over the 50 independent trajectories of length 10^6 iterations (10% of these iterations are discarded as a burn-in period).

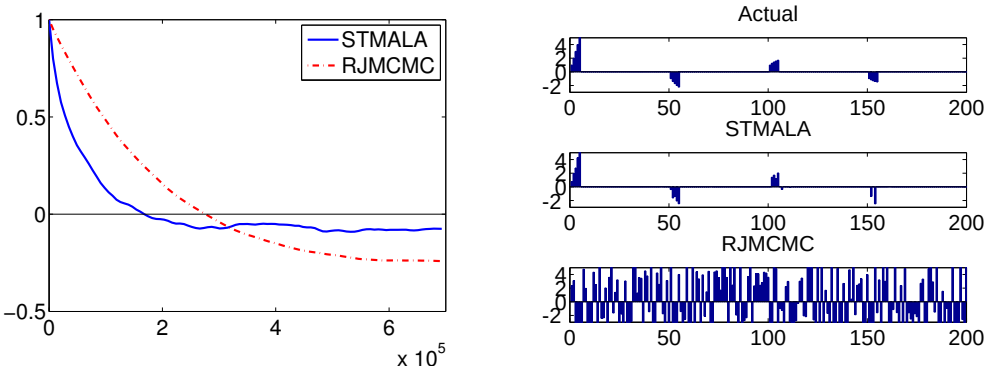

Fig. 8: (left) Mean autocorrelation function of X_1 for block-STMALA and RJMCMC. (right) Regression vectors estimated by block-STMALA and RJMCMC.

Figure 8 (right) shows the true regression vector X and its estimates by block-STMALA and RJMCMC: it shows that block-STMALA provides a sparse estimation while RJMCMC needs a lot of components to explain the observations. This is probably because RJMCMC is more or less equivalent to test each model in turn, which yields slow convergence in high dimensional settings. This slow convergence is also illustrated in Figure 9. Figure 9 (left) shows the evolution of the mean number of active components $|m|$. RJMCMC has not converged after the 300.000 iterations while the mean number of active components of block-STMALA is stable after few iterations. Figure 9 (right) displays the boxplots of

the estimation of the first component X_1 estimated by block-STMALA and RJMCMC as a function of the number of iterations.

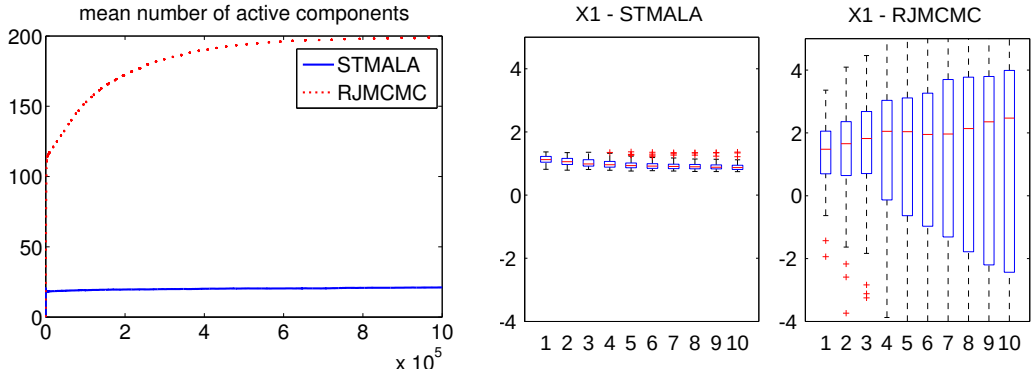

Fig. 9: (left) Evolution of the mean number of active components for STMALA and RJMCMC. (right) Evolution of the estimation of X_1 (mean over iterations) for block-STMALA and RJMCMC.

Figure 10 (left) shows the signal $G\hat{X}$ estimated by block-STMALA and RJMCMC as a function of the actual emitted signal GX (blue circles), where $\hat{X}$ is the mean regression vector over a trajectory. To highlight over fitting effects, a test sample $Y_{\text{test}} = G_{\text{test}}X + \sqrt{1/\vartheta}E_{\text{test}}$, where $G_{\text{test}} \in \mathbb{R}^{100\times 200}$ and $E_{\text{test}} \in \mathbb{R}^{100}$ are generated exactly as G and E, is also used. With green circles, $G_{\text{test}}\hat{X}$ as a function of $G_{\text{test}}X$ are displayed. This test data set is also used to compute a test error, which is given by

$$\mathcal{E}_{\text{test}} \stackrel{\text{def}}{=} \frac{\|G_{\text{test}}\hat{X} - G_{\text{test}}X\|_2^2}{100} .$$

The evolution of the mean test error ϵ_{test} over 100 independent runs, is displayed in Figure 10 (right). Both figures show that RJMCMC is subject to some over fitting, which is not the case of block-STMALA.

5.3. Regression for spectroscopy data

We use the biscuits data set composed of near infrared absorbance spectra of 70 cookies with different water, fat, flour and sugar contents studied in Brown et al. (2001) and Caron & Doucet (2008). The data are divided into a training data set containing measurements for $N = 39$ cookies, and a test data set containing measurements for 31 cookies. Each row of the design matrix consists of absorbance measurements for $P = 300$ different wavelengths from 1202 nm to 2400 nm with gaps of 4 nm. We compare the results obtained by block-STMALA with those obtained by RJMCMC for the prediction of fat content (*i.e.* $T = 1$).

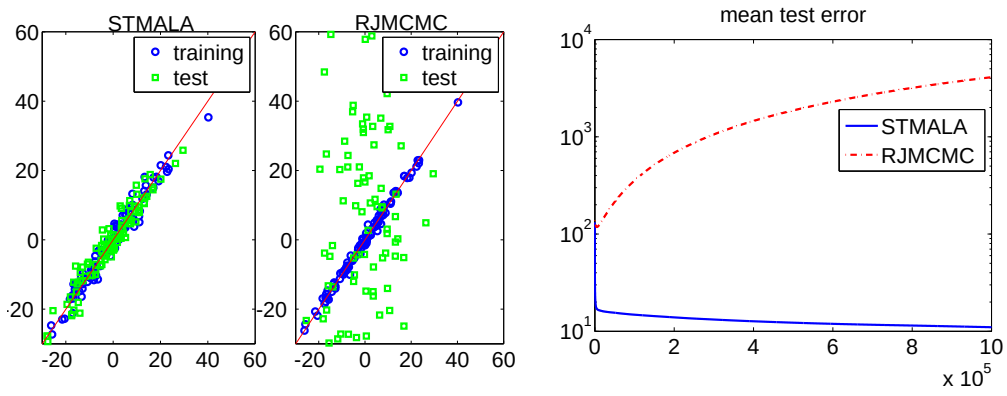

Fig. 10: (left) Emitted signal $G\hat{X}$ estimated by block-STMALA and RJMCMC versus actual emitted signal GX. (right) Evolution of the mean test error for RJMCMC and block-STMALA.

To improve the stability of the algorithm, the columns of the matrix G containing the measurements are centered and a column with each entry being equal to one is added.

The model used here is the one presented in Section 2.2 with the unknown noise parameter set to $\tau = 0.5$. The parameters of the algorithms are given by $\sigma = 2\sqrt{2/L_g}$, $\eta = 15$, $\gamma = 0.35$ for block-STMALA, and by $\sigma_{\mathrm{RJ}} = 0.9$ for RJMCMC. The gradient in block-STMALA is truncated as suggested in Roberts & Tweedie (1996), so that the norm of the truncated gradient does not exceed 0.7. The computations are made over 100 independent trajectories of $N_{it} = 2.10^6$ iterations, with $B = 10^5$. We choose the parameters so that the two algorithms have similar acceptance-rejection ratios (the final ratios are about 45% for block-STMALA and 42% for RJMCMC). Figure 11 shows the regression vectors $\hat{X}$ estimated by block-STMALA and RJMCMC after one trajectory (left) and the mean regression vector estimated by block-STMALA and RJMCMC over the 100 independent trajectories (right).

The regression vector estimated by block-STMALA with STVS operator has a spike around 1726 nm, which is known to be in a fat absorbance region (see Brown et al. (2001), Caron & Doucet (2008)), in almost all the trajectories. The regression vector estimated by RJMCMC has also a spike close to this region, but its location is very unstable over the different trajectories. This instability explains the differences between block-STMALA and RJMCMC, even if the mean regression vectors estimated by the two algorithms over the 100 independent trajectories are quite similar.

Figure 12 displays the boxplots of the 100 independent values of the components of the regression vectors estimated by block-STMALA and RJMCMC associated to 9 wavelengths

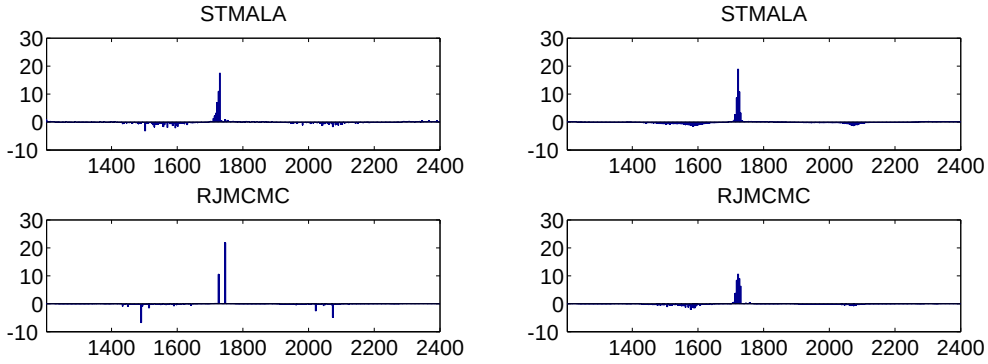

Fig. 11: (left) Regression vectors estimated by block-STMALA and RJMCMC after one trajectory. (right) Mean regression vectors estimated by block-STMALA and RJMCMC over 100 independent trajectories.

close to 1726 nm. It illustrates that the location of the spike retrieved by RJMCMC, when is it retrieved, is not stable, while block-STMALA retrieves a spike centered at 1726 nm in almost every trajectory, even if the height of this spike fluctuates.

