Majorize-Minimize Adapted Metropolis-Hastings Algorithm

Yosra Marnissi, Emilie Chouzenoux, Amel Benazza-Benyahia,
Jean-Christophe Pesquet

Majorize-Minimize Adapted Metropolis–Hastings Algorithm

Yosra Marnissi, *Student Member, IEEE,* Emilie Chouzenoux, *Member, IEEE,*
Amel Benazza-Benyahia, *Member, IEEE,* and Jean-Christophe Pesquet, *Fellow, IEEE*

***Abstract*—The dimension and the complexity of inference problems have dramatically increased in statistical signal processing. It thus becomes mandatory to design improved proposal schemes in Metropolis-Hastings algorithms, providing large proposal transitions that are accepted with high probability. The proposal density should ideally provide an accurate approximation to the target density with a low computational cost. In this paper, we derive a novel Metropolis-Hastings proposal, inspired from Langevin dynamics, where the drift term is preconditioned by an adaptive matrix constructed through a Majorization-Minimization strategy. We propose several variants of low-complexity curvature metrics applicable to large scale problems. We demonstrate the geometric ergodicity of the resulting chain for the class of super-exponential distributions. The proposed method is shown to exhibit a good performance in two signal recovery examples.**

***Index Terms*—MCMC methods, Langevin diffusion, Majorization-Minimization, signal recovery**

I. Introduction

FINDING a solution to an inverse problem consists of estimating an unknown signal from measurements based on the direct model linking the target signal to the observed one. However, perfect measurements are generally not available due to the presence of some random parasite signals that make difficult the extraction of useful information. In this work, we consider the following observation model:

$$\mathbf{z} = \mathcal{D}(\mathbf{H}\overline{\mathbf{x}}) \tag{1}$$

where $\overline{\mathbf{x}} \in \mathbb{R}^Q$ denotes the target signal, $\mathbf{z} \in \mathbb{R}^N$ is the measured data, $\mathbf{H} \in \mathbb{R}^{N\times Q}$ is an observation matrix describing the linear acquisition model, and $\mathcal{D}$ expresses the noise model, related to measurements errors (additive Gaussian or multiplicative noise, for instance). Such model arises in several signal processing applications (deblurring, denoising, super resolution, reconstruction, compressive sensing, inpainting) with appropriate definitions of the operator $\mathbf{H}$ and the noise model $\mathcal{D}$ [1]–[4].

The Bayesian framework has been widely adopted to perform the task of retrieving an estimate of the target signal given the data $\mathbf{z}$ and the model matrix $\mathbf{H}$. Bayesian modeling considers the parameters of interest as random variables rather than deterministic ones. Hence, this approach requires to specify a prior probability density $\mathsf{p}(\mathbf{x})$ that describes what is known about the sought signal before data are observed. Estimates are then computed relying on the posterior law that takes into account the prior $\mathsf{p}(\mathbf{x})$ combined with information about observations $\mathsf{p}(\mathbf{z}|\mathbf{x})$ via Baye's rule:

$$\mathsf{p}(\mathbf{x}|\mathbf{z}) = \frac{\mathsf{p}(\mathbf{z}|\mathbf{x})\mathsf{p}(\mathbf{x})}{\int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{z}|\mathbf{x}')\mathsf{p}(\mathbf{x}')\mathrm{d}\mathbf{x}'}. \tag{2}$$

A major challenge in Bayesian methods is the calculation of the posterior distribution, or more precisely, its exploration. In addition, nowadays, it is common in many fields such as medicine, astronomy and microscopy, to handle huge amounts of data with increasingly sophisticated models [5]. In these challenging settings, even if the prior and the observation model are simple, the posterior law is almost always intractable in the sense that it can only be known up to a multiplicative constant and/or has a complicated form which requires massive computing resources for handling it. Regarding the difficulty in directly dealing with the posterior distribution, many methods have been proposed [6]. In this paper, we are interested in Markov chain Monte Carlo (MCMC) simulation based techniques for large scale signal processing problems.

MCMC methods are stochastic simulation methods that allow to approximate a given target distribution such as the posterior law, by relying on Markov chain theory and Monte Carlo integration. They proceed in two main steps. First, a Markov chain is built with a given transition rule so that its stationary states follow the posterior law [7]–[10]. Once the Markov chain has reached its stationary distribution, Monte Carlo approximation is used to infer the posterior characteristics. A famous MCMC method is the Metropolis-Hastings (MH) one. The Metropolis algorithm was first proposed in [11]. Later, in [7], the MH algorithm was introduced as a generalization of the Metropolis algorithm to handle broader classes of problems. In order to draw a sample from a target distribution $\mathsf{p}(\mathbf{x}|\mathbf{z})$, two steps are applied alternately. First, a sample $\tilde{\mathbf{x}}^{(t)}$ is generated according to some proposal distribution of density $\mathsf{g}(.|\mathbf{x}^{(t)})$ that may depend on the current state $\mathbf{x}^{(t)}$ at each iteration t and should be easy to draw from. The proposed variable is then accepted or rejected according to the following acceptance probability:

$$\alpha(\mathbf{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}) = \min\left(1, \frac{\mathsf{p}(\tilde{\mathbf{x}}^{(t)}|\mathbf{z})\mathsf{g}(\mathbf{x}^{(t)}|\tilde{\mathbf{x}}^{(t)})}{\mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z})\mathsf{g}(\tilde{\mathbf{x}}^{(t)}|\mathbf{x}^{(t)})}\right). \tag{3}$$

Y. Marnissi is with Safran TECH, Groupe Safran 1, Rue Jeunes Bois, 78772 Châteaufort, France (e-mail: marnissi.yosra@gmail.com).

E. Chouzenoux and J.-C. Pesquet are with University Paris-Saclay, CentraleSupélec, Inria, Centre de Vision Numérique, 3 rue Joliot Curie, 91190 Gif sur Yvette, France (e-mail: first.last@centralesupelec.fr).

