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[*], Gersende Fort[*], Sylvain Le Corff[†], Eric Moulines[*]

September 11, 2015

Abstract

This paper introduces a new Markov Chain Monte Carlo method for Bayesian variable selection in high dimensional settings. The algorithm is a Hastings-Metropolis sampler with a proposal mechanism which combines a Metropolis Adjusted Langevin (MALA) step to propose local moves associated with a shrinkage-thresholding step allowing to propose new models. The geometric ergodicity of this new trans-dimensional Markov Chain Monte Carlo sampler is established. An extensive numerical experiment, on simulated and real data, is presented to illustrate the performance of the proposed algorithm in comparison with some more classical trans-dimensional algorithms.

1 Introduction

We focus on variable selection in regression problems: the objective is to explain a response variable with a (possibly very) large number of explanatory variables, which can be either discrete or continuous. In many applications, it is known that only a small fraction of explanatory variables explains a large fraction of the observations, and using this information is crucial for inference. Variable selection is particularly challenging in high dimensional settings.

A variety of algorithms to explore the collection of models and criteria for selecting among competing models has been proposed. In the Bayesian framework, the variable selection problem is transformed into posterior inference: rather than searching a highly hypothetical "best" model, Bayesian analysis attempts to estimate the joint posterior distribution of the collection of all subsets of parameters. In high dimension, this aim is often overly ambitious: estimating the marginal posterior probability that a variable should be included in the model is already challenging.

In the last three decades, Markov Chain Monte Carlo (MCMC) methods have been the most commonly used computational procedures to sample posterior distributions [1]. An early attempt to perform variable selection is the Reversible Jump MCMC (RJMCMC) introduced in [2]. RJMCMC is a trans-dimensional sampler which produces a Markov chain evolving between spaces of different dimensions. The dimension of the sample varies at each iteration as active (nonzero) 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 trans-dimensional moves. RJMCMC requires ingenuity in designing appropriate jumping rules to produce computationally efficient and theoretically effective methods. Despite many attempts [3, 4], this algorithm is prone to fail when the dimension of the parameter space is large (as illustrated in our numerical section).

[5] considers 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, sampling jointly all models is of course out of reach. A more efficient algorithm, the Metropolized Carlin and Chib (MCC), simultaneously proposed by [6, 7] and later improved by [8], does not require to sample from the whole collection of models and therefore can be implemented in practice. The mixing of this algorithm depends critically on the specification of pseudo-priors, which requires also a fair amount of tuning.

[1]LTCI, Telecom ParisTech & CNRS, 46 rue Barrault 75634 Paris Cedex 13 France
[2]Laboratoire de Mathématiques, Université Paris-Sud and CNRS, UMR 8628, Orsay, France

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 (nonzero) components of the regression vector. The objective is to estimate probabilities of activation for each regression parameter. In [9] 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 [10, 11]. Samples from the posterior distribution of the models are obtained in [12] and in [13] with particle filters.

In this paper, we introduce a novel algorithm, the Shrinkage-Thresholding Metropolis-Adjusted Langevin Algorithm (STMALA) to perform sparse regression in high dimensional models. This algorithm might be seen as a trans-dimensional MCMC method relying on the MALA algorithm (see [14]). The proposal distribution in the STMALA algorithm goes as follows:

- compute a noisy gradient step of the logarithm of the smooth part of the target distribution;
- apply a shrinkage-thresholding operator to ensure sparsity and to shrink values of the regression parameters toward zero;
- use an accept-reject step to guarantee the convergence to the correct target distribution.

Each iteration of the STMALA algorithm may be seen as a randomized version of the Shrinkage-Thresholding algorithm (see [15]) to guide variable selection. The Shrinkage-Thresholding algorithm (and its accelerated version FISTA) is one of the most effective method to solve sparse inverse problems. Our intuition is that a single iteration of the Shrinkage-Thresholding algorithm (with some additional noise added to ensure irreducibility) is a sensible way to visit collection of models. This intuition is supported both by very promising experimental results obtained in a variety of challenging situations and by some theoretical results. In particular, we have established the geometric ergodicity of the STMALA algorithm for a large class of target distributions. To our best knowledge, it is the first result providing a rate of convergence for a trans-dimensional MCMC algorithm (like RJMCMC and MCC); usually, only Harris recurrence is proved, see [16].

Our algorithm is closely related to the proximal MCMC algorithm of [17]; the main difference stems from the fact that our algorithm is designed to sample jointly the models and their parameters, whereas [17] is a method to sample from high-dimensional posterior distributions with sparsity inducing priors.

This paper is organized as follows. STMALA and its application to Bayesian variable selection is described in Section 2. The geometric ergodicity of the STMALA algorithm is addressed in Section 3. Numerical experiments on simulated and real data sets to assess the performance of STMALA are given in Section 4. All the proofs are postponed to Section 6.

2 The Shrinkage-Thresholding MALA algorithm

This section introduces the Shrinkage-Thresholding MALA algorithm which is designed to sample from a target distribution defined on $\mathbb{R}^p$, $p \in \mathbb{N}^*$. Denote by $\mathcal{M} \stackrel{\text{def}}{=} \{0,1\}^p$ the set of binary vectors. For any $m = (m_1, \ldots, m_p) \in \mathcal{M}$, set

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

the family of active, i.e. nonzero, variables. For any $m \in \mathcal{M}$, denote by S_m the subset of $\mathbb{R}^p$ defined by

$$S_m \stackrel{\text{def}}{=} \{z \in \mathbb{R}^p, z_i \neq 0, i \in I_m, z_j = 0, j \notin I_m\} \tag{2}$$

and by $|m| \stackrel{\text{def}}{=} \sum_{i=1}^p m_i$ the number of non-zero components in m. $\{S_m, m \in \mathcal{M}\}$ is a partition of $\mathbb{R}^p$ and we assume that the target distribution may be written as

$$\pi(\mathrm{d}x) = \sum_{m \in \mathcal{M}} \omega_m \pi_m(x) \mathbb{1}_{S_m}(x) \nu_m(\mathrm{d}x)\ , \tag{3}$$

where $\{\omega_m, m \in \mathcal{M}\}$ is the prior probability of the models and $\pi_m(x)\nu_m(\mathrm{d}x)$ is the distribution of x conditionally to the model m. We consider situations when $\nu_m(\mathrm{d}x) = \prod_{i \in I_m} \mathrm{d}x_i \prod_{j \notin I_m} \delta_0(\mathrm{d}x_j)$

and $\pi_m(x) \propto \exp(-U_m(x) - V_m(x))$ with $x \mapsto U_m(x)$ continuously differentiable and $x \mapsto V_m(x)$ possibly non-smooth (a penalization term).

Two different shrinkage-thresholding operators are considered to sample sparse vectors, namely the Proximal one (Prox) Ψ_1 and the soft thresholding operator with vanishing shrinkage (STVS) Ψ_2: for any $\gamma > 0$, $1 \leq i \leq p$ and $u = (u_1, \cdots, u_p) \in \mathbb{R}^p$,

$$(\Psi_1(u))_i \stackrel{\text{def}}{=} u_i \left(1 - \gamma/|u_i|\right)_+ , \tag{4}$$

$$(\Psi_2(u))_i \stackrel{\text{def}}{=} u_i \left(1 - \gamma^2/|u_i|^2\right)_+ , \tag{5}$$

where for $a \in \mathbb{R}$, a_+ denotes the positive part of a: $a_+ \stackrel{\text{def}}{=} \max(a, 0)$. Lemma 2.1 shows that the

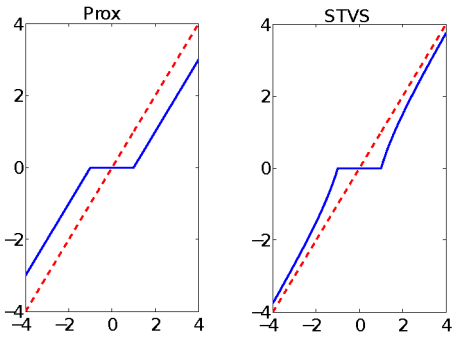

Figure 1: Continuous line : Shrinkage-Thresholding functions associated with Prox (left) and STVS (right) in one dimension. Dashed lines : the identity function.

soft thresholding operator with vanishing shrinkage Ψ_2, known as the empirical Wiener operator (see [18]), compromises between minimizing a (non-convex) function h and being close to u.