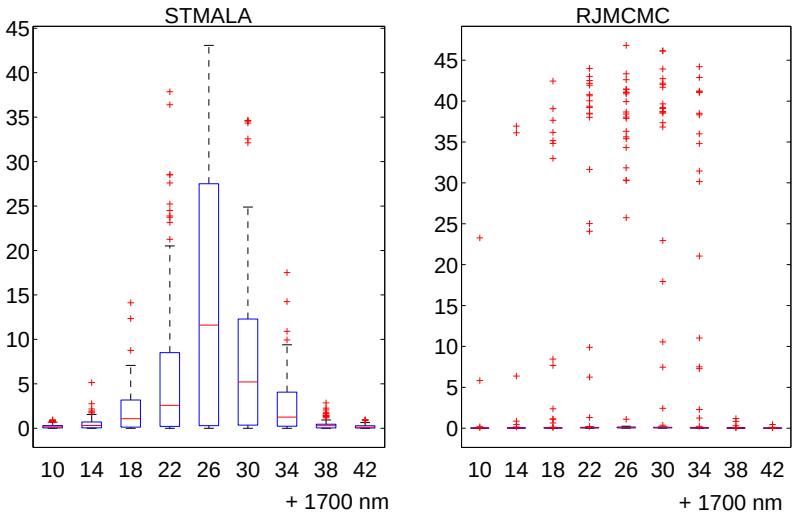

Fig. 12: Boxplots of the 100 independent values of the components of the regression vectors estimated by block-STMALA and RJMCMC associated to 9 wavelengths close to 1726 nm.

Figure 13 (left) shows the emitted signal $G\hat{X}$ estimated over one trajectory by block-STMALA and RJMCMC as a function of the observations Y. In this numerical experiment, block-STMALA provides better results than RJMCMC for both the training set and the test set. This is confirmed by Figure 13 (right) which displays the evolution of the mean

square error (MSE) on the test dataset, defined by

$$\mathrm{MSE} = \frac{\|G_{\mathrm{test}}\hat{X} - Y_{\mathrm{test}}\|_2^2}{31} ,$$

as a function of the number of iterations (mean over 100 independent trajectories). The mean MSE after 2.10^6 iterations is about 0.75 for block-STMALA and about 1.6 times greater for RJMCMC.

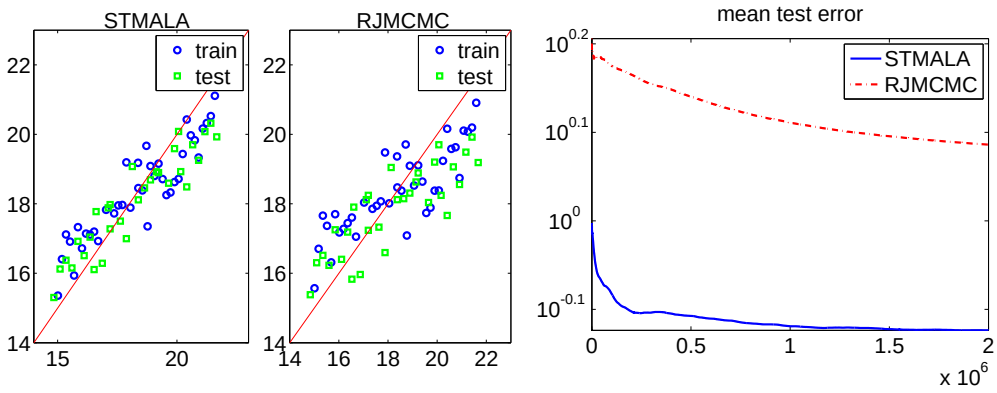

Fig. 13: (left) Emitted signal $G\hat{X}$ estimated by block-STMALA and RJMCMC versus the observations Y. (right) Evolution of the mean MSE (over 100 independent trajectories) on the test data set for RJMCMC and block-STMALA.

6. Proofs

6.1. *Proof of Eq.* (6)

LEMMA 6.1. *The function defined on* $\mathbb{R}^{|m|\times T}$ *by*

$$x \mapsto \left(\frac{\lambda^T \Gamma\left(\frac{T}{2}\right)}{2\pi^{T/2}(T-1)!}\right)^{|m|} \exp(-\lambda\|x\|_{2,1}) ,$$

is a probability density function with respect to the Lebesgue measure. Γ *denotes the standard Gamma function defined on* $(0,+\infty)$ *by* $\Gamma : x \mapsto \int_0^{+\infty} t^{x-1}\mathrm{e}^{-t}\mathrm{d}t$.

PROOF. By definition of $\|\cdot\|_{2,1}$,

$$\int_{\mathbb{R}^{|m|\times T}} \exp(-\lambda\|x\|_{2,1})\mathrm{d}x = \left(\int_{\mathbb{R}^T} \exp(-\lambda\|x\|_2)\mathrm{d}x\right)^{|m|} . \tag{20}$$

Let $S_T \stackrel{\mathrm{def}}{=} \left\{x \in \mathbb{R}^T;\ \|x\|_2 = 1\right\}$ and σ_{T-1} be the $T-1$ dimensional Hausdorff measure on S_T. On one hand, we have

$$\int_{\mathbb{R}^T} \mathrm{e}^{-\lambda\|x\|_2}\mathrm{d}x = \sigma_{T-1}(S_T)\int_{\mathbb{R}_+^\star} r^{T-1}\mathrm{e}^{-\lambda r}\mathrm{d}r = \sigma_{T-1}(S_T)\frac{(T-1)!}{\lambda^T} .$$

On the other hand,

$$\pi^{T/2} = \int_{\mathbb{R}^T} \mathrm{e}^{-\|x\|_2^2} \mathrm{d}x = \sigma_{T-1}(S_T) \int_{\mathbb{R}_+^\star} \mathrm{e}^{-r^2} r^{T-1} \mathrm{d}r = \frac{\sigma_{T-1}(S_T)}{2} \Gamma\left(\frac{T}{2}\right) .$$

which implies that $\sigma_{T-1}(S_T) = 2\pi^{T/2}/\Gamma(T/2)$. This concludes the proof.

6.2. Proofs of Section 3

6.2.1. Proof of Lemma 3.1

For any $1 \leq i \leq P$, $1 \leq j \leq T$, and any $y \in \mathbb{R}^{P \times T}$, define the proximal operator

$$\left(\mathrm{prox}_{\gamma\|\cdot\|_{2,1}}(y)\right)_{ij} = \begin{cases} 0 & \text{if } i \text{ such that } \|y_{i\cdot}\|_2 \leq \gamma \ , \\ y_{i,j}\left(1 - \gamma/\|y_{i\cdot}\|_2\right) & \text{otherwise} \ . \end{cases}$$

Let φ be a bounded continuous function on $\mathbb{R}^{P \times T}$.

$$\mathbb{E}[\varphi(Z)] = \left(2\pi\sigma^2\right)^{-TP/2} \int_{\mathbb{R}^{P\times T}} \varphi\left(\mathrm{prox}_{\gamma\|\cdot\|_{2,1}}(y)\right) \prod_{i=1}^{P} \exp\left(-\frac{\|y_{i\cdot} - \mu_{i\cdot}\|^2}{2\sigma^2}\right) \mathrm{d}y \ .$$

For $y \in \mathbb{R}^{|m|\times T}$, denote by $\overline{y}$ the $(|m| \times T)$-matrix defined by $\overline{y}_{i\cdot} = y_{i\cdot}\left(1 - \gamma/\|y_{i\cdot}\|_2\right)$. Fubini's theorem yields

$$\begin{aligned} \mathbb{E}[\varphi(Z)] = \left(2\pi\sigma^2\right)^{-TP/2} \sum_{m\in\mathcal{M}} \prod_{i\notin I_m} \int \mathbf{1}_{\|y_{i\cdot}\|_2 \leq \gamma} \exp\left(-\frac{\|y_{i\cdot} - \mu_{i\cdot}\|^2}{2\sigma^2}\right) \mathrm{d}y_{i\cdot} \\ \times \int_{\mathbb{R}^{|m|\times T}} \varphi\left((m, \overline{y})\right) \left(\prod_{k=1}^{|m|} \mathbf{1}_{\|\overline{y}_{k\cdot}\|_2 > \gamma}\right) \exp\left(-\frac{\|\overline{y} - \mu_{m\cdot}\|_2^2}{2\sigma^2}\right) \mathrm{d}\overline{y} \ , \\ = \left(2\pi\sigma^2\right)^{-TP/2} \sum_{m\in\mathcal{M}} \prod_{i\notin I_m} p\left(\mu_{i\cdot}\right) \int_{\mathbb{R}^{|m|\times T}} \varphi\left((m, \overline{y})\right) \left(\prod_{k=1}^{|m|} \mathbf{1}_{\|\overline{y}_{k\cdot}\|_2 > \gamma}\right) \\ \exp\left(-\frac{\|\overline{y} - \mu_{m\cdot}\|_2^2}{2\sigma^2}\right) \mathrm{d}\overline{y} \ . \end{aligned}$$