A. Benazza-Benyahia is with University of Carthage, COSIM Lab., SUP'COM, Cité Technologique des Communications, 2080, Tunisia (e-mail: benazza.amel@supcom.rnu.tn).

This work was supported by the Institut Universitaire de France and by the Agence Nationale de la Recherche under grants ANR-14-CE27-0001 GRAPHSIP and ANR-17-CE40-0004-01 MAJIC.

However, the performance of the MH algorithm is obviously strongly related to the choice of the proposal distribution. This issue becomes especially critical in large scale problems. In general, when selecting a proposal in MH algorithms, one should consider two issues. First, whilst MH algorithms are guaranteed to yield samples from the target distribution after some burn-in period, the number of iterations required to reach convergence may be infeasibly large. Second, the generated samples after convergence may be correlated. This correlation originates from two main sources: (i) the correlation introduced by keeping unchanged the parameter value because the newly generated one was rejected; (ii) the correlation between successive samples for non-independent proposals. A poorly mixed chain tends to generate samples that are highly correlated leading to an incomplete summary of the target distribution and highly biased estimators. Consequently, more samples are needed to achieve the same precision as i.i.d methods (e.g., importance sampling [12], rejection sampling [13]). In [14], the efficiency of MH algorithms is discussed with respect to the acceptance probability. In general, a good proposal should be a good approximation (or at least, a good local approximation) to the target density without being costly to sample from. This problem is often tackled in an empirical manner. However, it is also possible to determine theoretically an optimal proposal scaling [14] or to use adaptive algorithms in order to find a local approximation of the target distribution automatically [15]. One typical approach is the Random Walk (RW) whose adaptive proposal law takes the form of a Gaussian distribution centered at the current state [16]. The popularity of this algorithm is mainly related to its simplicity of implementation. However, the RW usually takes too many steps to reach stability for high dimensional models. Furthermore, slow convergence together with bad mixing behavior could make the Markov chain more likely to get trapped into some regions and thus fail to explore efficiently the whole target space [17]. Intuitively, a good proposal density should take advantage of the local properties of the target distribution to accelerate the exploration of regions with high probability values. In particular, it should reflect the dependence structure of the target distribution for large scale problems. In this respect, a large amount of works has been devoted to construct proposals in MH algorithms in an attempt to meet these requirements [18]–[26]. In this work, we are interested in proposals based on the Euler discretization of the Langevin stochastic differential equation where the drift term accounts for the slope and curvature of the target law. Our main contribution is to propose a preconditioned version of the standard Metropolis Hastings adapted Langevin algorithm using an adaptive matrix based on a Majorize-Minimize strategy.

This paper is organized as follows. In Section II, we formulate the problem and we give a brief overview of the Langevin diffusion process. In Section III, we describe the new Majorize-Minimize adapted MH algorithm. In Section IV, a particular attention is paid to the convergence proof of the proposed algorithm. Section V is devoted to experimental results. Finally, some concluding remarks are drawn in Section VI.

II. Problem statement and related work

A. Langevin diffusion

A Q-dimensional Langevin diffusion is a continuous time Markov process $(\mathbf{x}(t))_{t\in[0,+\infty[}$ with values in $\mathbb{R}^Q$ defined as the solution to the following stochastic differential equation [27]:

$$(\forall t \in [0,+\infty[) \quad \mathrm{d}\mathbf{x}(t) = \mathbf{b}(\mathbf{x}(t))\mathrm{d}t + \mathbf{V}(\mathbf{x}(t))\mathrm{d}\mathbf{B}(t), \quad (4)$$

where $\mathbf{x}(0) \in \mathbb{R}^Q$, $(\mathbf{B}(t))_{t\in[0,+\infty[} \in \mathbb{R}^Q$ is a Brownian motion, and for every $\mathbf{x} \in \mathbb{R}^Q$, $\mathbf{V}(\mathbf{x}) \in \mathbb{R}^{Q\times Q}$ is the volatility matrix and $\mathbf{b}(\mathbf{x}) = (b_i(\mathbf{x}))_{i=1}^Q$ is the drift term defined as follows:

$$\begin{aligned}(\forall i \in \{1,\ldots,Q\}) \quad b_i(\mathbf{x}) &= \frac{1}{2}\sum_{j=1}^{Q} A_{i,j}(\mathbf{x})\frac{\partial \log \pi(\mathbf{x})}{\partial x_j} \\ &\quad + |\mathbf{A}(\mathbf{x})|^{\frac{1}{2}} \sum_{j=1}^{Q} \frac{\partial}{\partial x_j}\left(A_{i,j}(\mathbf{x})|\mathbf{A}(\mathbf{x})|^{-\frac{1}{2}}\right)\end{aligned} \quad (5)$$

where $\mathbf{A}(\mathbf{x}) = \mathbf{V}(\mathbf{x})\mathbf{V}(\mathbf{x})^\top = (A_{i,j}(\mathbf{x}))_{1\leq i,j\leq Q}$ is a symmetric definite positive matrix and $|\mathbf{A}(\mathbf{x})|$ denotes its determinant. Note that this process attains asymptotically a stationary distribution whose density is π. Moreover, if a state $\mathbf{x}(t_0)$ follows the distribution of density π, all subsequent states $\mathbf{x}(t_0+\tau)$, $\tau > 0$ also follow the same distribution. Density π is assumed here to be differentiable. Thereby, when $\pi = \mathsf{p}(\cdot \mid \mathbf{z})$, one can construct a Langevin Markov chain whose stationary law is the target posterior distribution. In the following, this choice for π is made.