Lemma 2.1. *For any $\gamma > 0$ and $u \in \mathbb{R}^p$,*

$$\Psi_2(u) = \text{argmin}_{x \in \mathbb{R}^p} \left(t(x) + \frac{1}{2}\|x - u\|^2 \right) ,$$

where

$$t(x) = \gamma^2 \left[\text{asinh}\left(\|x\|/(2\gamma)\right) - (1/2) \exp\left(-2 \,\text{asinh}\left(\|x\|/(2\gamma)\right)\right) \right] .$$

Proof. The proof is postponed to Section 6.1. □

From a current state $X^n \in \mathbb{R}^p$ the algorithm proposes a new point Z defined by

$$Z = \Psi \left(X^n + \frac{\sigma^2}{2} h(X^n) + \sigma \xi^{n+1} \right) , \tag{6}$$

where $\sigma > 0$, $\xi^{n+1} \sim \mathcal{N}_p(0, I)$ and for all $x \in \mathbb{R}^p$ and $D > 0$,

$$h(x) \stackrel{\text{def}}{=} \sum_{m \in \mathcal{M}} \mathbb{1}_{S_m}(x) \frac{D \, \nabla U_m(x)}{D \vee \|\nabla U_m(x)\|} , \tag{7}$$

with $a \vee b = \max(a, b)$. The following lemma shows that this proposal mecanism is equivalent to sampling a new binary vector $m' \in \mathcal{M}$ conditionally to x; and then sampling a new vector with non null components in $\mathbb{R}^{|m'|}$ conditionally to (m', x). Define

$$\mu(x) \stackrel{\text{def}}{=} x + \sigma^2 h(x)/2 . \tag{8}$$

Lemma 2.2. *Let $x \in \mathbb{R}^p$, $D, \gamma, \sigma > 0$. Let $\Psi \in \{\Psi_1, \Psi_2\}$. The random vector $\Psi(\mu(x) + \sigma\xi)$ where $\xi \sim \mathcal{N}_p(0, I)$ has a density with respect to $\sum_{m\in\mathcal{M}} \nu_m$ given by $z \mapsto q_\Psi(x, z)$ with*

$$q_\Psi(x, z)\mathbb{1}_{S_m}(z) = \left(\prod_{i\notin I_m} \rho(\mu_i(x))\right) \times \left(\prod_{i\in I_m} f_\Psi(\mu_i(x), z_i)\right) \tag{9}$$

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

$$\begin{aligned}
&\rho(c) \stackrel{\text{def}}{=} \mathbb{P}\left\{|c + \sigma\zeta| \le \gamma\right\} \ , \text{ with } \zeta \sim \mathcal{N}(0, 1) \ , \\
&f_{\Psi_1}(c, z) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-1/2} \times \exp\left\{-\left|\left(1 + \gamma|z|^{-1}\right) z - c\right|^2 /(2\sigma^2)\right\} \ ;
\end{aligned}$$

and

$$f_{\Psi_2}(c, z) \stackrel{\text{def}}{=} \left(2\pi\sigma^2\right)^{-1/2} g\left(\gamma^2|z|^{-2}\right) \tilde{g}\left(\gamma^2|z|^{-2}\right) \exp\left(-\left|g\left(\gamma^2|z|^{-2}\right)\ z - c\right|^2 /(2\sigma^2)\right) \ ,$$

with

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

Proof. The proof is postponed to Section 6.2. □

For any x, $z \mapsto q_\Psi(x, z)$ consists in

(i) sampling each component of a new model $m' \in \mathcal{M}$ as independent $\{0, 1\}$-Bernoulli random variable with success parameter $\rho(\mu_i(x))$, $1 \le i \le p$;

(ii) for $i \notin I_{m'}$, set $z_i = 0$; conditionally to (m', x), sample independent components such that for any $i \in I_{m'}$, the distribution of z_i on $\mathbb{R}^\star$ is $f_\Psi(\mu_i(x), z_i)$.

The proposal (6) is then accepted and $X^{n+1} = Z$ with probability $\alpha_\Psi(X^n, Z)$ given by

$$\alpha_\Psi(x, z) \stackrel{\text{def}}{=} 1 \wedge \frac{\pi(z)\, q_\Psi(z, x)}{\pi(x)\, q_\Psi(x, z)} \ ; \tag{10}$$

otherwise, $X^{n+1} = X^n$. In high dimensional settings, STMALA may encounter some difficulties to accept the proposed moves. Following [19], we introduce a variant of the algorithm in which only a fixed number η of components of X^n is updated at each iteration n. This is achieved by combining STMALA and a Gibbs sampler in a STMALA-within-Gibbs algorithm.

3 V-Geometric ergodicity of the L_1 proximal STMALA

In this section, we address the V-geometric ergodicity of the STMALA chain $(X^n)_{n\ge0}$ under the following assumptions: for any $m \in \mathcal{M}$,

A1 (i) $\omega_m > 0$ and $\pi_m > 0$ on S_m.
(ii) π_m is continuous on S_m.
(iii) $\pi_m(x)\mathbb{1}_{S_m}(x) \to 0$ when $\|x\| \to \infty$.

A2 for any $s > 0$,

$$\lim_{r\to\infty} \sup_{x\in S_m, \|x\|\ge r} \pi_m(x + s\, n(x))/\pi_m(x) = 0 \ ,$$

where $n(x) \stackrel{\text{def}}{=} x/\|x\|$.

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

$$W_m(x) \stackrel{\text{def}}{=} \{(\|x\| - u)n(x) - s\zeta : s \in (0, b-u)\ ;\ \zeta \in S_m, \|\zeta\| = 1, \|\zeta - n(x)\| \leq \epsilon\}\ . \tag{11}$$

$W_m(x)$ is the cone of S_m with apex $x - u\ n(x)$ and aperture 2ϵ. We will prove (see Lemma 6.6) that A3 guarantees that, the probability to accept a move from x to any point of $W_m(x)$ converges to one as $\|x\|$ goes to infinity.

A3 There exist $b, R, \epsilon > 0$ and $u \in (0, b)$ such that for any $m \in \mathcal{M}$, for any $x \in S_m \cap \{\|x\| \geq R\}$, for all $y \in S_m \cap W_m(x)$: $\pi_m(x - u\ n(x)) \leq \pi_m(y)$.

When for any $m \in \mathcal{M}$, π_m is differentiable on S_m, A2 and A3 are satisfied if (see for details), for all $m \in \mathcal{M}$,

$$\lim_{x \in S_m, \|x\| \to \infty} \langle n(x), \nabla \log(\pi_m(x)) \rangle = -\infty\ ,$$
$$\limsup_{x \in S_m, \|x\| \to \infty} \langle n(x), n(\nabla \pi_m(x)) \rangle < 0\ ;$$

(see [20, Section 4 and the proof of Theorem 4.3] for details).

Let P_Ψ denote the transition kernel associated to the Hastings-Metropolis move with proposal (6) and acceptance-rejection ratio (10).

Theorem 3.1. *Assume A1-3 hold. Then, for any $\Psi \in \{\Psi_1, \Psi_2\}$, for any $\beta \in (0, 1)$, there exist $C > 0$ and $\lambda \in (0, 1)$ such that for any $n \geq 0$ and any $x \in \mathbb{R}^p$,*

$$\|P_\Psi^n(x, .) - \pi\|_V \leq C\, V(x)\, \lambda^n\ , \tag{12}$$

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

Proof. By definition of the acceptance-rejection ratio, π is invariant with respect to P_Ψ. The V-uniform geometric ergodicity follows from Proposition 6.4 and Proposition 6.8 given in Section 6.3: Proposition 6.4 establishes that the chain is psi-irreducible and aperiodic and shows that any Borel set $C \subset \mathbb{R}^p$ such that $C \cap S_m$ is a compact subset of S_m is a small set for P_Ψ; Proposition 6.8 shows that there exists an accessible small set $C \subset \mathbb{R}^p$ and constants $c_1 \in (0, 1)$ and $c_2 < \infty$ such that for any $x \in \mathbb{R}^p$, $P_\Psi V(x) \leq c_1 V(x) + c_2 \mathbb{1}_C(x)$. The proof is then concluded by [21, Theorem 15.0.2]. □

4 Numerical illustrations

In this section, STMALA[1] is compared to the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. For any $\ell \times \ell'$ matrix A and any $1 \leq j \leq \ell$, $1 \leq k \leq \ell'$, $A_{.,k}$ (resp. $A_{j,.}$) denotes the k-th column (resp. the j-th row) of A. In all the sequel, only the performance of STMALA with Ψ_2 is considered due to lack of space. It has been experimentally observed in all the considered scenarios that Ψ_2 performs significantly better than Ψ_1, because it avoids to shrink the significative components of x.

In the examples below, π is the posterior distribution of a regression vector in a logistic regression model; π_m is the conditional distribution of the regression vector conditionally to the observations and to the model m.