By Fubini's theorem, it is sufficient to compute integrals of the form

$$\int_{\mathbb{R}^T} \widetilde{\varphi}\left(v\left(1 - \frac{\gamma}{\|v\|_2}\right)\right) \mathbf{1}_{\|v\|_2 > \gamma} \exp\left(-\frac{\|v - \mu_{i,\cdot}\|_2^2}{2\sigma^2}\right) \mathrm{d}v \ ,$$

for a generic function $\widetilde{\varphi}$. Consider the change of variable $\mathbb{R}^T \to \mathbb{R}^T$ $z = v\left(1 - \frac{\gamma}{\|v\|_2}\right)$. Note that $\|z\|_2 = \|v\|_2 - \gamma$ and that $v = \psi(z)$, where for any $z \in \mathbb{R}^T$, $\psi(z) \stackrel{\text{def}}{=} \frac{\|z\|_2 + \gamma}{\|z\|_2} z$. We now determine the Jacobian matrix of ψ. Hereafter, z and h are elements of $\mathbb{R}^T$. For any h, z such that $z \neq 0$,

$$\|z + h\|_2 = \|z\|_2 + \left\langle \frac{z}{\|z\|_2}, h \right\rangle + o\left(\|h\|_2\right) \ .$$

Then,

$$\frac{1}{\|z + h\|_2} = \frac{1}{\|z\|_2} \frac{1}{1 + \left\langle \frac{z}{\|z\|_2^2}, h \right\rangle + o\left(\|h\|_2\right)} = \frac{1}{\|z\|_2} \left(1 - \left\langle \frac{z}{\|z\|_2^2}, h \right\rangle + o\left(\|h\|_2\right)\right) \ .$$

Therefore,

$$\psi(z+h) = \left(1+\frac{\gamma}{\|z+h\|_2}\right)(z+h) = \psi(z) + \left\{\left(1+\frac{\gamma}{\|z\|_2}\right) I_T - \frac{\gamma}{\|z\|_2^3} zz^\star\right\} h + o(\|h\|_2)$$

and the Jacobian matrix of ψ at z is

$$J\psi(z) = \left(1+\frac{\gamma}{\|z\|_2}\right) I_T - \frac{\gamma}{\|z\|_2^3} zz^\star \; .$$

Define the unit vector $\omega \stackrel{\text{def}}{=} z/\|z\|_2$. Then, the determinant of $J\psi(z)$ is given by

$$\begin{aligned}\mathrm{Det}\left(J\psi(z)\right) &= \left(1+\frac{\gamma}{\|z\|_2}\right)^T \mathrm{Det}\left(I_T - \frac{\gamma}{\gamma+\|z\|_2}\omega\omega^\star\right) \; , \\ &= \left(1+\frac{\gamma}{\|z\|_2}\right)^T \left(1-\frac{\gamma}{\gamma+\|z\|_2}\right) = \left(1+\frac{\gamma}{\|z\|_2}\right)^{T-1} \; .\end{aligned}$$

Finally,

$$\begin{aligned}&\int_{\mathbb{R}^T} \widetilde{\varphi}\left(v\left(1-\frac{\gamma}{\|v\|_2}\right)\right) \mathbf{1}_{\|v\|_2>\gamma} \exp\left(-\frac{\|v-\mu_{i\cdot}\|_2^2}{2\sigma^2}\right) \mathrm{d}v \\ &\qquad = \int_{\mathbb{R}^T} \widetilde{\varphi}\left(v\right) \exp\left(-\frac{\|\psi(v)-\mu_{i\cdot}\|_2^2}{2\sigma^2}\right) \left(\frac{\gamma+\|v\|_2}{\|v\|_2}\right)^{T-1} \mathrm{d}v \; .\end{aligned}$$

This concludes the proof.

6.2.2. Proof of Lemma 3.3

The proof of Lemma 3.3 follows the same lines as the proof of Lemma 3.1, with the function ψ replace by $\widetilde{\psi}(z) = g\left(\gamma^2/\|z\|_2^2\right)\; z$. We detail the computation of the Jacobian. We have

$$\nabla\widetilde{\psi}(z) = g\left(\frac{\gamma^2}{\|z\|_2^2}\right)\; I + g'\left(\frac{\gamma^2}{\|z\|_2^2}\right)\left(-\frac{\gamma^2}{\|z\|_2^4}\right)\; 2zz^\star \; ,$$

and for any $u>0$, $g'(u) = 1/\sqrt{1+4u}$. Th proof follows upon noting that for any a, b, $\mathrm{Det}(aI - bzz^\star) = a^T - a^{T-1}b\|z\|_2^2$.

6.2.3. Proof of Lemma 3.4

Proof in the case $\ell = 1$. We first compute the derivative of h on $]0,\infty[$ (note that h is symmetric). For any $x \in]0,+\infty[$,

$$h'(x) = \gamma^2\left[\frac{1}{\sqrt{x^2+4\gamma^2}} + \frac{1}{\sqrt{x^2+4\gamma^2}} \exp\left(-2\mathrm{asinh}\left(\frac{x}{2\gamma}\right)\right)\right] \; .$$

As $\exp(-2t) = 2sh^2(t) + 1 - 2\sqrt{1+sh(t)^2}sh(t)$, this yields

$$h'(x) = \frac{-x+\sqrt{x^2+4\gamma^2}}{2} \quad \text{for any} \quad x>0 \; .$$

Since h is symmetric,

$$h'(x) = \frac{-x - \sqrt{x^2 + 4\gamma^2}}{2} \quad x < 0 \, .$$

Set $\psi_u(x) \stackrel{\text{def}}{=} h(x) + (x-u)^2/2$. Since we have $\psi_{-u}(x) = \psi_u(-x)$, we only have to consider the case when $u \geq 0$. Hereafter, $u \geq 0$. It is easily proved that on $]0, \infty[$, the derivative ψ'_u is strictly increasing to infinity, and a solution to the equation $\psi'_u(x) = 0$ exists on $]0, \infty[$ iff $u > \gamma$. In this case, this solution is $u - \gamma^2/u$, and $\psi_u(u - \gamma^2/u) \leq \psi_u(0)$. When $u \in [0, \gamma)$, $\inf_{x>0} \psi_u(x) = \psi_u(0)$. Moreover, it can be proved that $\psi'_u(x) = 0$ has no solution on $]-\infty, 0[$, and therefore that $\inf_{x<0} \psi_u(x) = \psi_u(0)$ whatever $u > 0$ is. Hence, the minimum is reached at 0 if $u \in [0, \gamma[$ and at $u - \gamma^2/u$ if $u > \gamma$. This concludes the proof.

Proof in the case $\ell > 1$. Set $x \in \mathbb{R}^\ell$ of the form $x = r\xi$ where $r > 0$ and ξ is on the unit sphere of $\mathbb{R}^\ell$. Since the function h only depends on the radius r, the minimum over $\mathbb{R}^\ell$ of $x \mapsto h(x) + \|x - u\|^2/2$ is reached in the direction $\xi_\star = u/\|u\|$. Then, finding the minimum in this direction is equivalent to find the minimum of the function ψ_u on $\mathbb{R}^+$, which yields $r_\star = 0$ if $\|u\| \leq \gamma$ and $r_\star = (1 - \gamma^2/\|u\|^2)$ otherwise. This concludes the proof.

6.3. Proof of Theorem 4.1

In this section, let $\psi : \bigcup_{m \in \mathcal{M}} \left(\{m\} \times (\mathbb{R}^\star)^{|m|}\right) \to \mathbb{R}^P$ denote the one-to-one map such that for any $m \in \mathcal{M}$, $\psi(m, .) : (\mathbb{R}^\star)^{|m|} \to \mathbb{R}^P$ is the function such that

$$\psi(m, x) = y \quad \text{with} \quad y_{m\cdot} = x \quad \text{and} \quad y_{-m\cdot} = 0 \, . \tag{21}$$

Set

$$\tilde{\mu}(x) \stackrel{\text{def}}{=} x - \frac{\sigma^2}{2} \frac{D \, \nabla g(x)}{\max\left(D, \|\nabla g(x)\|_2\right)} \, . \tag{22}$$

To make the notations easier, we denote by q the proposal distribution. Since $T = 1$, Lemma 3.1 shows that for any $m \in \mathcal{M}$ and $y \in S_m$

$$q(x, y) = \prod_{i \notin I_m} p\left(\tilde{\mu}_i(x)\right) \prod_{i \in I_m} f\left(\tilde{\mu}_i(x), y_i\right) \tag{23}$$

where p is given by Lemma 3.1 and

$$f(c, y) = (\sqrt{2\pi}\sigma)^{-1} \exp\left(-\left|y + \gamma \, \text{sign}(y) - c\right| / (2\sigma^2)\right) \, . \tag{24}$$

We start with a preliminary lemma which will be fundamental for the proofs since it allows to compare the proposal distribution q to gaussian proposals.