The Langevin diffusion describes a dynamic in continuous time. However, one can still approximate this equation by discretizing time. This is done by splitting the time interval into a series of small intervals of length $\Delta t = \varepsilon^2$. The smaller the value of ε, the closer the approximation to the dynamic in continuous time. Numerous procedures have been developed for time discretization [28]. We focus here on Euler's scheme. Then, the Langevin diffusion reads

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \varepsilon^2\mathbf{b}(\mathbf{x}^{(t)}) + \varepsilon\mathbf{A}^{1/2}(\mathbf{x}^{(t)})\boldsymbol{\omega}^{(t+1)} \quad (6)$$

where $\varepsilon > 0$ is the stepsize resulting from Euler's discretization and $\left(\boldsymbol{\omega}^{(t)}\right)_{t\in\mathbb{N}} \in \mathbb{R}^Q$ is a realization of zero-mean white noise with covariance matrix $\mathbf{I}_Q$. Scheme (6) is referred to as the Unadjusted Langevin Algorithm (ULA) [18]. As pointed out in [18], [23], the time discretization error induces a bias in the ULA algorithm. Therefore, the Markov chain following ULA scheme may sway away from the target stationary distribution providing only an approximation of it, unless the stepsize decreases to zero in which case the algorithm is exact. This discrepancy can be corrected by adding a Metropolis acceptance test at each iteration to guarantee the reversibility of the chain with respect to the posterior distribution. The resulting sampler can be seen as an MH algorithm where, for every $t \in \mathbb{N}$, $\mathsf{g}(\cdot \mid \mathbf{x}^{(t)})$ is the density of a Gaussian distribution with mean $\mathbf{x}^{(t)} + \varepsilon^2\mathbf{b}(\mathbf{x}^{(t)})$ and covariance matrix $\varepsilon^2\mathbf{A}(\mathbf{x}^{(t)})$. Note that convergence properties have also been obtained for some variants of ULA in [18], [29], [30].

It is worth noting that two scale parameters play an important role in (6): ε determines the length of the proposed jumps, whereas the scale matrix $\mathbf{A}(\cdot)$ controls their direction. Various classes of algorithms have been developed from this diffusion model depending on the choice of this matrix. In the subsequent subsection, we will review the most popular ones.

B. Choice of the scale matrix

The standard Metropolis adjusted Langevin algorithm (MALA) is the simplest form of diffusion (6) when $\mathbf{A}(\cdot)$ equals $\mathbf{I}_Q$ [18]:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\nabla \log \mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z}) + \varepsilon \boldsymbol{\omega}^{(t+1)}. \quad (7)$$

It can be proved that MALA has $\mathsf{p}(\mathbf{x}|\mathbf{z})$ as its stationary distribution and is more likely to accept proposed values than a standard RW [18]. Indeed, the gradient information of the target distribution allows the chain to be guided toward regions of higher probability where most of the samples should lie and hence, it enables to achieve high acceptance rates [14], [31]. As a consequence, in several applications, MALA was shown to generally require less iterations to converge than the standard RW [22]. Moreover, it should be noted that a bad adjustment of ε can significantly affect the convergence rate especially for high dimensional problems [26]. For this reason, many methods have focused on how to choose a suitable stepsize in order to make the asymptotic average acceptance rate bounded away from zero in high dimensions [21], [26]. Despite these improvements, when the variables of interest are strongly correlated with strongly differing variances, MALA algorithm fails to explore efficiently the target space. In fact, since the third term in the MALA update is an isotropic Brownian motion, the discretization stepsize ε in such a parameter space, is generally constrained to take very small values in order to deal with the directions with smallest variances, which may result in a slow convergence of the algorithm, poor mixing of the chain and highly correlated samples [23]. The performance of MALA can be improved by introducing a scale matrix different from the identity matrix [27]. Some approaches have been proposed to accelerate the algorithm by preconditioning the proposal density with a constant scale matrix [20], according to the following scheme:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\mathbf{A}\nabla \log \mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z}) + \varepsilon \mathbf{A}^{1/2}\boldsymbol{\omega}^{(t+1)}, \quad (8)$$

where $\mathbf{A} \in \mathbb{R}^{Q\times Q}$ is a constant symmetric positive definite matrix. While the stepsize ε can easily be tuned with respect to the asymptotic acceptance rate, there is no clear guiding strategies for the selection of the constant matrix in the absence of some knowledge about the moments of the target density which are usually unknown. Furthermore, the use of the same preconditioning matrix in the whole algorithm may be inefficient since optimal scaling in the burn-in period may differ from that in the stationary phase [32]. Therefore, rather than employing a fixed global scale matrix in the proposal density, a position dependent matrix may be employed to take into account the local structure of the target density at each state of the Markov chain. In that respect, many algorithms [15], [22]–[25], [33] rely on adaptive procedures where $\mathbf{A}(\cdot)$ is tuned automatically according to the past behavior of the Markov chain resorting to some deterministic optimization tools. For example, when setting $\mathbf{A}(\mathbf{x})$ to the inverse of the Hessian matrix of $-\log \mathsf{p}(\mathbf{x}|\mathbf{z})$ at every $\mathbf{x} \in \mathbb{R}^Q$, and assuming a locally constant curvature, the term involving the derivatives of the scale matrix in (5) reduces to zero. Consequently, the generated sample at each iteration $t \in \mathbb{N}$ reads:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\mathbf{A}(\mathbf{x}^{(t)})\nabla \log \mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z}) + \varepsilon \mathbf{A}^{1/2}(\mathbf{x}^{(t)})\boldsymbol{\omega}^{(t+1)} \quad (9)$$

where, for every $\mathbf{x} \in \mathbb{R}^Q$, $\mathbf{A}^{-1}(\mathbf{x}) = -\nabla^2 \log \mathsf{p}(\mathbf{x}|\mathbf{z})$ that is, for every $i \in \{1,\dots,Q\}$ and, for every $j \in \{1,\dots,Q\}$, $\left[\mathbf{A}^{-1}\right]_{i,j}(\mathbf{x}) = -\dfrac{\partial^2 \log \mathsf{p}(\mathbf{x}|\mathbf{z})}{\partial x_i \partial x_j}$. Consequently, the computation of the drift term $\mathbf{b}(\cdot)$ becomes a scaled Newton step for minimizing $-\log \mathsf{p}(\cdot|\mathbf{z})$. Thus, a new sample of the Newton-based MCMC is more likely drawn from a highly probable region and then more likely accepted, which can speed up the convergence of the sampling process [24], [25], [33]. Note that, in practice, this method has a high computational cost since it requires the computation of the full Hessian matrix and its inverse at each iteration. This is especially critical for large scale problems and/or when the Hessian matrix is not positive definite. One appealing solution is to replace the Hessian by a scale matrix that provides information similar to the Hessian with a lower computational cost. In particular, several methods rely on the Fisher information matrix as a preconditioning matrix in the Langevin diffusion [22], [23], and can thus be interpreted as the discretization of the MALA algorithm directly on the natural Riemannian manifold where the parameters live. In the following, we propose a new approach where the scale matrix of the Langevin diffusion is chosen according to a Majorize-Minimize strategy.