4.1 Logistic regression

Let G be a known $N \times p$ design matrix. We have N independent observations $Y = (Y_1, \ldots, Y_N)$ such that for all i, Y_i is a Bernoulli random variable with parameter $\exp(G_{i,.}X)/(1 + \exp(G_{i,.}X))$. Conditionally to a model $m \in \mathcal{M}$, the prior on the nonzero components of the regression vector $X \in S_m$ is $\mathcal{N}(0, c(G_m' G_m)^{-1})$, where c is a known scaling parameter, and G_m denotes the matrix

[1] MATLAB codes for STMALA are available at the address http://www.math.u-psud.fr/∼lecorff/software.html

with columns $\{G_{.,i}, i \in I_m\}$. The prior on the models ω_m is equal to $\theta_\star^{|m|}(1-\theta_\star)^{p-|m|}$ for $\theta_\star \in (0,1)$. In this experiment, we choose $p = 50$ and $N = 100$ to assess the performance of STMALA in a simple framework; the components of G are i.i.d. $\mathcal{N}(0,1)$ and $\theta_\star = 0.05$. The algorithm is run with $c = 100$, $\sigma = 0.3$ and $\eta = 5$. The choice of the threshold γ in Ψ_2 is crucial (if γ is too large, few nonzero samples are proposed and the algorithm converges slowly and if γ is too small, the algorithm proposes non-sparse solutions that are not likely to be accepted): γ is set to 0.4 to get a mean acceptance rate of around 20%.

STMALA is used to estimate the posterior probabilities of activation of the components of X, defined for all $1 \leq i \leq p$ as the conditional probability of the event $\{X_i \neq 0\}$ given the observations Y. The estimation is given by $\sum_{n=B}^{N_{it}+B} \mathbb{1}_{\{X_i^n \neq 0\}}/N_{it}$ where N_{it} is the number of iterations of the algorithm and B denotes the number of iterations discarded as a burn-in period. We choose $N_{it} = 50.000$ and $B = 10.000$. Figure 2 (top) provides the true regression vector, the posterior mean of the regression vector given by STMALA and the estimated activation probabilities over 100 independent Monte Carlo runs. This experiment highlights the ability of STMALA to choose the good model (the 3 nonzero components of X are recovered) and to get high posterior probabilities of activation for the selected components of X.

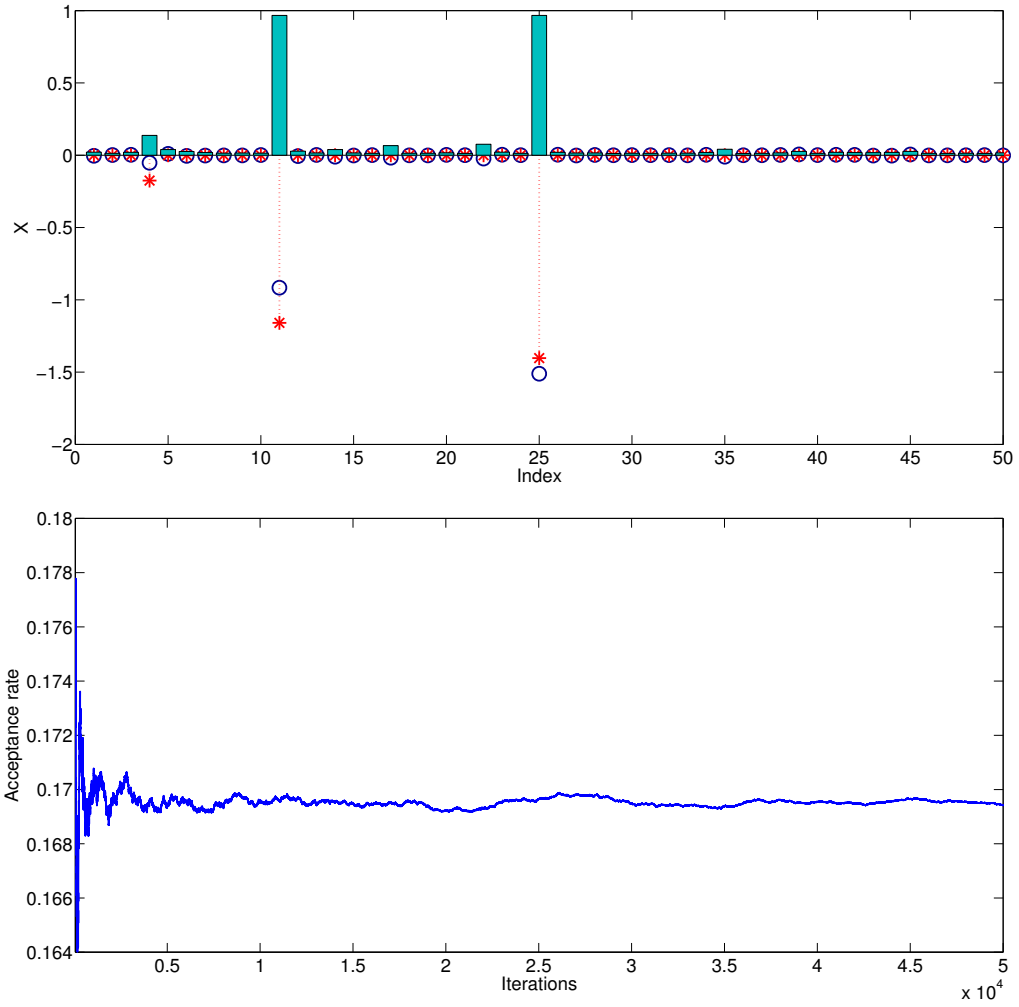

Figure 2: (top) True regression vector (stars), mean regression vector (circles) and estimated activation probabilities (bars). (bottom) Mean acceptance rate as a function of the number of iterations.

4.2 Linear regression

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

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

where G is a $N \times p$ (known) design matrix, E is a Gaussian random vector with i.i.d. standard entries and τ is the (known) precision. The prior on the models is $\omega_m = \theta_\star^{|m|}(1-\theta_\star)^{p-|m|}$ for some (known) $\theta_\star \in (0,1)$. The conditional distribution of X given the observations Y and the model m is given by

$$\pi_m(x) \propto \exp\left(-\frac{\tau}{2}\|Y - Gx\|^2\right) \times \prod_{\ell=1}^{p} \left\{ \left(1 + \frac{x_\ell^2}{2aK}\right)^{-(a+1/2)} \mathbb{1}_{\{m_\ell=1\}} + \delta_0(x_l)\, \mathbb{1}_{\{m_\ell=0\}} \right\} .$$

Such a posterior distribution can be obtained from the following hierarchical model: *(i)* given $m \in \mathcal{M}$ and positive precisions $(\vartheta_1, \ldots, \vartheta_p)$, the entries $X = (X_1, \ldots, X_p)$ are independent with distribution

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

(ii) the precision parameters $\vartheta = (\vartheta_1, \ldots, \vartheta_p)$ are i.i.d. with Gamma distribution $\mathrm{Ga}(a, aK)$, where a, K are fixed.

The performance of STMALA is illustrated with the model introduced in [22] and presented in [23, Section 8]. We choose $N = 100$ and $p = 200$. The covariates $(G_{\cdot,1}, \ldots, G_{\cdot,p})$ are sampled from a Gaussian distribution with $\mathbb{E}[G_{\cdot,i}] = 0$ and $\mathbb{E}[G_{ji}G_{ki}] = (0.3)^{|j-k|}$; $\tau = 1$. To produce the observations, we choose the nonzero coefficients of X in 4 clusters of 5 adjacent variables such that, for all $k \in \{1,2,3,4\}$ and all $j \in \{1,2,3,4,5\}$, $X_{50*(k-1)+j} = (-1)^{k+1}\, j^{1/k}$. Below, this *true* value of the regression vector is denoted by $X^\star$.
$\theta_\star = 0.1$, $a = 2$ and $K = 0.08$. STMALA is run with $\eta = 20$ and $\gamma = 0.35$.

The standard deviation of the RJMCMC proposal is chosen so that STMALA and RJMCMC have similar acceptance rates (between 15% and 20%). Figure 3 shows the true regression vector

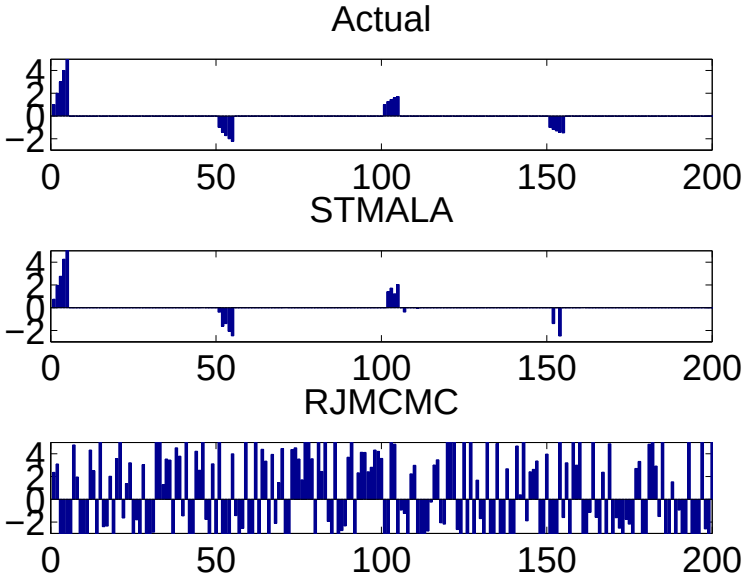

Figure 3: Regression vectors estimated by STMALA and RJMCMC.