LEMMA 6.2. *Denote by g_ϵ the one-dimensional centered Gaussian density with variance ϵ. Set $\epsilon_1 \stackrel{\text{def}}{=} \sigma^2/2, \epsilon_2 \stackrel{\text{def}}{=} 2\sigma^2$ and*

$$k_1 \stackrel{\text{def}}{=} \exp\left(-\left(\gamma/\sigma^2 + D/2\right)^2\right) / \sqrt{2} \, , \quad k_2 \stackrel{\text{def}}{=} \exp\left(\left(\gamma/(2\sigma^2) + D/4\right)^2\right) \sqrt{2} \, .$$

(i) For any $x, y \in \mathbb{R}^P$ and any $1 \leq i \leq P$,

$$k_1 \; g_{\epsilon_1}(y_i - x_i) \leq f(\tilde{\mu}_i(x), y_i) \leq k_2 \; g_{\epsilon_2}(y_i - x_i) \; . \tag{25}$$

(ii) For any $x \in \mathbb{R}^P$ and $y \in S_m$, $q(x,y) \leq k_2^{|m|} \prod_{i \in I_m} g_{\epsilon_2}(y_i - x_i)$. Therefore, there exists a constant $C > 0$ such that for any $x, y \in \mathbb{R}^P$, $q(x,y) \leq C$.

PROOF. Let x and y be in $\mathbb{R}^P$ and $i \in \{1, \cdots, P\}$. By definition of $\tilde{\mu}$ (see 22), we have $|\tilde{\mu}_i(x) - x_i| \leq \|\tilde{\mu}(x) - x\|_2 \leq D\sigma^2/2$. Thus, on one hand,

$$\begin{aligned} |y_i - x_i| &\leq |y_i + \gamma \; \mathrm{sign}(y_i) - \tilde{\mu}_i(x)| + \gamma + |\tilde{\mu}_i(x) - x_i| \; , \\ &\leq |y_i + \gamma \; \mathrm{sign}(y_i) - \tilde{\mu}_i(x)| + \gamma + \frac{D\sigma^2}{2} \; , \end{aligned}$$

which implies $|y_i + \gamma \; \mathrm{sign}(y_i) - \tilde{\mu}_i(x)|^2 \geq \frac{1}{2} |y_i - x_i|^2 - \left(\gamma + D\sigma^2/2\right)^2$. Similarly, it holds $|y_i + \gamma \; \mathrm{sign}(y_i) - \tilde{\mu}_i(x)|^2 \leq 2 \, |y_i - x_i|^2 + 2 \left(\gamma + D\sigma^2/2\right)^2$. This concludes the proof of (i). The second statement follows trivially from (23) since $p(\tilde{\mu}_i(x)) \leq 1$.

The proof of Theorem 4.1 also requires a lower bound on the probability that a component of the proposed point will be set to zero. Such a bound is given in Lemma 6.3.

LEMMA 6.3. *Let p and $\tilde{\mu}$ be given by Lemma 3.1 and (22). It holds*

$$\inf_{z \in \mathbb{R}^P} \min_{i \notin I_{m_z}} p(\tilde{\mu}_i(z)) > 0 \; .$$

PROOF. For any $z \in \mathbb{R}^P$, $\|z - \tilde{\mu}(z)\|_2 \leq D\sigma^2/2$ by definition of $\tilde{\mu}(z)$ (see (22)). Then, $|\tilde{\mu}_i(z)| \leq D\sigma^2/2$ for any $i \notin I_{m_z}$. Hence, there exists a constant $C > 0$ such that

$$\inf_{z \in \mathbb{R}^P} \min_{i \notin I_{m_z}} \mathbb{P}(|\tilde{\mu}_i(z) + \xi| \leq \gamma) \geq C \; , \tag{26}$$

with $\xi \sim \mathcal{N}(0,1)$.

PROPOSITION 6.4. *(i) Let C be a Borel set of $\mathbb{R}^P$ such that for any $m \in \mathcal{M}$, $C \cap S_m$ is a compact set of S_m, where S_m is defined by (5). Then, C is a one-small set for the kernel P_{trunc}: there exists a positive measure $\tilde{\nu}$ on $\mathbb{R}^P$ such that $P_{trunc}(x, A) \geq \tilde{\nu}(A) \mathbf{1}_C(x)$.*

(ii) The Markov kernel P_{trunc} is psi-irreducible and aperiodic.

PROOF. (i): Let C and K be two Borel sets of $\mathbb{R}^P$ such that $\nu(K) > 0$ and for any $m \in \mathcal{M}$, $C \cap S_m$ and $K \cap S_m$ are compact subsets of S_m. Since $\mathbb{R}^P = \bigcup_{m \in \mathcal{M}} S_m$, we have

$$\inf_{x \in C} P_{trunc}(x, A) = \inf_{m \in \mathcal{M}} \inf_{x \in C \cap S_m} P_{trunc}(x, A) \; ,$$

so that it is enough to establish a minorization on the kernel for any $x \in C \cap S_{m_\star}$ whatever $m_\star \in \mathcal{M}$. Let $m_\star \in \mathcal{M}$. By definition of P_{trunc}, ψ (see (21)), q (see (23)) and ν (see (2))

$$\begin{aligned}
P_{trunc}(x, A) &\geq \int_{A\cap K} \alpha(x,y) q(x,y) \mathrm{d}\nu(y) \ , \\
&\geq \sum_{m\in\mathcal{M}} \int_{A\cap K} \alpha(x,y) \prod_{i\notin I_m} p(\tilde{\mu}_i(x)) \delta_0(\mathrm{d}y_i) \prod_{i\in I_m} f(\tilde{\mu}_i(x), y_i) \mathrm{d}y_i \ , \\
&\geq \sum_{m\in\mathcal{M}} \int_{(A\cap K)\cap S_m} \alpha(x,y) \prod_{i\notin I_m} p(\tilde{\mu}_i(x)) \delta_0(\mathrm{d}y_i) \prod_{i\in I_m} f(\tilde{\mu}_i(x), y_i) \mathrm{d}y_i \ , \\
&\geq \sum_{m\in\mathcal{M}} \prod_{i\notin I_m} p(\tilde{\mu}_i(x)) \int_{A\cap K\cap S_m} \alpha(x, \psi(m, y_m)) \prod_{i\in I_m} f(\tilde{\mu}_i(x), y_i) \mathrm{d}y_i \ , \\
&\geq \sum_{m\in\mathcal{M}} k_1^{|m|} \prod_{i\notin I_m} p(\tilde{\mu}_i(x)) \int_{A\cap K\cap S_m} \alpha(x, \psi(m, y_m)) \prod_{i\in I_m} g_{\epsilon_1}(x_i - y_i) \mathrm{d}y_i \ ,
\end{aligned}$$

where the last inequality is a consequence of Lemma 6.2(i). For any $x \in S_{m_\star}$ and $y \in S_m$, we have

$$\alpha(x,y) = 1 \wedge \frac{\pi_m(y)}{\pi_{m_\star}(x)} \frac{\prod_{i\notin I_{m_\star}} p\left(\tilde{\mu}_i(y)\right) \ \prod_{i\in I_{m_\star}} f\left(\tilde{\mu}_i(y), x_i\right)}{\prod_{i\notin I_m} p\left(\tilde{\mu}_i(x)\right) \ \prod_{i\in I_m} f\left(\tilde{\mu}_i(x), y_i\right)} \ .$$

There exists a compact set of $\mathbb{R}$ such that for any $x \in C \cap S_m$ and $y \in K \cap S_m$, $\tilde{\mu}_i(x)$ and $\tilde{\mu}_i(y)$ are in this compact for any i. Hence, A2(i) and Lemma 6.2(i) imply that there exists $\varepsilon_m > 0$ such that for any $x \in C \cap S_m$ and $y \in K \cap S_m$,

$$\alpha(x,y) \geq \varepsilon_m \ , \qquad \inf_{i\in I_m} g_{\epsilon_1}(x_i - y_i) \geq \varepsilon_m \ .$$

This yields for any $x \in C \cap S_{m_\star}$, $P_{trunc}(x, A) \geq \inf_{m\in\mathcal{M}} \varepsilon_m \int_A \mathbf{1}_K(y) \mathrm{d}\nu(y)$, thus concluding the proof.

(ii): By (Mengersen & Tweedie 1996, Lemma 1.1), the Markov chain $(X_n)_{n\geq 0}$ is psi-irreducible since for any $x, y \in \mathbb{R}^P$, $q(x,y) > 0$ and $\pi(x) > 0$ as a consequence of Lemma 6.2(i) and A1. The chain is strongly aperiodic since by Proposition 6.4(i) it possesses a one-small set with positive ν-measure, which concludes the proof.

For any measurable function $f : \mathbb{R}^p \to \mathbb{R}^+$, $P_{trunc} f : \mathbb{R}^P \to \mathbb{R}^+$ is a measurable function defined by $x \mapsto \int P_{trunc}(x, \mathrm{d}z) f(z)$. P_{trunc} is a Hastings-Metropolis kernel with proposal distribution $q(x,y)\mathrm{d}\nu(y)$ given by (23) and target distribution $\pi(y)\mathrm{d}\nu(y)$.

Fix $\beta \in (0,1)$ and set $V : \mathbb{R}^P \to [1, \infty)$, $x \mapsto c_\beta \pi^{-\beta}(x)$. Note that such a constant c_β exists under A1. Define the possible rejection region $R(x)$ by

$$R(x) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^P : \alpha(x,y) < 1\} = \{y \in \mathbb{R}^P : \pi(x) q(x,y) > \pi(y) q(y,x)\} \ .$$

We have

$$\begin{aligned}\frac{P_{trunc}V(x)}{V(x)} &\leq \int \alpha(x,y)\frac{\pi^{-\beta}(y)}{\pi^{-\beta}(x)}q(x,y)\mathrm{d}\nu(y) + \int_{R(x)} q(x,y)\mathrm{d}\nu(y)\ ,\\ &\leq \sum_{m\in\mathcal{M}} T_m(x) + \int_{R(x)} q(x,y)\mathrm{d}\nu(y)\ ,\end{aligned}$$

where

$$T_m(x) \stackrel{\text{def}}{=} \int_{\mathbb{R}^{|m|}} \alpha(x,\psi(m,z))\,\frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)}\,q(x,\psi(m,z))\,\mathrm{d}z\ . \tag{27}$$

PROPOSITION 6.5. *Assume A1 to A4 hold. For $\beta \in (0,1)$, set $V(x) \propto \pi^{-\beta}(x)$. Then*

$$\limsup_{\|x\|\to\infty} \frac{P_{trunc}V(x)}{V(x)} < 1\ .$$

The proof of Proposition 6.5 is detailed below: by Lemma 6.6, $\limsup_{\|x\|_2\to\infty} T_m(x) = 0$ for all $m \in \mathcal{M}$; and by Lemma 6.8, $\limsup_{\|x\|\to\infty} \int_{R(x)} q(x,y)\mathrm{d}\nu(y) < 1$.