III. Proposed algorithm

A. Majorize-Minimize framework

The Majorization-Minimization (MM) principle is a powerful tool for designing optimization algorithms. The idea behind the MM approach is to replace an original complicated minimization problem with successive minimizations of some well chosen surrogate functions, satisfying the so-called tangent majorant conditions [34]–[36]:

Definition III.1. *Tangent majorant.*
Let $\mathcal{J}\colon \mathbb{R}^Q \to \mathbb{R}$ and let $\mathbf{x}' \in \mathbb{R}^Q$. A function $f(\mathbf{x}', \cdot)$ is said to be a tangent majorant *function of $\mathcal{J}$ at $\mathbf{x}'$ if*

$$\begin{cases} P_1: \ f(\mathbf{x}', \mathbf{x}') = \mathcal{J}(\mathbf{x}'), \\ P_2: \ f(\mathbf{x}', \mathbf{x}) \geqslant \mathcal{J}(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^Q). \end{cases} \quad (10)$$

Let $\mathbf{x}^{(0)} \in \mathbb{R}^Q$ be an arbitrary initial value and let $\left(\mathbf{x}^{(t)}\right)_{t\geq 1}$ be the sequence constructed as follows

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \underset{\mathbf{x}\in\mathbb{R}^Q}{\operatorname{argmin}} f(\mathbf{x}^{(t)}, \mathbf{x}). \quad (11)$$

According to (10), the scheme (11) will produce a monotically decreasing sequence $(\mathcal{J}(\mathbf{x}^{(t)}))_{t\in\mathbb{N}}$ since we have

$$\mathcal{J}(\mathbf{x}^{(t)}) \underset{(a)}{=} f(\mathbf{x}^{(t)},\mathbf{x}^{(t)}) \underset{(b)}{\geqslant} f(\mathbf{x}^{(t)},\mathbf{x}^{(t+1)}) \underset{(c)}{\geqslant} \mathcal{J}(\mathbf{x}^{(t+1)}) \tag{12}$$

where (a) follows from the tangency property P_1, (b) from the minimization step (11), and (c) from the majorization property P_2 (see Figure 1). Then, under mild assumptions, the sequence can be shown to converge to a stationary point of $\mathcal{J}$ [37].

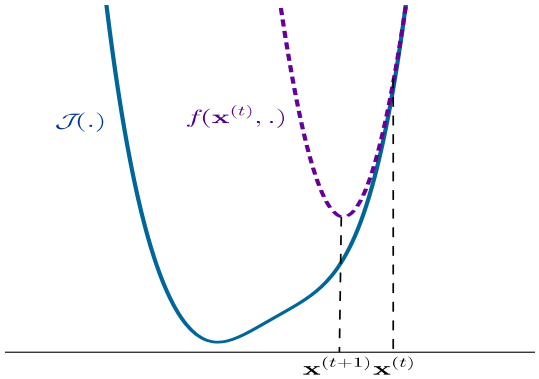

Fig. 1: MM algorithm: the new iterate $\mathbf{x}^{(t+1)}$ is the minimizer of the tangent majorant $f(\mathbf{x}^{(t)},.)$ of $\mathcal{J}$ at $\mathbf{x}^{(t)}$.

The performance of MM algorithms depends crucially on the choice of the surrogate function f. In particular, it has to be chosen so that a minimizer of it is easy to compute. A simple choice is to adopt an MM quadratic strategy, which consists in assuming the existence, for every $\mathbf{x}' \in \mathbb{R}^Q$, of a positive definite matrix $\mathbf{Q}(\mathbf{x}') \in \mathbb{R}^{Q\times Q}$ such that the following quadratic function defined, for every $\mathbf{x} \in \mathbb{R}^Q$, by

$$f(\mathbf{x}',\mathbf{x}) = \mathcal{J}(\mathbf{x}')+(\mathbf{x}-\mathbf{x}')^\top\nabla\mathcal{J}(\mathbf{x}')+\frac{1}{2}(\mathbf{x}-\mathbf{x}')^\top\mathbf{Q}(\mathbf{x}')(\mathbf{x}-\mathbf{x}') \tag{13}$$

is a tangent majorant of $\mathcal{J}$ at $\mathbf{x}'$. Then, the MM optimization algorithm reduces to building a sequence $(\mathbf{x}^{(t)})_{t\in\mathbb{N}}$ through the following preconditioned gradient scheme:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{x}^{(t)})\nabla\mathcal{J}(\mathbf{x}^{(t)}) \tag{14}$$

with $\varepsilon \in]0,\sqrt{2}]$ is a relaxation stepsize. Note that (14) implies that inequality (b) in (12) is satisfied, by noticing that $2\varepsilon^{-2}\mathbf{Q}(\mathbf{x}') \succeq \mathbf{Q}(\mathbf{x}')$, for every $\mathbf{x}' \in \mathbb{R}^Q$ and every $\varepsilon \in]0,\sqrt{2}]$.