X and its estimates obtained by STMALA and RJMCMC; these estimates $\hat{X}$ are defined as the posterior mean along a trajectory of length 10^6 (the first 10% samples are discarded). It shows

that 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 4. 50 independent trajectories of length 10^6 are run; Figure 4 (top) shows the evolution of the mean number (over the 50 runs) of active components $|m|$. RJMCMC has not converged after the 300.000 iterations while the mean number of active components of STMALA is stable after few iterations. Figure 4 (bottom) displays the boxplots of the estimation of the first component X_1 estimated by STMALA and RJMCMC as a function of the number of iterations.

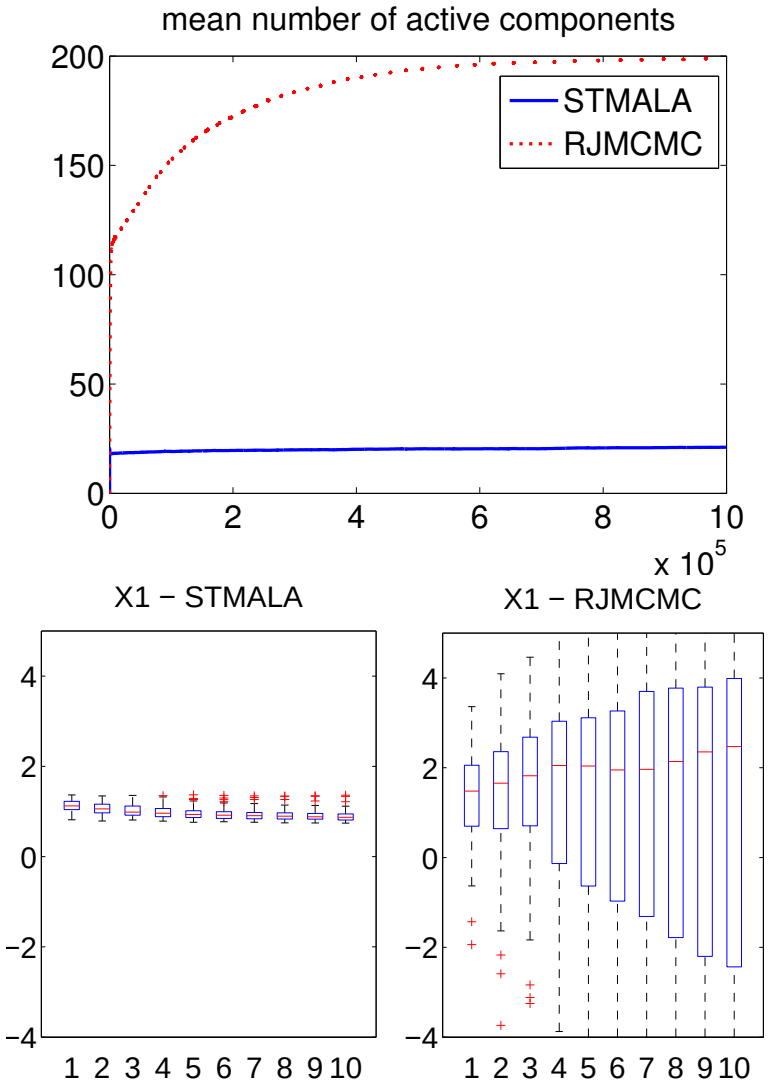

Figure 4: (top) Evolution of the mean number of active components for STMALA and RJMCMC. (bottom) Evolution of the estimation of $X_1^\star$ (mean over iterations) for STMALA and RJMCMC.

Figure 5 (top) shows the signal $G\hat{X}$ estimated by 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 $Y_{\text{test}} = G_{\text{test}}X^\star + \tau^{-1/2}E_{\text{test}}$, 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 $G_{\text{test}}X^\star$ are displayed. This test data set is also used to compute a test error, which is given by

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

The evolution of the mean test error $\mathcal{E}_{\text{test}}$ over 100 independent runs, is displayed in Figure 5 (bottom). Both figures show that RJMCMC is subject to some over fitting, which is not the case of STMALA.

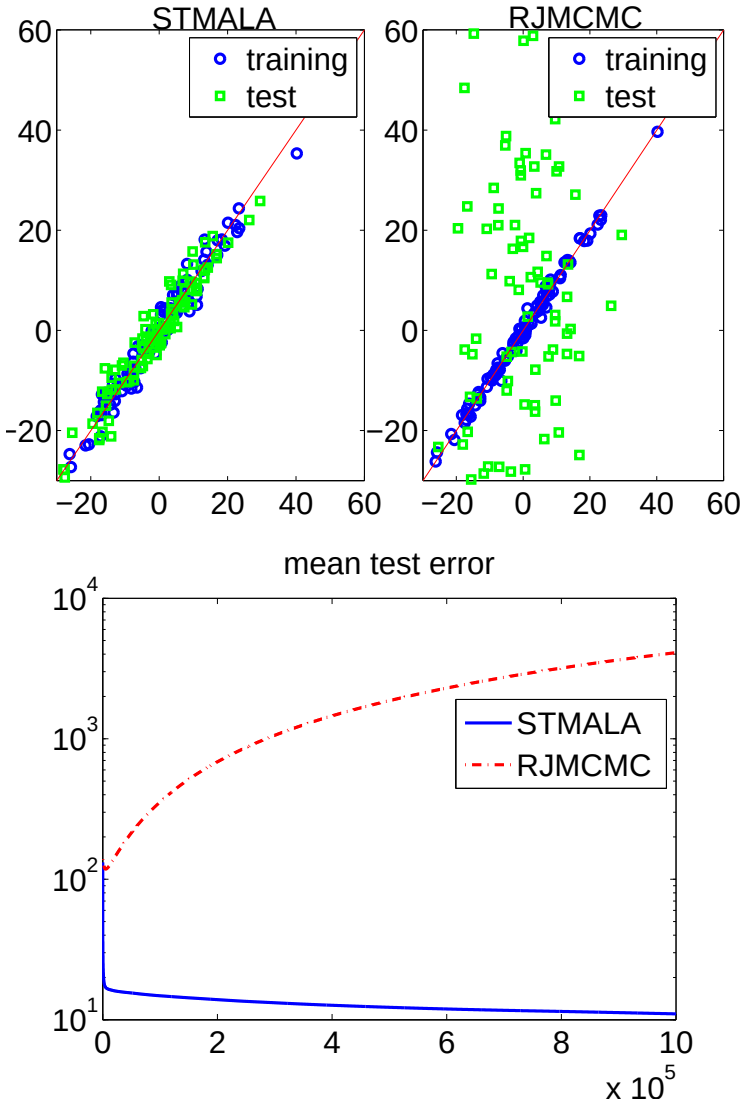

Figure 5: (top) Emitted signal $G\hat{X}$ estimated by STMALA and RJMCMC versus actual emitted signal GX. (bottom) Evolution of the mean test error for RJMCMC and STMALA.

4.3 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 [9] and [24]. 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. The observation model is given by

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

where G is the design matrix, X is the unknown regression vector and $E \sim \mathcal{N}(0, I)$ is the measurement noise. 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 STMALA with those obtained by RJMCMC for the prediction of fat content. 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 parameters of the algorithms are given by $\tau = 0.5$, $\eta = 15$, $\gamma = 0.35$ for STMALA. The computations are made over 100 independent trajectories of $N_{it} = 2.10^6$ iterations, with a burn-in $B = 10^5$. The design parameters of STMALA and RJMCMC are chosen so that the two algorithms have similar acceptance-rejection ratios (the final ratios are about 45% for STMALA and 42% for RJMCMC). Figure 6 shows the regression vectors $\hat{X}$ obtained by STMALA and RJMCMC, and computed as the posterior mean along one trajectory (left) and the mean regression vector estimated by STMALA and RJMCMC over 100 independent trajectories (right).

The regression vector estimated by STMALA has a spike around 1726 nm, which is known to be in a fat absorbance region (see [9, 24]), in almost all the trajectories.

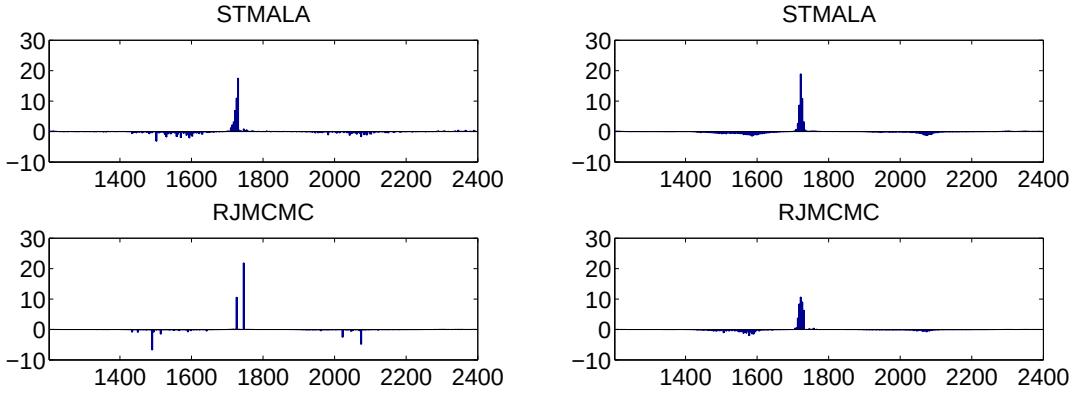

Figure 6: (left) Regression vectors estimated by STMALA and RJMCMC. (right) Mean regression vectors estimated STMALA and RJMCMC over 100 independent trajectories.