LEMMA 6.6. *Assume A1, A2 and A3 hold. Then for any $m \in \mathcal{M}$, $\limsup_{\|x\|_2\to\infty} T_m(x) = 0$.*

PROOF. The proof is adapted from Jarner & Hansen (2000) and Atchadé (2006) who respectively address the geometric ergodicity of a symmetric Random Walk Hastings-Metropolis algorithm and the geometric ergodicity of MALA. Let $m \in \mathcal{M}$ and $\epsilon > 0$ be fixed. Define

$$\begin{aligned}\mathcal{B}(x_m,a) &\stackrel{\text{def}}{=} \{z\in\mathbb{R}^{|m|}, \|z-x_m\|_2 \leq a\}\ ,\\ \mathcal{C}_m(x) &\stackrel{\text{def}}{=} \{z\in\mathbb{R}^{|m|}, \pi(\psi(m,z)) = \pi(x)\}\ ,\\ \mathcal{C}_m(x,u) &\stackrel{\text{def}}{=} \{z+sn(z), |s|\leq u, z\in\mathcal{C}_m(x)\}\ ,\\ A_m(x) &\stackrel{\text{def}}{=} \{z\in\mathbb{R}^{|m|}, \pi(\psi(m,z))q(\psi(m,z),x) \geq \pi(x)q(x,\psi(m,z))\}\ ,\\ R_m(x) &\stackrel{\text{def}}{=} \mathbb{R}^{|m|}\setminus A_m(x)\ .\end{aligned}$$

We write $T_m(x) \leq T_{m,1}(x,a) + \sum_{j=2}^4 T_{m,j}(x,a,u)$ with

$$\begin{aligned}T_{m,1}(x,a) &\stackrel{\text{def}}{=} \int_{\mathcal{B}^c(x_m,a)} \alpha(x,\psi(m,z))\frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)}q(x,\psi(m,z))\mathrm{d}z\ ,\\ T_{m,2}(x,a,u) &\stackrel{\text{def}}{=} \int_{\mathcal{B}(x_m,a)\cap\mathcal{C}_m(x,u)} \alpha(x,\psi(m,z))\frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)}q(x,\psi(m,z))\mathrm{d}z\ ,\\ T_{m,3}(x,a,u) &\stackrel{\text{def}}{=} \int_{A_m(x)\cap\mathcal{B}(x_m,a)\cap\mathcal{C}_m^c(x,u)} \frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)}q(x,\psi(m,z))\mathrm{d}z\ ,\\ T_{m,4}(x,a,u) &\stackrel{\text{def}}{=} \int_{R_m(x)\cap\mathcal{B}(x_m,a)\cap\mathcal{C}_m^c(x,u)} \frac{\pi^{1-\beta}(\psi(m,z))}{\pi^{1-\beta}(x)}q(\psi(m,z),x)\mathrm{d}z\ .\end{aligned}$$

We prove that there exist positive constants C, M such that $\sup_{\|x\|\geq M} T_m(x) \leq C\epsilon$. Since ϵ is arbitrarily small, this yields the lemma. Note that for any $z \in \mathbb{R}^{|m|}$,

$$\alpha(x, \psi(m,z)) \frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)} \leq \left(\frac{q(\psi(m,z),x)}{q(x,\psi(m,z))} \right)^{\beta} . \tag{28}$$

Control of $T_{m,1}$: By (28), $T_{m,1}(x,a) \leq \int_{\mathcal{B}^c(x_m,a)} q(x,\psi(m,z))^{1-\beta} q(\psi(m,z),x)^{\beta} \mathrm{d}z$. By (23) and Lemma 6.2, there exists a constant $C > 0$ such that

$$\begin{aligned} T_{m,1}(x,a) &\leq C k_2^{|m|(1-\beta)} \int_{\mathcal{B}^c(x_m,a)} \prod_{i \in I_m} g_{\epsilon_2}(x_i - y_i)^{1-\beta} \mathrm{d}y_i \\ &\leq C k_2^{|m|(1-\beta)} \int_{\mathcal{B}^c(0,a)} \prod_{i \in I_m} g_{\epsilon_2}(y_i)^{1-\beta} \mathrm{d}y_i \ . \end{aligned}$$

Therefore, there exists $a > 0$ such that $\sup_{x \in \mathbb{R}^P} T_{m,1}(x,a) \leq \epsilon$.

Control of $T_{m,2}$: By (28), $T_{m,2}(x,a,u) \leq \int_{\mathcal{B}(x_m,a) \cap \mathcal{C}_m(x,u)} q(x,\psi(m,z))^{1-\beta} q(\psi(m,z),x)^{\beta} \mathrm{d}z$. By A3, the Lebesgue measure of $\mathcal{B}_m(x,a_m) \cap \mathcal{C}_m(x,u)$ can be made arbitrarily small - independently of $x \in \mathbb{R}^P$ - when u is small enough (see (Jarner & Hansen 2000, Proof of Theorem 4.1) for details). Therefore, since q is bounded (see Lemma 6.2(ii)), there exists $u > 0$ such that $\sup_{x \in \mathbb{R}^P} T_{m,2}(x,a,u) \leq \epsilon$.

Control of $T_{m,3}$: Set $d_r(u) \stackrel{\text{def}}{=} \sup_{\|x\|_2 \geq r} \pi(x + u\, n(x))/\pi(x)$. By A3, choose r large enough so that $(d_{r-u}(u))^{1-\beta} \vee (d_r(u))^{\beta} \leq \epsilon$. By A1 and A2(i), $\sup_{z \in \mathcal{B}(0,r)} \pi(\psi(m,z))^{-\beta} < \infty$, so that by Lemma 6.2(ii)

$$\sup_{x \in \mathbb{R}^P} \int_{A_m(x) \cap \mathcal{B}(x_m,a) \cap \mathcal{C}_m^c(x,u) \cap \mathcal{B}(0,r)} q(x,\psi(m,z)) \pi^{-\beta}(\psi(m,z)) \mathrm{d}z < \infty \ .$$

A2(ii) implies that

$$\limsup_{\|x\| \to \infty} \int_{A_m(x) \cap \mathcal{B}(x_m,a) \cap \mathcal{C}_m^c(x,u) \cap \mathcal{B}(0,r)} \frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)} q(x,\psi(m,z)) \mathrm{d}z = 0 \ .$$

Moreover, by definition of $A_m(x)$, for any $z \in A_m(x)$ it holds

$$\frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)} q(x,\psi(m,z)) \leq \frac{\pi^{1-\beta}(\psi(m,z))}{\pi^{1-\beta}(x)} q(\psi(m,z),x) \ ;$$

by Lemma 6.2(ii), there exists a constant C such that for any $x \in \mathbb{R}^P$ and $z \in A_m(x)$

$$\frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)} q(x,\psi(m,z)) \leq C \left(\frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)} \wedge \frac{\pi^{1-\beta}(\psi(m,z))}{\pi^{1-\beta}(x)} \right) .$$

This yields

$$\begin{aligned} &\int_{A_m(x) \cap \mathcal{B}(x_m,a) \cap \mathcal{C}_m^c(x,u) \cap \mathcal{B}^c(0,r)} \frac{\pi^{-\beta}(\psi(m,z))}{\pi^{-\beta}(x)} q(x,\psi(m,z)) \mathrm{d}z \\ &\leq C \left(\sup_{z \in \mathcal{C}_m^c(x,u) \cap \mathcal{B}^c(0,r)} \frac{\pi^{\beta}(x)}{\pi^{\beta}(\psi(m,z))} \wedge \sup_{z \in \mathcal{C}_m^c(x,u) \cap \mathcal{B}^c(0,r)} \frac{\pi^{1-\beta}(\psi(m,z))}{\pi^{1-\beta}(x)} \right) . \end{aligned}$$

Let $z \in \mathcal{C}_m^c(x,u) \cap \{z : \pi(\psi(m,z)) < \pi(x)\}$ and set $\bar{x} \stackrel{\text{def}}{=} \psi(m, y_m) - u\ n(\psi(m, y_m))$. By A2(i), $h : s \mapsto \pi(\psi(m,z) - s\ n(\psi(m,z))) - \pi(x)$ is continuous, and by definition of $\mathcal{C}_m^c(x,u)$, $h(s) \neq 0$ for any $0 \leq s \leq u$. Since $h(0) < 0$ (we assumed that $\pi(\psi(m,z)) < \pi(x)$), this implies that $h(u) < 0$ i.e. $\pi(\bar{x}) \leq \pi(x)$. Then,

$$\sup_{z \in \mathcal{C}_m^c(x,u) \cap \mathcal{B}^c(0,r)} \frac{\pi(\psi(m,z))}{\pi(x)} = \frac{\pi(\psi(m,z))}{\pi(\bar{x})} \frac{\pi(\bar{x})}{\pi(x)} \leq \frac{\pi(\psi(m,z))}{\pi(\bar{x})} \leq d_{r-u}(u)\ .$$

If $z \in \mathcal{C}_m^c(x,u) \cap \{z : \pi(\psi(m,z)) \geq \pi(x)\}$, we set $\bar{x} \stackrel{\text{def}}{=} \psi(m, y_m) + u\ n(\psi(m, y_m))$ and obtain similarly that $\pi(x)/\pi(\psi(m,z)) \leq d_r(u)$. Hence, we established that

$$\sup_{z \in \mathcal{C}_m^c(x,u) \cap \mathcal{B}^c(0,r)} \frac{\pi(\psi(m,z))}{\pi(x)} \leq d_r(u)\ .$$

As a conclusion, there exist constants C, M such that $\sup_{\|x\| \geq M} T_{m,3}(x,a,u) \leq C\epsilon$.