B. Proposed sampling algorithm

In this work, we propose to extend the idea behind the aforementioned MM quadratic strategy to the context of stochastic samplers. More specifically, our idea is to push the proposal distribution of the MH algorithm at each iteration from the current state to a region with high density value. Unlike the RW where the proposal is centered on the current state, we propose to pick the mean of the proposal density using an MM search step of the form (14), and then to explore the space around this center according to an MM curvature matrix $\mathbf{Q}(\mathbf{x}^{(t)})$ that should well describe the local curvature of the target distribution. This results in a preconditioned Langevin proposal where the scale matrix $\mathbf{A}(\mathbf{x}^{(t)})$ in (9), equal to the inverse of the curvature matrix $\mathbf{Q}(\mathbf{x}^{(t)})$, is constructed according to the MM strategy. Similarly to Newton-based MCMC methods, the drift term, when assuming zero curvature changes, leads, from a current state $\mathbf{x}^{(t)}$, to a state with a higher value of $\log \mathsf{p}(\mathbf{x}|\mathbf{z})$ since it results from an iteration of MM algorithm minimizing $\mathcal{J}(\mathbf{x}) = -\log \mathsf{p}(\mathbf{x}|\mathbf{z})$. Consequently, the obtained proposal reduces to a stochastically perturbed version of an MM iteration for minimizing $-\log \mathsf{p}(\mathbf{x}|\mathbf{z})$. The proposed sample is then subjected to the accept/reject rule of the MH algorithm. The resulting sampler called 3MH is described by Algorithm 1.

Algorithm 1 Majorize-Minimize adapted Metropolis–Hastings (3MH) algorithm

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\varepsilon \in]0,\sqrt{2}]$

for $t = 0, 1, \ldots$ **do**

Generate

$$\tilde{\mathbf{x}}^{(t)} \sim \mathcal{N}\left(\mathbf{m}(\mathbf{x}^{(t)}), \varepsilon^2\mathbf{Q}^{-1}(\mathbf{x}^{(t)})\right)$$

where

$$\mathbf{m}(\cdot) = \cdot + \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\cdot)\nabla \log \mathsf{p}(\cdot \mid \mathbf{z})$$

Acceptance-Rejection:

Generate $u \sim \mathcal{U}(0,1)$

Compute

$$\alpha(\mathbf{x}^{(t)},\tilde{\mathbf{x}}^{(t)}) = \min\left(1, \frac{\mathsf{p}(\tilde{\mathbf{x}}^{(t)}|\mathbf{z})\mathsf{g}(\tilde{\mathbf{x}}^{(t)}|\mathbf{x}^{(t)})}{\mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z})\mathsf{g}(\mathbf{x}^{(t)}|\tilde{\mathbf{x}}^{(t)})}\right)$$

where, for every $\mathbf{v} \in \mathbb{R}^Q$,

$$\mathsf{g}(\cdot \mid \mathbf{v}) \propto |\mathbf{Q}(\mathbf{v})|^{\frac{1}{2}} \exp\left(-\frac{1}{2\varepsilon^2}\|\cdot - \mathbf{m}(\mathbf{v})\|^2_{\mathbf{Q}(\mathbf{v})}\right)$$

if $u < \alpha(\mathbf{x}^{(t)},\tilde{\mathbf{x}}^{(t)})$ **then**

Accept: $\mathbf{x}^{(t+1)} = \tilde{\mathbf{x}}^{(t)}$

else

Reject: $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$

end if

end for

The metric $\mathbf{Q}(\cdot)$ is thus the precision matrix of the Gaussian proposal distribution, the choice of which is crucial for the efficiency of the sampling algorithm. We propose to set $\mathbf{Q}(\mathbf{x}^{(t)})$ at each iteration $t \in \mathbb{N}$ such that (13) is a tangent majorant to the minus logarithm of the posterior density at the current state $\mathbf{x}^{(t)}$, i.e. it should satisfy Properties P_1 and P_2 in (10). Furthermore, for practical efficiency, it must be chosen so as to provide a good approximation to the local curvature of the posterior distribution. In the following, we propose a general procedure for building such a set of suitable preconditioning matrices $\{\mathbf{Q}(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^Q}$ under some mild conditions on the posterior distribution.

C. Construction of the tangent majorant

We focus on the case when the minus-log of the target density function $\mathcal{J} = -\log \mathsf{p}(\cdot \mid \mathbf{z})$ can be expressed up to an

additive constant as

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}) = \Phi(\mathbf{H}\mathbf{x} - \mathbf{z}) + \Psi(\mathbf{x}) \tag{15}$$

where $\mathbf{z} \in \mathbb{R}^N$, $\mathbf{H} \neq \mathbf{0}_{N\times Q}$, and

$$\Psi(\mathbf{x}) = \sum_{s=1}^{S} \psi_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|) \tag{16}$$

with $(\forall s \in \{1,\dots,S\})$ $\mathbf{V}_s \in \mathbb{R}^{P_s\times Q}$, $\mathbf{c}_s \in \mathbb{R}^{P_s}$, and $(\psi_s)_{1\leqslant s\leqslant S}$ is a set of nonnegative continuous functions. Note that this form of posterior density is very versatile. It is frequently encountered in inverse problems where $\mathbf{z}$ is the observation, Φ is the data fidelity term and Ψ is the minus logarithm of the prior density involving some linear operators $\mathbf{V}_1,\dots,\mathbf{V}_S$. For instance, $(\mathbf{V}_s)_{1\leq s\leq S}$ may be matrices computing the horizontal and vertical discrete gradients (or higher order differences) between neighboring pixels, which are useful for edge preserving in image restoration problems. In this case, by setting $P_s = 1$ and $\psi_s = |\cdot|$, we recover the anisotropic total variation while for $P_s = 2$ and ψ_s equal to the ℓ_2 norm, we obtain the isotropic form of it [38]. Another important choice, is the analysis frame regularization where $\mathbf{V} = [\mathbf{V}_1^\top,\dots,\mathbf{V}_S^\top]^\top$ is a frame of $\mathbb{R}^Q$. For example, $\mathbf{V}_1$ may be the operator that computes low frequency wavelet coefficients and ψ_1 a function enforcing smooth solutions, while the remaining operators give the high frequency ones that can be well described using suitable heavy tailed functions ψ_s such as the ℓ_p^p penalties for $p < 1$, the Cauchy, or the Bernoulli-Gaussian models [38], [39]. As Langevin based algorithms require the use of differentiable regularizations, one can either rely on approaches based on Moreau-Yoshida regularisation [30] or use smoothed approximations of these functions that have a quadratic behavior near 0 [40]–[43].