Figure 7 displays the boxplots of the 100 independent values of the components of the regression vectors estimated by STMALA and RJMCMC associated to 9 wavelengths close to 1726 nm. It illustrates that the location of the spike retrieved by RJMCMC is not stable, while STMALA retrieves a spike centered at 1726 nm in almost every trajectory.

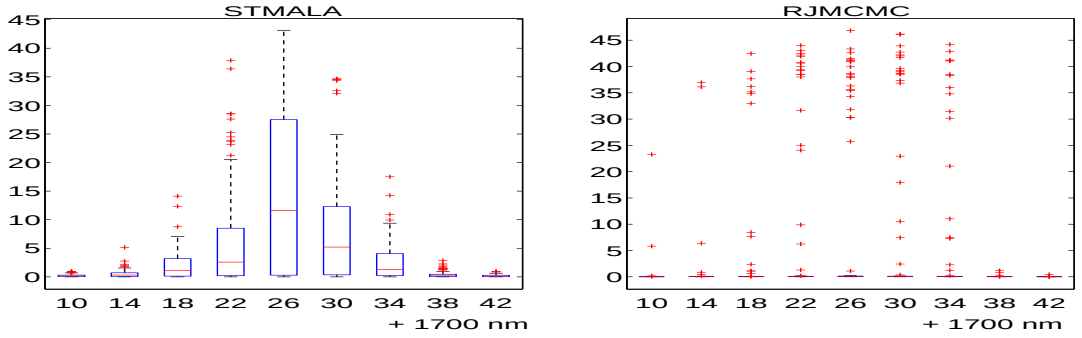

Figure 7: Boxplots of the 100 independent values of the components of the regression vectors estimated by STMALA and RJMCMC associated with 9 wavelengths close to 1726 nm.

Figure 8 (top) shows the estimated emitted signal $G\hat{X}$ obtained by STMALA and RJMCMC as a function of the observations Y. In this numerical experiment, STMALA provides better results than RJMCMC for both the training set and the test set. This is confirmed by Figure 8 (bottom)

which displays the evolution of the mean square error (MSE) on the test dataset, defined by

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

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 STMALA and about 1.6 times greater for RJMCMC.

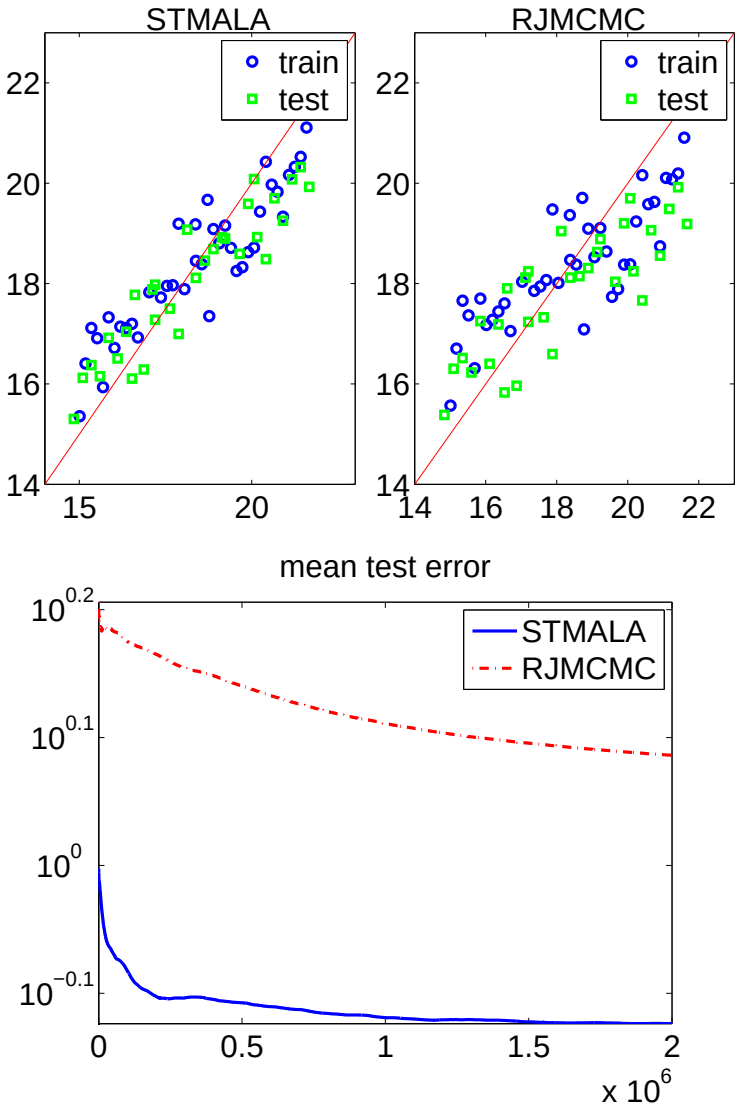

Figure 8: (top) Emitted signal $G\hat{X}$ estimated by STMALA and RJMCMC versus the observations Y. (bottom) Evolution of the mean MSE (over 100 independent trajectories) on the test data set for RJMCMC and STMALA.

5 Conclusions

In this paper, we propose a new trans-dimensional MCMC algorithm to perform Bayesian variable selection in a high-dimensional regression setting. This algorithm is closely related to [17] but is adapted to sample models which are exactly sparse in the sense that a certain number of components are equal to zero. In addition, under fairly weak assumptions, the STMALA algorithm is shown to

be geometrically ergodic. In the high-dimensional setting, the STMALA algorithm outperforms the RJMCMC algorithm which is considered as the state of the art. The performance of the STMALA algorithm depends on the tuning of a set of parameters: an adaptive version is currently under investigation. Also, the algorithm has still to be adapted to the ultra large scale framework, which likely requires additional specific procedures.

6 Proofs

For all $m \in \mathcal{M}$, define $\kappa^{[m]} = (\kappa_1^{[m]}, \cdots, \kappa_p^{[m]})$ as the indices of nonzero coefficients of m: $\kappa_1^{[m]} \stackrel{\text{def}}{=} \inf\{1 \leq i \leq p \ ; \ m_i = 1\}$ and for $2 \leq j \leq |m|$, $\kappa_j^{[m]} \stackrel{\text{def}}{=} \inf\{i > \kappa_{j-1}^{[m]} : m_i = 1\}$. Then, for all $x \in \mathbb{R}^{|m|}$, let $x^{[m]}$ be the vector of $\mathbb{R}^p$ such that for all $1 \leq i \leq |m|$, $x^{[m]}_{\kappa_i^{[m]}} = x_i$ and for all $i \notin \{\kappa_1^{[m]}, \ldots, \kappa_{|m|}^{[m]}\}$, $x_i^{[m]} = 0$. For all $y \in \mathbb{R}^p$ and all $m \in \mathcal{M}$, let $y_{[m]}$ be the vector of $\mathbb{R}^{|m|}$ such that for all $1 \leq i \leq |m|$, $(y_{[m]})_i = y_{\kappa_i^{[m]}}$.

6.1 Proof of Lemma 2.1

Consider first the case $p = 1$. We first compute the derivative of t on $(0, \infty)$ (note that t is symmetric). For any $x \in (0, \infty)$,

$$t'(x) = \gamma^2 \left[(x^2 + 4\gamma^2)^{-1/2} + (x^2 + 4\gamma^2)^{-1/2} \exp\left(-2 \operatorname{asinh}\left(x/(2\gamma)\right)\right) \right] .$$

Using straightforward computations, we get

$$t'(x) = (-x + \operatorname{sign}(x)\sqrt{x^2 + 4\gamma^2})/2 .$$

Set $\psi_u(x) \stackrel{\text{def}}{=} t(x) + \ (x-u)^2/2$. 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 $(0, \infty)$ if and only if $u > \gamma$. 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 $u > \gamma$.

Consider now the case $p > 1$. Set $x \in \mathbb{R}^p$ of the form $x = r\xi$ where $r > 0$ and ξ is on the unit sphere of $\mathbb{R}^p$. Since the function t only depends on the radius r, the minimum over $\mathbb{R}^p$ of $x \mapsto t(x) + \|x - u\|^2/2$ 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.