Control of $T_{m,4}$ Following the same lines as for the control of $T_{m,3}(x,a,u)$, it can be shown that there exist constants C, M such that $\sup_{\|x\| \geq M} T_{m,4}(x,a,u) \leq C\epsilon$.

LEMMA 6.7. *Assume A1, A3 and A4 hold. Let u, b, ϵ, R be given by A4 and $W_m(x)$ be defined by* (15). *There exists $r > R$ such that for any $m \in \mathcal{M}$ and $x \in S_m \cap \{\|x\|_2 \geq r\}$, $W_m(x) \subset \{y \in \mathbb{R}^{|m|}, \alpha(x, \psi(m,y)) = 1\}$.*

PROOF. The proof is adapted from Jarner & Hansen (2000). Let $m \in \mathcal{M}$ and $x \in S_m$ such that $\|x\| \geq r$ for some $r > R$ to be fixed later (the constant R is given by A4). We first prove that there exists a positive constant C_b such that

$$\frac{\pi(x)}{\pi(x - un(x))} \leq C_b \leq \inf_{z \in \mathcal{B}(x_m, b)} \frac{q(\psi(m,z), x)}{q(x, \psi(m,z))}\ . \tag{29}$$

By (23), Lemma 6.2(i) and Lemma 6.3, there exist $C, C_b > 0$ - independent of $x \in S_m$ - such that

$$\inf_{z \in \mathcal{B}(x_m,b)} \frac{q(\psi(m,z),x)}{q(x,\psi(m,z))} \geq C^{P-|m|} k_1^{|m|} k_2^{-|m|} \inf_{z \in \mathcal{B}(x_m,b)} \prod_{i \in I_m} \frac{g_{\epsilon_1}(x_i - z_i)}{g_{\epsilon_2}(x_i - z_i)} \geq C_b\ .$$

By A3, we can choose r large enough so that $\pi(x)/\pi(x - un(x)) \leq C_b$. This yields (29).

Let $z \in W_m(x)$. Then, $\|z - x_m\|_2 \leq b$ so that $z \in \mathcal{B}(x_m, b)$. Hence, by (29), $q(\psi(m,z),x)/q(x,\psi(m,z)) \geq C_b$. In addition,

$$\frac{\pi(\psi(m,z))}{\pi(x)} = \frac{\pi(\psi(m,z))}{\pi(x - un(x))} \frac{\pi(x - un(x))}{\pi(x)} \geq \frac{\pi(\psi(m,z))}{\pi(x - un(x))} \frac{1}{C_b} \geq \frac{1}{C_b}\ ,$$

where in the last inequality we used A4. Hence,

$$\alpha(x, \psi(m,z)) = \frac{\pi(\psi(m,z))}{\pi(x)} \frac{q(\psi(m,z),x)}{q(x,\psi(m,z))} \geq 1$$

thus showing the lemma.

LEMMA 6.8. *Assume A1 to A4 hold. Then* $\limsup_{\|x\|_2 \to \infty} \int_{R(x)} q(x,y)\mathrm{d}\nu(y) < 1$.

PROOF. Set $A_m(x) \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{|m|}, \alpha(x, \psi(m,z)) = 1\}$. By definition of $\mathrm{d}\nu$, by Lemma 6.2(i) and by Lemma 6.3, there exists a constant $C > 0$ such that

$$\begin{aligned}
1 - \int_{R(x)} q(x,y)\mathrm{d}\nu(y) &= \sum_{m \in \mathcal{M}} \int_{A_m(x)} q(x, \psi(m,z))\mathrm{d}z \ , \\
&\geq \sum_{m \in \mathcal{M}} k_1^{|m|} \prod_{i \notin I_m} p(\tilde{\mu}_i(x)) \int_{A_m(x)} \prod_{i \in I_m} g_{\epsilon_1}(x_i - y_i)\mathrm{d}y_i \\
&\geq k_1^{|m_x|} \prod_{i \notin I_{m_x}} p(\tilde{\mu}_i(x)) \int_{A_{m_x}(x)} \prod_{i \in I_{m_x}} g_{\epsilon_1}(x_i - y_i)\mathrm{d}y_i \\
&\geq C\, k_1^{|m_x|} \int_{A_{m_x}(x)} \prod_{i \in I_{m_x}} g_{\epsilon_1}(x_i - y_i)\mathrm{d}y_i \ .
\end{aligned}$$

By Lemma 6.7, for any x large enough,

$$1 - \int_{R(x)} q(x,y)\mathrm{d}\nu(y) \geq C\, k_1^{|m_x|} \int_{W_{m_x}(x)} \prod_{i \in I_{m_x}} g_{\epsilon_1}(x_i - y_i)\mathrm{d}y_i \ .$$

where $W_m(x)$ is defined in (15). We have

$$\int_{W_{m_x}(x)} \prod_{i \in I_{m_x}} g_{\epsilon_1}(x_i - y_i)\mathrm{d}y_i = \int_{W_{m_x}(x) - x_{m_x}} \prod_{i \in I_{m_x}} g_{\epsilon_1}(y_i)\mathrm{d}y_i \ , \tag{30}$$

where $A - x \stackrel{\text{def}}{=} \{z, z + x \in A\}$. Observe that

$$W_{m_x}(x) - x_{m_x} = \{-un(x_{m_x}) - s\xi; 0 < s < b - u, \xi \in \mathbb{R}^{|m_x|}, \|\xi\|_2 = 1, \|\xi - n(x_{m_x})\|_2 \leq \epsilon\} \ ,$$

so that the integrals in (30) depend on x only through m_x. Since $\mathcal{M}$ is finite, there exists a constant $C' > 0$ independent of x such that

$$\int_{W_{m_x}(x) - x_{m_x}} g_{\epsilon_1}(y_{m_x})\mathrm{d}y_{m_x} \geq C' \ .$$

A. Example of a target density satisfying assumptions A1 to A4

Let $v > 0$ and $\lambda \geq 0$. We prove that the density π defined on $\mathbb{R}^P$ by

$$\pi(x) \propto w_m\, c_\lambda^{-|m|} \exp\left(-\frac{1}{2\tau}\|Y - Gx\|_2^2 - \lambda\|x\|_1 - v\|x\|_2^2\right) \ , \tag{31}$$

for any $x \in S_m$, satisfies the conditions A1 to A4. Note that since $T = 1$, $\|x\|_{2,1} = \|x\|_1$.

For any $m \in \mathcal{M}$ and $x \in S_m$, we have $\pi(x) \propto \exp(-g(x) - \bar{g}(x))$ with

$$g(x) = \frac{1}{2\tau}\|Y - Gx\|_2^2 \quad \text{and} \quad \bar{g}(x) = \lambda\|x\|_1 + v\|x\|_2^2 - \log\left(w_m c_\lambda^{-|m|}\right) \ .$$

Hence A1 is satisfied. It is clear that for any $m \in \mathcal{M}$, the restriction of π_m is continuous. Moreover, for any $x \in \mathbb{R}^P$,

$$\pi(x) \leq C \max_{|m| \in \{1,\cdots,P\}} c_\lambda^{-|m|} \exp\left(-\lambda \|x\|_1 - v\|x\|_2^2\right)$$

and as $\lambda \geq 0$ and $v > 0$, the RHS tends to zero when $\|x\|_2 \to \infty$. Therefore, A2 is satisfied. A3 and A4 are proved in Lemma A.1 and Lemma A.2.