We further make the following assumptions:

Assumption III.1.

(i) *Φ is a continuous coercive differentiable function with an L-Lipschitzian gradient, that is, for every $(\mathbf{u},\mathbf{v}) \in (\mathbb{R}^N)^2$,*

$$\|\nabla\Phi(\mathbf{u}) - \nabla\Phi(\mathbf{v})\| \leqslant L\|\mathbf{u}-\mathbf{v}\|,$$

with $L \in]0,+\infty[$.

(ii) *$(\forall s \in \{1,\dots,S\})$ ψ_s is a differentiable function the derivative of which is denoted by $\dot{\psi}_s$.*

(iii) *$(\forall s \in \{1,\dots,S\})$ $\psi_s(\sqrt{\cdot})$ is concave over $\mathbb{R}^+$.*

(iv) *$(\forall s \in \{1,\dots,S\})$ $(\exists \bar{\omega}_s > 0)$ such that $(\forall u > 0)$, $0 \leqslant \dot{\psi}_s(u) \leqslant \bar{\omega}_s u$ and $\lim_{u\to 0} \dot{\psi}_s(u)/u \in \mathbb{R}$.*

Assumption III.1(i) holds for a large number of data fidelity terms. This includes for example the Gaussian noise model, the Huber function which may be useful for limiting the influence of outliers present in some datasets [44], the Cauchy model [45], and the signal-dependent Gaussian model generally used as a second order approximation of mixed Poisson-Gaussian noise [46], as well as the exact Poisson-Gaussian likelihood [47]. More examples can be found in [35]. Furthermore, Assumptions III.1(ii)-(iv) are satisfied for several commonly used prior models such as the Student-t distribution, the Gaussian distribution as well as smoothed approximation of ℓ_p^p regularization functions for $p \leqslant 2$ and $\ell_2 - \ell_0$ penalties (asymptotically constant with a quadratic behavior near 0) used to approximate the ℓ_0 pseudo-norm [35], [48]–[50].[1]

Under Assumptions III.1(i)-(iv), convex quadratic tangent majorants of (15) can be obtained by setting (see [35]):

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_1(\mathbf{x}) = \mu\mathbf{H}^\top\mathbf{H} + \mathbf{V}^\top\mathrm{Diag}\{\boldsymbol{\omega}(\mathbf{x})\}\mathbf{V} + \zeta\,\mathbf{I}_Q \tag{17}$$

where $\mu \in [L,+\infty[$, $\mathbf{V} = \left[\mathbf{V}_1^\top,\dots,\mathbf{V}_S^\top\right]^\top$ and $\boldsymbol{\omega}(\mathbf{x}) = (\omega_i(\mathbf{x}))_{1\leq i\leq P}$ is such that, for every $s \in \{1,\dots,S\}$ and $p \in \{1,\dots,P_s\}$,

$$\omega_{P_1+P_2+\dots+P_{s-1}+p}(\mathbf{x}) = \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x}-\mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x}-\mathbf{c}_s\|}. \tag{18}$$

Hereabove, ζ is a nonnegative constant that can be useful to ensure the invertibility of $\mathbf{Q}_1(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^Q$.

The numerical efficiency of the proposed algorithm relies on the use of quadratic majorants that provide tight approximations of the target density but also whose curvature matrices are simple to compute. However, sampling from the proposal constructed by the MM strategy when using the curvature matrix $\mathbf{Q}_1(\cdot)$ given by (17) is often very difficult because of the high computational cost of each iteration and/or memory limitations. In fact, similarly to Newton MCMC samplers, the main computational cost is related to the computation of the inverse of (17) and sampling from the associated high-dimensional Gaussian distribution at each iteration. In the following, we will propose alternative choices of the curvature matrix, when matrix $\mathbf{Q}_1(\cdot)$ given by (17) leads to an intractable scheme. The practical efficiency of the different metric strategies will be analyzed in our experimental part.

Constant curvature matrix: We can resort to the following constant curvature matrix which can be seen as a majorant of (17) constant with respect to variable $\mathbf{x}$:

$$\mathbf{Q}_2 = \mu\mathbf{H}^\top\mathbf{H} + \max_{1\leqslant s\leqslant S} \bar{\omega}_s\,\mathbf{V}^\top\mathbf{V} + \zeta\mathbf{I}_Q. \tag{19}$$

It can be noted that in the special case when $\mathbf{H}$ is circulant and $\mathbf{V}^\top\mathbf{V} = \nu\mathbf{I}_Q$ with $\nu > 0$, which is the case for example when $\mathbf{V}$ is a tight frame analysis operator, then $\mathbf{Q}_2$ is easily invertible in the Fourier domain. More generally, when $\mathbf{H}$ and $\mathbf{V}$ can be diagonalized in the same basis, the inversion and the square root decomposition of (19) can be easily performed in this basis.

Diagonal curvature matrix: Using the convexity property of quadratic function and Jensen's equality [34], one can also derive an alternative choice for the majorant matrix that can be understood as a diagonal approximation of $\mathbf{Q}_1$:

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_3(\mathbf{x}) = \mathrm{Diag}\left(\mu\mathbf{L}^\top\mathbf{1}_N + \mathbf{P}^\top\boldsymbol{\omega}(\mathbf{x})\right) + \zeta\mathbf{I}_Q \tag{20}$$

where $\mathbf{L} \in \mathbb{R}^{N\times Q}$, $\mathbf{P} \in \mathbb{R}^{P\times Q}$, with $P = \sum_{s=1}^{S} P_s$, are matrices whose elements are given, respectively, by

$$(\forall i \in \{1,\dots,N\})(\forall j \in \{1,\dots,Q\}) \quad L_{i,j} = |H_{i,j}| \sum_{k=1}^{Q} |H_{i,k}|, \tag{21}$$

[1]Note that, in this work, improper prior laws are allowed provided that the resulting posterior distribution is proper.

and

$$(\forall i \in \{1,\dots,P\})(\forall j \in \{1,\dots,Q\}) \quad P_{i,j} = |V_{i,j}| \sum_{k=1}^{Q} |V_{i,k}|. \tag{22}$$

A complete proof for the construction of this diagonal majorant can be found in [51, Lemma 4.1].