6.2 Proof of Lemma 2.2

Let φ be a bounded continuous function on $\mathbb{R}^p$. Then,

$$\mathbb{E}[\varphi(Z)] = \left(2\pi\sigma^2\right)^{-p/2} \int_{\mathbb{R}^p} \varphi\left(\Psi_1(y)\right) \times \prod_{i=1}^{p} \exp\left(-\frac{|y_i - \mu_i(x)|^2}{2\sigma^2}\right) \mathrm{d}y .$$

For $m \in \mathcal{M}$ and $y \in \mathbb{R}^{|m|}$, set $\overline{y} = (\overline{y}_1, \cdots, \overline{y}_{|m|})$ where $\overline{y}_i \stackrel{\text{def}}{=} y_i\left(1 - \gamma/|y_i|\right)$. Fubini's theorem yields

$$\mathbb{E}[\varphi(Z)] = \left(2\pi\sigma^2\right)^{-|m|/2} \sum_{m \in \mathcal{M}} \prod_{i \notin I_m} \rho\left(\mu_i(x)\right) \times \int_{\mathbb{R}^{|m|}} \varphi\left(\overline{y}^{[m]}\right) \left(\prod_{k=1}^{|m|} \mathbb{1}_{\{|y_k| > \gamma\}}\right) \times \exp\left(-\frac{\|y - \mu(x)_{[m]}\|^2}{2\sigma^2}\right) \mathrm{d}y .$$

It is sufficient to compute integrals of the form

$$I(\widetilde{\varphi}) = \int_{\mathbb{R}} \widetilde{\varphi}\left(v\left(1-\frac{\gamma}{|v|}\right)\right) \mathbb{1}_{\{|v|>\gamma\}} \times \exp\left(-\frac{|v-\mu|^2}{2\sigma^2}\right) \mathrm{d}v \ ,$$

for a generic function $\widetilde{\varphi}$. Consider the change of variable $\mathbb{R} \setminus [-\gamma, \gamma] \to \mathbb{R}^\star$: $z = v\left(1-\frac{\gamma}{|v|}\right)$. Note that $|z| = |v| - \gamma$ and $v = \psi(z)$, where for any $z \in \mathbb{R}^\star$, $\psi(z) \stackrel{\text{def}}{=} (1+\gamma/|z|)z$. Then,

$$I(\widetilde{\varphi}) = \int_{\mathbb{R}^\star} \widetilde{\varphi}(v) \exp\left(-\frac{|\psi(v)-\mu|^2}{2\sigma^2}\right) \mathrm{d}v \ .$$

This concludes the proof for Ψ_1. The proof for Ψ_2 follows the same lines as the proof of Lemma 2.2, with the function ψ replaced by $\widetilde{\psi}(z) = g\left(\gamma^2/\|z\|^2\right)\ z$.

6.3 Proof of Theorem 3.1

For ease of notations, we denote by q the proposal distribution. Lemma 2.2 shows that for any $m \in \mathcal{M}$ and $y \in S_m$

$$q(x,y) = \prod_{i \notin I_m} \rho\left(\mu_i(x)\right) \prod_{i \in I_m} f\left(\mu_i(x), y_i\right) \ , \tag{13}$$

where ρ and f are given by Lemma 2.2 and $\mu(x) = (\mu_1(x), \cdots, \mu_p(x))$ is given by (8). 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. Denote by g_σ the one-dimensional centered Gaussian density with standard deviation σ.

Lemma 6.1. *There exist k_1, k_2, σ_1 and σ_2 such that For any $x, y \in \mathbb{R}^p$ and any $1 \le i \le p$,*

$$k_1\ g_{\sigma_1}(y_i - x_i) \le f(\mu_i(x), y_i) \le k_2\ g_{\sigma_2}(y_i - x_i) \ ,$$

Proof. Assume first that $\Psi = \Psi_1$. Let $x, y \in \mathbb{R}^p$ and $i \in \{1, \cdots, p\}$. By definition of μ (see (8)), we have $|\mu_i(x) - x_i| \le \|\mu(x) - x\| \le D\sigma^2/2$. Thus,

$$|y_i - x_i| \le |y_i + \gamma\ \mathrm{sign}(y_i) - \mu_i(x)| + \gamma + \frac{D\sigma^2}{2} \ ,$$

which implies $|y_i + \gamma\ \mathrm{sign}(y_i) - \mu_i(x)|^2 \ge \frac{1}{2}\,|y_i - x_i|^2 - \left(\gamma + D\sigma^2/2\right)^2$. Similarly, $|y_i + \gamma\ \mathrm{sign}(y_i) - \mu_i(x)|^2 \le 2\,|y_i - x_i|^2 + 2\left(\gamma + D\sigma^2/2\right)^2$.

Assume now that $\Psi = \Psi_2$ and let $x, y \in \mathbb{R}^p$ and $i \in \{1, \cdots, p\}$. First,

$$g(\gamma^2|y|^{-2})\tilde{g}(\gamma^2|y|^{-2}) = (1 + 1/\sqrt{1+4\gamma^2|y|^{-2}})/2 \ ,$$

which yields $1/2 \le g(\gamma^2|y|^{-2})\tilde{g}(\gamma^2|y|^{-2}) \le 1$. Furthermore,

$$|g(\gamma^2|y|^{-2})y - \mu(x)| \le |g(\gamma^2|y|^{-2})y - y| + |y - x| + |x - \mu(x)| \le \gamma + |y - x| + D\sigma^2/2 \ ,$$

On the other hand,

$$|y - x| \le |g(\gamma^2|y|^{-2})y - \mu(x)| + |g(\gamma^2|y|^{-2})y - y| + |x - \mu(x)| \le |g(\gamma^2|y|^{-2})y - \mu(x)| + \gamma + D\sigma^2/2 \ .$$

□

Corollary 6.2. *For any $x \in \mathbb{R}^p$ and $y \in S_m$, $q(x,y) \le k_2^{|m|} \prod_{i \in I_m} g_{\sigma_2}(y_i - x_i)$. Therefore, there exists a constant $C > 0$ such that for any $x, y \in \mathbb{R}^p$, $q(x,y) \le C$.*

The proof of Theorem 3.1 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 ρ and μ be given by Lemma 2.2 and (8). It holds*

$$\inf_{m \in \mathcal{M}} \inf_{z \in S_m} \inf_{i \notin I_m} \rho(\mu_i(z)) > 0 \ .$$

Proof. For $i \notin I_m$, by (8), $|\mu_i(z)| \leq D\sigma^2/2$. Hence, there exists a constant $C > 0$ such that

$$\inf_{z \in \mathbb{R}^p} \min_{i \notin I_m} \mathbb{P}(|\mu_i(z) + \sigma\xi| \leq \gamma) \geq C \;, \tag{14}$$

where $\xi \sim \mathcal{N}(0,1)$. □

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. Then, C is a one-small set for the kernel P_Ψ: there exists a positive measure $\tilde{\nu}$ on $\mathbb{R}^p$ such that $P_\Psi(x, A) \geq \tilde{\nu}(A)\mathbb{1}_C(x)$.*

(ii) The Markov kernel P_Ψ is psi-irreducible and aperiodic.

Proof. For notation simplicity, we drop the dependency in Ψ (i). We set $\nu = \sum_{m \in \mathcal{M}} \nu_m$. 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

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

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, q (see (13)) and ν

$$P(x, A) \geq \int_{A \cap K} \alpha(x, y) q(x, y) \mathrm{d}\nu(y)$$

where, for any $x \in S_{m_\star}$ and $y \in S_m$, we have

$$q(x, y) = \prod_{i \notin I_m} \rho(\mu_i(x)) \prod_{i \in I_m} f_\Psi\left(\mu_i(x), y_i\right) \;.$$

The latter inequality implies

$$P(x, A) \geq \sum_{m \in \mathcal{M}} k_1^{|m|} \prod_{i \notin I_m} \rho(\mu_i(x)) \times \int_{A \cap K \cap S_m} \alpha(x, y) \prod_{i \in I_m} g_{\sigma_1}(x_i - y_i) \mathrm{d}y_i \;,$$

where the last inequality follows from Lemma 6.1. For any $x \in S_{m_\star}$ and $y \in S_m$, we have

$$\alpha_\Psi(x, y) = 1 \wedge \frac{\omega_m \pi_m(y) q(y, x)}{\omega_{m_\star} \pi_{m_\star}(x) q(x, y)} \;.$$

There exists a compact set of $\mathbb{R}$ such that for any $x \in C \cap S_{m_\star}$ and $y \in K \cap S_m$, $\mu_i(x)$ and $\mu_i(y)$ are in this compact for any i. Hence, A1(i)-(ii) and Lemmas 6.1 and 6.3 imply that there exists $\varepsilon_m > 0$ such that for any $x \in C \cap S_{m_\star}$ and $y \in K \cap S_m$,

$$\alpha_\Psi(x, y) \geq \varepsilon_m \;, \qquad \inf_{i \in I_m} g_{\sigma_1}(x_i - y_i) \geq \varepsilon_m \;.$$