LEMMA A.1. *Let π be given by* (31). *Then,* $\lim_{r\to\infty} \sup_{\|x\|_2 \geq r} \pi_m(x + s\, n(x))/\pi_m(x) = 0$ *for any $m \in \mathcal{M}$ and $s > 0$.*

PROOF. Let $m \in \mathcal{M}$, $s > 0$ and $x \in S_m$. Since $x + s\ n(x) \in S_m$, we have

$$\frac{\pi_m(x + sn(x))}{\pi_m(x)} = \exp\left(\frac{1}{2\tau}\left(\|Y - Gx\|_2^2 - \|Y - G(x + sn(x))\|_2^2\right)\right) \times \exp\left(\lambda\left(\|x\|_1 - \|x + sn(x)\|_1\right) + v\left(\|x\|_2^2 - \|x + sn(x)\|_2^2\right)\right) .$$

First, $\|x\|_1 - \|x + sn(x)\|_1 = -s\|x\|_1/\|x\|_2 \leq -s$. Moreover,

$$\|Y - Gx\|_2^2 - \|Y - G(x + sn(x))\|_2^2 = -s^2\|Gn(x)\|_2^2 + 2s\Big\langle Y - Gx, Gn(x)\Big\rangle .$$

And finally,

$$\|x\|_2^2 - \|x + sn(x)\|_2^2 = -2s\Big\langle x, n(x)\Big\rangle - s^2\|n(x)\|_2^2 \leq -2s\|x\|_2 .$$

This implies that for any x,

$$\begin{aligned}
\frac{\pi_m(x + sn(x))}{\pi_m(x)} &\leq \exp\left(\frac{s}{\tau}\Big\langle Y - Gx, Gn(x)\Big\rangle\right)\exp\left(-2sv\|x\|_2\right) , \\
&\leq \exp\left(\frac{s}{\tau}\Big\langle Y, Gn(x)\Big\rangle\right)\exp\left(-\frac{s\|x\|_2}{\tau}\|Gn(x)\|_2^2\right)\exp\left(-2sv\|x\|_2\right) , \\
&\leq \exp\left(\frac{s}{\tau}\sup_{\|z\|_2=1}\Big\langle Y, Gz\Big\rangle\right)\exp\left(-2sv\|x\|_2\right) .
\end{aligned}$$

As $v > 0$, the RHS tends to zero as $\|x\|_2 \to \infty$, uniformly for $\|x\|_2 \geq r$.

LEMMA A.2. *The target density π defined by* (31) *satisfies the condition A4.*

PROOF. Let $b > 0$ and $u \in (0, b)$ be fixed. Let $\epsilon > 0$ be such that

$$\epsilon \leq \inf_{z,\|z\|=1} \|Gz\|_2^2 / \sup_{z,\|z\|=1} \|G'Gz\|_2 \quad \text{and} \quad \epsilon < 1 . \tag{32}$$

Let $m \in \mathcal{M}$ and $x \in S_m$ such that $\|x\|_2 \geq b$. Set $x_\star \stackrel{\text{def}}{=} x - u\, n(x)$, and $y \stackrel{\text{def}}{=} x_\star - s\xi \in S_m$ such that $y_m \in W_m(x)$, where by definition of $W_m(x)$ in (15), $s \in (0, b-u)$ and $\xi \in S_m$ is such that $\|\xi\|_2 = 1$ and $\|\xi - n(x)\| \leq \epsilon$. First, note that when $\|x\|_2 > b$, $x_\star \in S_m$, and

$$\frac{\pi_m(y)}{\pi_m(x_\star)} = \exp\left(-\frac{1}{2\tau}\left(\|Y - G(x_\star - s\xi)\|_2^2 - \|Y - Gx_\star\|_2^2\right) - \lambda\left(\|x_\star - s\xi\|_1 - \|x_\star\|_1\right)\right) \times \exp\left(-v\left(\|x_\star - s\xi\|_2^2 - \|x_\star\|_2^2\right)\right) .$$

Moreover,

$$\begin{aligned}\|x_\star - s\xi\|_1 - \|x_\star\|_1 &= \|x_\star - s\ n(x_\star) + s\ n(x_\star) - s\xi\|_1 - \|x_\star\|_1\ ,\\ &\leq \|x_\star\|_1 \left(1 - \frac{s}{\|x_\star\|_2}\right) + s\|n(x_\star) - \xi\|_1 - \|x_\star\|_1\ ,\\ &\leq -s\frac{\|x_\star\|_1}{\|x_\star\|_2} + s\epsilon\frac{\|n(x_\star) - \xi\|_1}{\|n(x_\star) - \xi\|_2}\end{aligned}$$

since $y \in W_m(x)$. By equivalence of the norm, and as $\epsilon < 1$, there exists a constant C such that

$$\sup_{x\in S_m} \|x_\star - s\xi\|_1 - \|x_\star\|_1 \leq C\ .$$

Define $\phi : s \mapsto \|Y - G(x_\star - s\ \xi)\|_2^2$; ϕ is differentiable on $\mathbb{R}$, and

$$\begin{aligned}\phi'(s) &= 2\Big\langle Y - Gx_\star + sG\xi, G\xi\Big\rangle = 2\Big\langle Y, G\xi\Big\rangle - 2\Big\langle G\xi, Gx_\star\Big\rangle + 2s\|G\xi\|_2^2\ ,\\ &= 2\Big\langle Y, G\xi\Big\rangle - 2\|x_\star\|_2 \left(\Big\langle \xi - n(x_\star), G'Gn(x_\star)\Big\rangle + \|Gn(x_\star)\|_2^2\right) + 2s\|G\xi\|_2^2\ .\end{aligned}$$

By definition of ϕ, $\|Y - G(x_\star - s\xi)\|_2^2 - \|Y - Gx_\star\|_2^2 = \phi(s) - \phi(0)$. By the mean value theorem, there exists $\tau \in [0, s]$ such that

$$\begin{aligned}\phi(s) - \phi(0) &= s\ \phi'(\tau)\\ &\leq 2s\tau\|G\xi\|_2^2 + 2s\Big\langle Y, G\xi\Big\rangle - 2s\|x_\star\|_2 \left(\|Gn(x_\star)\|_2^2 + \Big\langle \xi - n(x_\star), G'Gn(x_\star)\Big\rangle\right)\ ,\\ &\leq 2s\tau\|G\xi\|_2^2 + 2s\Big\langle Y, G\xi\Big\rangle - 2s\|x_\star\|_2 \left(\inf_{z,\|z\|=1} \|Gz\|_2^2 - \epsilon \sup_{z,\|z\|=1} \|G'Gz\|_2\right)\ .\end{aligned}$$

Therefore, by (32), there exists a constant $C > 0$ independent of x such that,

$$\sup_{x\in S_m} \|Y - G(x_\star - s\xi)\|_2^2 - \|Y - Gx_\star\|_2^2 \leq C\ .$$

And finally,

$$\begin{aligned}&\|x_\star - s\xi\|_2^2 - \|x_\star\|_2^2\ ,\\ &= \|x_\star - s\ n(x_\star) + s\ n(x_\star) - s\xi\|_2^2 - \|x_\star\|_2^2\ ,\\ &= \|x_\star - s\ n(x_\star) + s\ n(x_\star) - s\xi\|_2^2 - \|x_\star - s\ n(x_\star)\|_2^2 + \|x_\star - s\ n(x_\star)\|_2^2 - \|x_\star\|_2^2\ ,\\ &= 2s\Big\langle n(x_\star) - \xi, x_\star - sn(x_\star)\Big\rangle + s^2\|n(x_\star) - \xi\|_2^2 - 2s\|x_\star\|_2 + s^2\ ,\\ &\leq s^2 - 2s\|x_\star\|\ (1 - \epsilon)\ .\end{aligned}$$

Therefore, as $\epsilon < 1$,

$$\|x_\star - s\xi\|_2^2 - \|x_\star\|_2^2 \underset{\|x\|_2\to\infty}{\longrightarrow} -\infty\ .$$

Therefore, there exists $\epsilon > 0$ and $R > 0$ such that for any $\|x\|_2 \geq R$ and any $y \in S_m$ with $y_m \in W_m(x)$, $\pi_m(y)/\pi_m(x_\star) \geq 1$.

B. Appendix for the numerical experiments

B.1. The RJMCMC algorithm

The RJMCMC algorithm proposed by Green (1995) is designed to sample from distributions defined on a parameter space which is a union of subspaces. Here we use it to sample from absolutely continuous distributions with respect to $\mathrm{d}\nu$ defined in (2), by iteratively sampling a binary vector $m \in \{0,1\}^P$ and a point $X \in S_m$. Here we denote by $\pi\mathrm{d}\nu$ the target distribution.

Each iteration of this algorithm is decomposed into two steps: *(i)* first propose a new binary vector m', in general "close" to the current one m, according to a transition probability $j(m,\cdot)$, and *(ii)* then compute a new point $X' \in S_{m'}$ from the current one X in a reversible way. This second step introduces an auxiliary random variable u (resp. u'), sampled from a proposal distribution q specified by the user to balance the components that are null in X (resp. X') and not in X' (resp. X). Here we choose $q(u) = \mathcal{N}_T(u; 0, \sigma_{rj}^2 I_T)$, for some $\sigma_{rj} > 0$ if $u \neq 0_T$ and $q(0_T) = 1$ where 0_T is the null $1 \times T$ matrix.

In this work, the transition probability $j(m, m')$ is defined as follows: m' is obtained from the current m by adding, or by deleting, or by both adding and removing one of the active components (ones) of m, or by keeping the same m. The probability of these four strategies is uniform, except if $|m| = 0$ (only the adding move can be chosen), or if $|m| = P$ (only the deleting move can be chosen). The proposed m' is then chosen uniformly between all the binary vectors that can be reached by doing the selected move from the current m.

Then, to choose the new point $X' \in S_{m'}$, we use the following dynamics:

- If the component k is added, *i.e.* if $m'_i = m_i$ for any $i \neq k$ and $m'_k = 1$, $m_k = 0$, then $X'_{m\cdot} = X_{m\cdot}$, $X'_{k\cdot} = u \sim \mathcal{N}_T(0, \sigma_{rj}^2 I_T)$, and the other components $X'_{j\cdot}$ are null. And $u' = 0_T$.

- If one of the active rows k is deleted , *i.e.* if $m'_i = m_i$ for any $i \neq k$ and $m'_k = 0$, $m_k = 1$, then $X'_{m'\cdot} = X_{m'\cdot}$ and the othe components $X'_{j\cdot}$ are null. This corresponds to setting $u' = X_{k\cdot}$ and $u = 0_T$.