IV. Convergence analysis

In this section, we establish the convergence of the proposed algorithm.

It can be first noticed that the drift term in Algorithm 1 is equivalent to

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{b}(\mathbf{x}) = \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{x})\mathbf{D}(\mathbf{x}) \tag{23}$$

where $\mathbf{D}(\mathbf{x})$ is the truncated gradient defined by

$$\mathbf{D}(\mathbf{x}) = \frac{d}{\max(d, \|\nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})\|)} \nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z}), \tag{24}$$

provided that the parameter $d > 0$ tends to $+\infty$. Similarly to [15], we will study the asymptotic behaviour of the algorithm when using the modified drift term (23).[2]

We further make the following assumptions:

Assumption IV.1. *$\mathsf{p}(\cdot \mid \mathbf{z})$ is the density of a super-exponential distribution that is $\mathsf{p}(\cdot \mid \mathbf{z})$ is positive and has continuous first derivatives such that*

$$\lim_{\|\mathbf{x}\| \to +\infty} \frac{\mathbf{x}^\top \nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})}{\|\mathbf{x}\|} = -\infty, \tag{25}$$

and

$$\limsup_{\|\mathbf{x}\| \to +\infty} \frac{\mathbf{x}^\top \nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})}{\|\mathbf{x}\| \, \|\nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})\|} < 0. \tag{26}$$

Assumption IV.2. *For every $\mathbf{x} \in \mathbb{R}^Q$, the preconditioning matrix $\mathbf{Q}(\mathbf{x})$ has a bounded spectrum i.e., there exist two constants $\nu_{\min} > 0$ and $\nu_{\max} > 0$ independent of $\mathbf{x}$ such that*

$$\left(\forall \mathbf{x} \in \mathbb{R}^Q\right) \quad \nu_{\min}\mathbf{I}_Q \preceq \mathbf{Q}(\mathbf{x}) \preceq \nu_{\max}\mathbf{I}_Q. \tag{27}$$

Remark IV.1. *Assumption IV.2 holds for all the curvature matrices $\mathbf{Q}_1(\cdot)$, $\mathbf{Q}_2$ and $\mathbf{Q}_3(\cdot)$ proposed in Section III-C provided that $\zeta > 0$. Furthermore, Assumption IV.2 together with* (24)*, imply that the drift term $\mathbf{b}$ is bounded that is*

$$\left(\forall \mathbf{x} \in \mathbb{R}^Q\right) \quad \|\mathbf{b}(\mathbf{x})\| \leqslant \frac{\varepsilon^2}{2}\nu_{\min}^{-1} d. \tag{28}$$

We now state sufficient conditions for Assumption IV.1 to be satisfied.

Proposition IV.1. *Consider Model* (15) *with $\Phi = \frac{1}{2}\|\cdot\|^2$ and $(\psi_s)_{1 \leqslant s \leqslant S}$ satisfying Assumptions III.1(ii)-(iv). Then, Assumption IV.1 is satisfied if one of the following properties holds:*

- *$\mathbf{H}$ is injective, for example $\mathbf{H} = \mathbf{I}_Q$ which is the case for denoising problems;*
- *there exists $s_0 \in \{1,\dots,S\}$ such that*
  - (i) $\mathrm{Ker}(\mathbf{H}) \cap \mathrm{Ker}(\mathbf{V}_{s_0}) = \{\mathbf{0}_Q\}$,[3]
  - (ii) $\lim_{u \to +\infty} \dot{\psi}_{s_0}(u)/u > 0$,

Proof. See Appendix A. □

Subsequently, we can establish the geometric ergodicity of the proposed algorithm based on the results concerning RW in [53] and an adaptation of the analysis in [15], [18], [54]. In particular, the fact that the drift term of 3MH algorithm is assumed to stay bounded enables the use of similar proofs as in [15], [55]. Since our algorithm appears as a special case of the MH algorithm, the chain $\left(\mathbf{x}^{(t)}\right)_{t \in \mathbb{N}}$ constructed by the 3MH algorithm has $\mathsf{p}(\mathbf{x}|\mathbf{z})$ as an invariant distribution. The first important step of the proof of geometric ergodicity is to compare the proposal density g to Gaussian proposals.

Proposition IV.2. *Under Assumption IV.2, there exists $(k_1, k_2, \sigma_1, \sigma_2) \in (]0,+\infty[)^4$ such that for every $(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^Q)^2$,*

$$k_1 \mathsf{n}\left(\mathbf{x}; \mathbf{y}, \sigma_1^2 \mathbf{I}_Q\right) \leqslant \mathsf{g}(\mathbf{x}|\mathbf{y}) \leqslant k_2 \, \mathsf{n}(\mathbf{x}; \mathbf{y}, \sigma_2^2 \mathbf{I}_Q) \tag{29}$$

where $\mathsf{n}(\cdot; \mathbf{y}, \sigma_i^2 \mathbf{I}_Q)$, is the density of the Gaussian distribution of mean $\mathbf{y}$ and covariance matrix $\sigma_i^2 \mathbf{I}_Q$, $i \in \{1, 2\}$.