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

(ii): By [25, Lemma 1.1], the Markov chain $(X^n)_{n \geq 0}$ is psi-irreducible since for any $x, y \in \mathbb{R}^p$, $q(x, y) > 0$ as a consequence of Lemma 6.1 and strongly aperiodic since by Proposition 6.4(i) it possesses an accessible 1-small set. □

For any measurable function $f : \mathbb{R}^p \to \mathbb{R}^+$, $Pf : \mathbb{R}^p \to \mathbb{R}^+$ denotes $Pf(x) = \int P(x, \mathrm{d}z) f(z)$. Fix $\beta \in (0, 1)$ and set $V : \mathbb{R}^p \to [1, \infty)$, $x \mapsto c_\beta \pi^{-\beta}(x)$. Define the possible rejection region $R(x)$ by

$$R(x) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^p : \pi(x) q(x, y) > \pi(y) q(y, x)\} \;.$$

We have

$$\frac{PV(x)}{V(x)} \leq \sum_{m \in \mathcal{M}} \left\{ T_m(x) + \int_{R(x) \cap S_m} q(x, y) \mathrm{d}\nu_m(y) \right\} \;, \tag{15}$$

where

$$T_m(x) \stackrel{\text{def}}{=} \int_{\mathbb{R}^{|m|}} \alpha_\Psi(x, z^{[m]}) \, \frac{\pi^{-\beta}(z^{[m]})}{\pi^{-\beta}(x)} \, q(x, z^{[m]}) \, \mathrm{d}z \;. \tag{16}$$

Lemma 6.5. *For any $m \in \mathcal{M}$, $\limsup_{\|x\| \to \infty} T_m(x) = 0$.*

Proof. The proof is adapted from [20] and [26]. Let $m \in \mathcal{M}$ be fixed. Define

$$\begin{aligned}
\mathcal{B}_m(x,a) &\stackrel{\text{def}}{=} \{z \in \mathbb{R}^{|m|}, \|z - x_{[m]}\| \leq a\} \;, \\
\mathcal{C}_m(x) &\stackrel{\text{def}}{=} \{z \in \mathbb{R}^{|m|}, \pi(z^{[m]}) = \pi(x)\} \;, \\
\mathcal{C}_m(x,u) &\stackrel{\text{def}}{=} \{z + sn(z), |s| \leq u, z \in \mathcal{C}_m(x)\} \;, \\
R_m(x) &\stackrel{\text{def}}{=} \mathbb{R}^{|m|} \setminus A_m(x) \;,
\end{aligned}$$

where

$$A_m(x) \stackrel{\text{def}}{=} \{z \in \mathbb{R}^{|m|}, \pi(z^{[m]}) q(z^{[m]}, x) \geq \pi(x) q(x, z^{[m]})\} \;.$$

We decompose as follows

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

where

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

We prove that we may choose the constant $C > 0$ large enough so that for any $\epsilon > 0$ there exists $M > 0$ 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 $z \in \mathbb{R}^{|m|}$,

$$\alpha_\Psi(x, z^{[m]}) \frac{\pi^{-\beta}(z^{[m]})}{\pi^{-\beta}(x)} \leq \left(\frac{q(z^{[m]}, x)}{q(x, z^{[m]})} \right)^\beta \;. \tag{17}$$

Control of $T_{m,1}$ By (17), $T_{m,1}(x,a) \leq \int_{\mathcal{B}_m^c(x,a)} q(x, z^{[m]})^{1-\beta} q(z^{[m]}, x)^\beta \mathrm{d}z$. By (13) and Lemma 6.1, there exists a constant $C > 0$ such that

$$\begin{aligned}
T_{m,1}(x,a) \leq C k_2^{|m|(1-\beta)} \times \int_{\mathcal{B}_m^c(x,a)} \prod_i g_{\sigma_2}\left((x_{[m]})_i - y_i\right)^{1-\beta} \mathrm{d}y_i \\
\leq C k_2^{|m|(1-\beta)} \int_{\mathcal{B}_m^c(0,a)} \prod_i g_{\sigma_2}(y_i)^{1-\beta} \mathrm{d}y_i \;.
\end{aligned}$$

Therefore, for any $\epsilon > 0$, 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 (17), $T_{m,2}(x,a,u) \leq \int_{\mathcal{B}_m(x,a) \cap \mathcal{C}_m(x,u)} q(x, z^{[m]})^{1-\beta} q(z^{[m]}, x)^\beta \mathrm{d}z$. By A2, the Lebesgue measure of $\mathcal{B}_m(x,a) \cap \mathcal{C}_m(x,u)$ can be made arbitrarily small - independently of $x \in \mathbb{R}^p$ - when u is small enough (see [20, Proof of Theorem 4.1] for details). Therefore, since q is bounded (see Corollary 6.2), for any $\epsilon > 0$, there exists $u > 0$ such that for any $a > 0$: $\sup_{x \in \mathbb{R}^p} T_{m,2}(x,a,u) \leq \epsilon$.

Control of $T_{m,3}$ Set $d_r(u) \stackrel{\text{def}}{=} \sup_{\|x\|\geq r} \pi(x + u\, n(x))/\pi(x)$. By A2, for any $\epsilon, u > 0$, there exists $r > 0$ large enough so that $(d_{r-u}(u))^{1-\beta} \vee (d_r(u))^{1-\beta} \leq \epsilon$. By A1, $\sup_{z\in\mathcal{B}_m(0,r)} \pi(z^{[m]})^{-\beta} < \infty$, so that by corollary 6.2

$$\sup_{x\in\mathbb{R}^p} \int_{\mathcal{I}_m(x,a,u,r)} q(x, z^{[m]})\pi^{-\beta}(z^{[m]})\mathrm{d}z < \infty \; ,$$

where

$$\mathcal{I}_m(x, a, u, r) \stackrel{\text{def}}{=} A_m(x) \cap \mathcal{B}_m(x, a) \cap \mathcal{C}_m^c(x, u) \cap \mathcal{B}_m(0, r) \; .$$

A1(iii) implies that

$$\limsup_{\|x\|\to\infty} \int_{\mathcal{I}_m(x,a,u,r)} \frac{\pi^{-\beta}(z^{[m]})}{\pi^{-\beta}(x)} q(x, z^{[m]})\mathrm{d}z = 0 \; .$$

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

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

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

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

This yields there exists $C_\star$ such that for any $a, u, r > 0$,

$$\int_{A_m(x)\cap\mathcal{B}_m(x,a)\cap\mathcal{J}_m(x,u,r)} \frac{\pi^{-\beta}(z^{[m]})}{\pi^{-\beta}(x)} q(x, z^{[m]})\mathrm{d}z \leq C_\star \left(\sup_{z\in\mathcal{J}_m(x,u,r)} \frac{\pi^{\beta}(x)}{\pi^{\beta}(z^{[m]})} \wedge \sup_{z\in\mathcal{J}_m(x,u,r)} \frac{\pi^{1-\beta}(z^{[m]})}{\pi^{1-\beta}(x)} \right) \; ,$$

where

$$\mathcal{J}_m(x, u, r) \stackrel{\text{def}}{=} \mathcal{C}_m^c(x, u) \cap \mathcal{B}_m^c(0, r) \; .$$

Let $z \in \mathcal{C}_m^c(x, u) \cap \{z : \pi(z^{[m]}) < \pi(x)\}$. By A1(ii), $h : s \mapsto \pi(z^{[m]} - s\, n(z^{[m]})) - \pi(x)$ 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 $\pi(z^{[m]}) < \pi(x)$), this implies that $h(u) < 0$ i.e. $\pi(z^{[m]} - sn(z^{[m]})) \leq \pi(x)$. Then,

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

If $z \in \mathcal{C}_m^c(x, u) \cap \{z : \pi(z^{[m]}) \geq \pi(x)\}$, we obtain similarly that $\pi(x)/\pi(z^{[m]}) \leq d_r(u)$. Hence, we established that

$$\sup_{z\in\mathcal{C}_m^c(x,u)\cap\mathcal{B}_m^c(0,r)} \frac{\pi(z^{[m]})}{\pi(x)} \leq d_r(u) \vee d_{r-u}(u) \; .$$

As a conclusion, there exists $C_\star > 0$ and for any $\epsilon, a, u > 0$, there exists $M > 0$ such that $\sup_{\|x\|\geq M} T_{m,3}(x, a, u) \leq C_\star\epsilon$.