- If an active row is deleted (say k) and a new active row is inserted (say ℓ) then $X'_{j\cdot} = X_{j\cdot}$ for $j \notin \{k, \ell\}$, $X'_{k\cdot} = 0$, $X'_{\ell\cdot} = u \sim \mathcal{N}_T(0, \sigma_{rj}^2 I_T)$ and $u' = X_{k\cdot}$.

- Finally, if $m' = m$, $X'_{m\cdot} - X_{m\cdot}$ are $|m|$ i.i.d. random variables $\mathcal{N}_T(0, \sigma_{rj}^2 I_T)$, and the other components $X'_{j\cdot}$ are null. In this case, we set $u = u' = 0_T$.

The candidate $(m', X'_{m'\cdot})$ is the accepted with probability

$$\alpha((m, X_{m\cdot}), (m', X'_{m'\cdot})) = 1 \wedge \left[\frac{\pi(X')j(m', m)q(u')}{\pi(X)j(m, m')q(u)}\right] .$$

Note that in general, a Jacobian term takes part in the acceptance probability of the RJMCMC. For the dynamics chosen here, this Jacobian term is equal to one.

B.2. Computational aspects

This section provides a method to compute the probability $p(c)$ defined in Lemma 3.1. Let $c \in \mathbb{R}^T$ and $\gamma > 0$. $p(c)$ is given by

$$p(c) = \mathbb{P}\left(\|c + \xi\|_2 \leq \gamma\right) = \mathbb{P}\left(\left\|\frac{c+\xi}{\sigma}\right\|_2^2 \leq \frac{\gamma^2}{\sigma^2}\right) ,$$

where $\xi \sim \mathcal{N}_T(0, \sigma^2 I_T)$. In this case, $\|(c+\xi)/\sigma\|_2^2$ follows a non-centered chi-squared distribution with noncentrality parameter given by $\|c\|_2^2/\sigma^2$. Then, $p(c)$ can be approximated by the Matlab ncx2cdf function with parameters $\left(\gamma^2/\sigma^2, T, \|c\|_2^2/\sigma^2\right)$. To obtain faster computations, the following approximation introduced in (Johnson et al. 1995, Section 8) may also be used:

$$\mathbb{P}\left(\left\|\frac{c+\xi}{\sigma}\right\|_2^2 \leq x\right) \approx \Phi\left(\frac{x^u(T+\ell)^{-u} - [1 + uv(u - 1 - 0.5(2-u)wv)]}{u\sqrt{2v(1+wv)}}\right) ,$$

where Φ is the cumulative distribution function of a standard Gaussian distribution and

$$\ell \stackrel{\text{def}}{=} \|c/\sigma\|_2^2, \quad u \stackrel{\text{def}}{=} 1 - \frac{2}{3}\frac{(T+\ell)(T+3\ell)}{(T+2\ell)^2}, \quad v \stackrel{\text{def}}{=} \frac{T+2\ell}{(T+\ell)^2}, \quad w \stackrel{\text{def}}{=} (u-1)(1-3u) .$$

References

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

Beauchamp, J. & Mitchell, T. (1988), 'Bayesian variable selection in linear regression (with discussion)', *J. Amer. Statist. Assoc.* **83**, 1023–1036.

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-tresholding algorithm for linear inverse problems', *SIAM J. Imaging Sci.* **2**(1), 183–202.

Bickel, P., Ritov, Y. & Tsybakov, A. (2009), 'Simultaneous analysis of Lasso and Dantzig selector', *Ann. Statist.* **37**, 1705–1732.

Breiman, L. (1992), 'The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error', *J. Amer. Statist. Assoc.* **87**, 738–754.

Brooks, S., Giudici, P. & Roberts, G. (2003), 'Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions', *J. of Royal Statist. Soc. B.* **61**(1), 3–39.

Brown, P., Fearn, T. & Vannucci, M. (2001), 'Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem', *J. Amer. Statist. Assoc.* **96**(454), 398–408.

Bunea, F., Tsybakov, A. & Wegkamp, M. (2007), 'Sparsity oracle inequalities for the LASSO', *Electron. J. Statist.* **1**, 169–194.

Carlin, B. & Chib, S. (1995), 'Bayesian model choice via Markov chain Monte Carlo methods', *J. of Royal Statist. Soc. B.* **157**, 473–484.

Caron, F. & Doucet, A. (2008), Sparse Bayesian nonparametric regression, *in* 'Proceedings of the 25th International Conference on Machine Learning (ICML'2008)', pp. 88–95.

Casella, G. & Park, T. (2008), 'Bayesian LASSO', *J. Amer. Statist. Assoc.* **103**(482), 681–686.

Dalalyan, A. & Tsybakov, A. (2012), 'Sparse regression learning by aggregation and Langevin Monte-Carlo', *J. Comput. System Sci.* **78**(5), 1423–1443.

Dellaportas, P., Forster, J. & Ntzoufras, I. (2002), 'On Bayesian model and variable selection using MCMC', *Stat. Comput.* **12**, 27–36.

George, E. & McCulloch, R. (1993), 'Variable selection via Gibbs sampling', *J. Amer. Statist. Assoc.* **88**(423), 881–889.

Green, P. (1995), ' Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–723.

Griffin, J. & Brown, P. (2011), 'Bayesian hyper-lassos with non-convex penalization', *Aust. N. Z. J. Stat.* **53**(4), 423–442.

Ishwaran, H. & Rao, J. (2005), 'Spike and slab variable selection: frequentist ans bayesian strategies', *Ann. Statist.* **33**(2), 730–773.

Jarner, S. & Hansen, E. (2000), 'Geometric ergodicity of Metropolis algorithms', *Stoch. Proc. Appl.* **85**(2), 341–361.

Ji, C. & Schmidler, S. (2013), 'Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection', *J. Comput. Graph. Statist.* **22**(3), 708–728.

Johnson, N., Kotz, S. & Balakrishnan, N. (1995), *Continuous Univariate Distributions, Volume 2*, Wiley Series in Probability and Statistics.

Karagiannis, G. & Andrieu, C. (2013), 'Annealed Importance Sampling Reversible Jump MCMC Algorithms', *J. Comput. Graph. Statist.* **22**(3), 623–648.

Lamnisos, D., Griffin, J. & Steel, M. (2013), 'Adaptive Monte Carlo for Bayesian Variable Selection in Regression Models', *J. Comput. Graph. Statist.* **22**(3), 729–748.

Lucka, F. (2012), 'Fast Markov chain Monte Carlo sampling for sparse Bayesian inference in high-dimensional inverse problems using L1-type priors', *Inverse Problems* **28**(12).

Malsiner-Walli, G. & Wagner, H. (2011), 'Comparing spike and slab priors for Bayesian variable selection', *Austrian Journal of Statistics* **40**(4), 241–264.

Mengersen, K. & Tweedie, R. (1996), 'Rates of convergence of the Hastings and Metropolis algorithms', *Ann. Statist.* **24**(1), 101–121.

Meyn, S. P. & Tweedie, R. L. (1993), *Markov Chains and Stochastic Stability*, Springer, London.

Neal, P. & Roberts, G. (2006), 'Optimal scaling for partially updating MCMC algorithms', *Ann. Appl. Probab.* **16**(2), 475–515.

Nott, D. & Kohn, R. (2005), 'Adaptive sampling for Bayesian variable selection', *Biometrika* **92**(4), 747–763.

O'Hara, R. & Sillanpää, M. (2009), 'A review of Bayesian variable selection methods: what, how and which', *Bayesian Anal.* **4**(1), 85–117.

Parikh, N. & Boyd, S. (2013), 'Proximal algorithms', *Foundation and trends in optimization* **1**(3), 123–231.

Pereyra, M. (2013), 'Proximal Markov chain Monte Carlo algorithms', *arXiv:1306.0187* .

Petralias, A. & Dellaportas, P. (2013), 'An MCMC model search algorithm for regression problems', *J. Statist. Comput. Simulation* **83**(9), 1722–1740.

Rigollet, P. & Tsybakov, A. B. (2012), 'Sparse Estimation by Exponential Weighting', *Statist. Sci.* **27**(4), 558–575.

Roberts, G. & Rosenthal, J. (2006), 'Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains', *Ann. Appl. Probab.* **16**(4), 2123–2139.

Roberts, G. & Tweedie, R. (1996), 'Exponential convergence of Langevin distributions and their discrete approximations', *Bernoulli* **2**(4), 341–363.

Schäfer, C. & Chopin, N. (2013), 'Sequential Monte Carlo on large binary sampling spaces', *Stat. Comput.* **23**(2), 163–184.

Shi, M. & Dunson, D. (2011), 'Bayesian Variable Selection via Particle Stochastic Search', *Stat Probab Lett.* **81**(2), 283–291.

Siedenburg, K. (2012), Persistent Empirical Wiener Estimation With Adaptive Threshold Selection For Audio Denoising, *in* 'Proceedings of the 9th Sound and Music Computing Conference', pp. 426–433.

Tan, X., Li, J. & Stoica, P. (2010), Efficient sparse Bayesian learning via Gibbs sampling, *in* 'International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 3634–3637.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the Lasso', *J. R. Statist. Soc. B* **58**(1), 267–288.

Van de Geer, S. (2009), 'High-dimensional generalized linear models and the LASSO', *Ann. Statist.* **37**, 1705–1732.

West, M. (2003), 'Bayesian Factor Regression Models in the "Large p, Small n" Paradigm', *Bayesian Statistics* **7**, 723–732.

Wipf, D., Rao, B. & Nagarajan, S. (2011), 'Latent Variable Bayesian Models for Promoting Sparsity', *IEEE Trans. Inform. Theory* **57**(9), 6236–6255.