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 exists $C_\star > 0$ and for any $\epsilon, a, u > 0$, there exists $M > 0$ such that $\sup_{\|x\|\geq M} T_{m,4}(x, a, u) \leq C_\star\epsilon$. □

Lemma 6.6. *Let u, b, ϵ, R be given by A3 and $W_m(x)$ be defined by (11). There exists $r > R$ such that for any $m \in \mathcal{M}$ and $x \in S_m \cap \{\|x\| \geq r\}$, $W_m(x) \subset \{y \in S_m, \alpha_\Psi(x, y) = 1\}$.*

Proof. The proof is adapted from [20]. 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 A3). We first prove that there exists a positive constant C_b such that

$$\frac{\pi(x)}{\pi(x - un(x))} \leq C_b \leq \inf_{z \in \mathcal{B}_m(x,b)} \frac{q(z^{[m]}, x)}{q(x, z^{[m]})} \; . \tag{18}$$

By (13), Lemma 6.1 and Lemma 6.3, there exist $C, C_b > 0$ - independent of $x \in S_m$ - such that

$$\inf_{z \in \mathcal{B}_m(x,b)} \frac{q(z^{[m]}, x)}{q(x, z^{[m]})} \geq C^{p-|m|} k_1^{|m|} k_2^{-|m|} \times \inf_{z \in \mathcal{B}_m(x,b)} \prod_{i \in I_m} \frac{g_{\sigma_1}(x_i - z_i)}{g_{\sigma_2}(x_i - z_i)} \geq C_b \; .$$

By A2, we can choose r large enough so that for all $\|x\| \geq r$, $\pi(x)/\pi(x - un(x)) \leq C_b$. This yields (18). Let $z \in W_m(x)$. Then, $\|z - x\| \leq b$ so that $z \in \mathcal{B}_m(x, b)$. Hence, by (18), $q(z^{[m]}, x)/q(x, z^{[m]}) \geq C_b$. In addition,

$$\frac{\pi(z^{[m]})}{\pi(x)} = \frac{\pi(z^{[m]})}{\pi(x - un(x))} \frac{\pi(x - un(x))}{\pi(x)} \geq \frac{\pi(z^{[m]})}{\pi(x - un(x))} \frac{1}{C_b} \geq \frac{1}{C_b} \; ,$$

where in the last inequality we used A3. Hence,

$$\frac{\pi(z^{[m]})}{\pi(x)} \frac{q(z^{[m]}, x)}{q(x, z^{[m]})} \geq 1 \; ,$$

and $\alpha_\Psi(x, z^{[m]}) = 1$ thus showing the lemma. □

Lemma 6.7. $\limsup_{\|x\| \to \infty} \int_{R(x)} q(x, y) \mathrm{d}\nu(y) < 1$, *where* $\mathrm{d}\nu = \sum_m \mathbb{1}_{S_m} \mathrm{d}\nu_m$.

Proof. Let $x \in S_{m_\star}$. By definition of $\mathrm{d}\nu$, by Lemma 6.1 and by Lemma 6.3, there exists a constant $C > 0$ such that

$$\begin{aligned} 1 - \int_{R(x)} q(x, y) \mathrm{d}\nu(y) = \sum_{m \in \mathcal{M}} \int_{A_m(x)} q(x, z^{[m]}) \mathrm{d}z &\geq \sum_{m \in \mathcal{M}} k_1^{|m|} \mathcal{G}_m(x) \prod_{i \notin I_m} \rho(\mu_i(x)) \; , \\ &\geq k_1^{|m_\star|} \; \mathcal{G}_{m_\star}(x) \prod_{i \notin I_{m_\star}} \rho(\mu_i(x)) \; , \\ &\geq C \, k_1^{|m_\star|} \; \mathcal{G}_{m_\star}(x) \; , \end{aligned}$$

where

$$\mathcal{G}_m(x) \stackrel{\text{def}}{=} \int_{A_m(x)} \prod_{i \in I_m} g_{\sigma_1}(x_i - y_i) \mathrm{d}y_i \; .$$

By Lemma 6.6, for any $x \in S_{m_\star}$ large enough,

$$1 - \int_{R(x)} q(x, y) \mathrm{d}\nu(y) \geq C \, k_1^{|m_\star|} I_{m_\star}(x)$$

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

$$I_{m_\star}(x) = \int_{W_{m_\star}(x) - x} \left(\prod_{i \in I_{m_\star}} g_{\sigma_1}(y_i) \mathrm{d}y_i \right) \times \left(\prod_{i \notin I_{m_\star}} \delta_0(\mathrm{d}y) \right) \; . \tag{19}$$

Note that

$$W_{m_\star}(x) - x = \{ -u \, n(x) - s\xi ; 0 < s < b - u, \, \zeta \in S_{m_\star}, \|\zeta\| = 1, \|\zeta - n(x)\| \leq \epsilon \} \; ,$$

so that the integrals in (19) depend on x only through $m_\star$. Since $\mathcal{M}$ is finite, there exists a constant $C' > 0$ independent of x such that for any $m \in \mathcal{M}$ and $x \in S_m$,

$$\int_{W_m(x)} \left(\prod_{i \in I_m} g_{\sigma_1}(x_i - y_i) \mathrm{d}y_i \right) \times \left(\prod_{i \notin I_m} \delta_0(\mathrm{d}y) \right) \geq C' \; .$$

□

Proposition 6.8. $\limsup_{\|x\|\to\infty} PV(x)/V(x) < 1$.

Proof. The result follows from (15) and Lemmas 6.5 and 6.7. □

References

[1] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer, 2004.

[2] P. Green, " Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–723, 1995.

[3] S. Brooks, P. Giudici, and G. Roberts, "Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions," *J. Roy. Statist. Soc. B*, vol. 61, no. 1, pp. 3–39, 2003.

[4] G. Karagiannis and C. Andrieu, "Annealed Importance Sampling Reversible Jump MCMC Algorithms," *J. Comput. Graph. Statist.*, vol. 22, no. 3, pp. 623–648, 2013.

[5] B. Carlin and S. Chib, "Bayesian model choice via Markov chain Monte Carlo methods," *J. Roy. Statist. Soc. B*, vol. 157, pp. 473–484, 1995.

[6] S. J. Godsill, "On the relationship between Markov chain Monte Carlo methods for model uncertainty," *J. Comput. Graph. Statist.*, vol. 10, no. 2, pp. 230–248, 2001.

[7] P. Dellaportas, J. Forster, and I. Ntzoufras, "On Bayesian model and variable selection using MCMC," *Stat. Comput.*, vol. 12, pp. 27–36, 2002.

[8] A. Petralias and P. Dellaportas, "A MCMC model search algorithm for regression problems," *J. Statist. Comput. Simulation*, vol. 83, no. 9, pp. 1722–1740, 2013.

[9] P. Brown, T. Fearn, and M. Vannucci, "Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem," *J. Amer. Statist. Assoc.*, vol. 96, no. 454, pp. 398–408, 2001.

[10] D. Nott and R. Kohn, "Adaptive sampling for Bayesian variable selection," *Biometrika*, vol. 92, no. 4, pp. 747–763, 2005.

[11] D. Lamnisos, J. Griffin, and M. Steel, "Adaptive Monte Carlo for Bayesian Variable Selection in Regression Models," *J. Comput. Graph. Statist.*, vol. 22, no. 3, pp. 729–748, 2013.

[12] M. Shi and D. Dunson, "Bayesian Variable Selection via Particle Stochastic Search," *Statist. Probab. Lett.*, vol. 81, no. 2, pp. 283–291, 2011.

[13] C. Schäfer and N. Chopin, "Sequential Monte Carlo on large binary sampling spaces," *Stat. Comput.*, vol. 23, no. 2, pp. 163–184, 2013.

[14] G. Roberts and R. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996.

[15] A. Beck and M. Teboulle, "A fast iterative shrinkage-tresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[16] G. Roberts and J. Rosenthal, "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains," *Ann. Appl. Probab.*, vol. 16, no. 4, pp. 2123–2139, 2006.

[17] M. Pereyra, "Proximal markov chain monte carlo algorithms," *Statistics and Computing*, pp. 1–16, 2015. [Online]. Available: `http://dx.doi.org/10.1007/s11222-015-9567-4`

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

[19] P. Neal and G. Roberts, "Optimal scaling for partially updating MCMC algorithms," *Ann. Appl. Probab.*, vol. 16, no. 2, pp. 475–515, 2006.

[20] S. Jarner and E. Hansen, "Geometric ergodicity of Metropolis algorithms," *Stoch. Proc. Appl.*, vol. 85, no. 2, pp. 341–361, 2000.

[21] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability.* London: Springer, 1993.

[22] L. Breiman, "The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error," *J. Amer. Statist. Assoc.*, vol. 87, pp. 738–754, 1992.

[23] H. Ishwaran and J. Rao, "Spike and slab variable selection: frequentist ans bayesian strategies," *Ann. Statist.*, vol. 33, no. 2, pp. 730–773, 2005.

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

[25] K. Mengersen and R. Tweedie, "Rates of convergence of the Hastings and Metropolis algorithms," *Ann. Statist.*, vol. 24, no. 1, pp. 101–121, 1996.

[26] Y. Atchadé, "An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift," *Methodol. Comput. Appl. Probab.*, vol. 8, pp. 235–254, 2006.