Proof. See Appendix B. □

Theorem IV.1. *Under Assumptions IV.1 and IV.2, the Markov chain defined by the 3MH algorithm using the truncated gradient* (24) *is geometrically ergodic with stationary distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$.*

Proof. From Algorithm 1 and Proposition IV.2, g is positive and, for every $(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^Q)^2$, $\mathsf{g}(\mathbf{x}|\mathbf{y}) > 0$. Our algorithm appears as a special case of the Metropolis-Hastings algorithm. Moreover, $\mathsf{p}(\cdot \mid \mathbf{z})$ is positive and continuous. We can then deduce from Lemma 1.2 of [56] that the chain is aperiodic and from Lemma 1.1 of [56] that the chain is $\mathsf{p}(\mathbf{x}|\mathbf{z})$- irreducible with unique invariant distribution $\mathsf{p}(\mathbf{x}|\mathbf{z})$.

Assumption IV.1 has been introduced in [53] as a sufficient condition for the geometric ergodicity of the RW algorithm. It has been shown that, under Assumption IV.1, when (29) holds, the MALA algorithm with truncated gradient (24) is geometrically ergodic [15]. More explicitly, the proofs of the geometric ergodicity of the RW MH in [53] and the MALA algorithm in [15] rely on the same Gaussian bounds of the proposal density as the one established in (29). It follows that the geometric ergodicity property for 3MH can be deduced by a straightforward adaptation of the proof in [15] for MALA algorithm with truncated drift. Note that the geometric ergodicity is actually obtained for any preconditioned MALA algorithm provided that the preconditioning metric has a bounded spectrum. □

[2] The truncation of the gradient is an assumption that has been used in a number of papers for the convergence analysis of Langevin-type sampling algorithms such as [15] although some recent works [52] suggest that, when the gradient of the target is not Lipchitz continuous, it may lead to a chain exhibiting poor mixing properties.

[3] $\mathrm{Ker}(\mathbf{H})$ and $\mathrm{Ker}(\mathbf{V}_{s_0})$ denote the nullspaces of $\mathbf{H}$ and $\mathbf{V}_{s_0}$, respectively.

Remark IV.2. *Let us point out that our Theorem IV.1 ensures the geometric ergodicity rate of the 3MH algorithm for distributions with tails decaying more rapidly than the exponential law. For distributions with heavier tails, the 3MH algorithm is likely to not be geometrically ergodic, since the Langevin diffusion itself fails to converge in an exponential rate in that case [18].*

V. Experimental results

To illustrate the benefits that can be drawn from the proposed algorithm, we will focus on two applicative examples, namely 1D signal deconvolution and multicomponent image denoising.

A. Sparse signal deconvolution with a Student-t prior

Our first example focuses on the deconvolution of a seismic signal. The original signal $\bar{\mathbf{x}}$ is a sparse vector of length $Q = 784$ composed of a sequence of spikes called primary reflection coefficients [57], [58] as depicted in Figure 2. The non-zero coefficients give information about the travel time of seismic waves between two seismic reflectors, and the amplitude of the seismic events reflected back to the sensor [57]. We assume that the signal is degraded by a known blur operator $\mathbf{H}^{Q\times Q}$ and further corrupted with an additive Gaussian noise. Thereby, the observation model (1) reduces to the following linear additive noise model: $\mathbf{z} = \mathbf{H}\overline{\mathbf{x}}+\mathbf{w}$, where $\mathbf{w}$ is an additive zero mean Gaussian noise with variance σ^2. The aim is then to retrieve an estimate $\widehat{\mathbf{x}}$ of $\overline{\mathbf{x}}$ from $\mathbf{H}$ and $\mathbf{z}$.

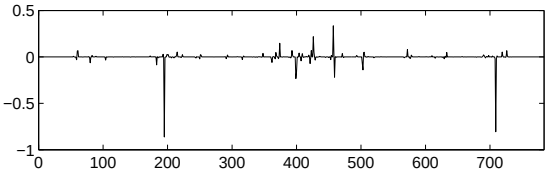

Fig. 2: Original signal.

1) Prior and posterior distributions: In order to promote the signal sparsity, we suppose that its coefficients are independent and identically distributed according to a Student-t ($\mathcal{ST}$) distribution. Then, we can identify the following functions:

$$\Phi = \frac{1}{2\sigma^2}\|\cdot\|^2 \text{ and } \Psi(\mathbf{x}) = \frac{\nu+1}{2}\sum_{i=1}^{Q}\log\left(\gamma^2+\frac{(x_i-\mu)^2}{\nu}\right). \tag{30}$$

Hereabove, $\Psi(\mathbf{x})$ refers to the negative logarithm of the $\mathcal{ST}$ prior on $\mathbf{x}$ of parameters ν, μ and γ. More specifically, $\nu > 0$ is the number of degrees of freedom determining the shape of the distribution, $\mu \in \mathbb{R}$ is the position parameter, and $\gamma > 0$ is the scale parameter [59]. It is worth to note that the Cauchy distribution is recovered as a particular case when $\nu = 1$. The $\mathcal{ST}$ distribution is often used in image reconstruction to model the distribution of wavelet coefficients [60]. This penalty has also been proposed in [61] as a tradeoff between the squared ℓ_2 norm and the non-convex approximation of the semi-norm ℓ_0 presented in [62], with the aim to enforce sparsity properties and better preserve discontinuities. Recall that the $\mathcal{ST}$ distribution can be written as a scale mixture of normal distribution where the hidden variable follows a gamma distribution with both parameters equal to $\nu/2$ [63]. In most Bayesian methods, it is generally used in this form: the unknown signal $\mathbf{x}$ and the hidden variable are estimated from their posterior joint distribution. In this work, we propose to directly use the expression defined in (30).

In the following, we assume that ν is known, and that we have only few prior information about the others parameters. Thus, the set of hyperparameters to be estimated jointly with $\mathbf{x}$ is $\boldsymbol{\Theta} = \{\mu, \gamma\}$. Since no explicit conjugate priors for these parameters are available, we propose to adopt simple and weakly informative priors. More specifically, uniform distributions are used for μ and γ defined on $[-\mu_m, \mu_M]$ and $[\gamma_m, \gamma_M]$ respectively, where μ_m, μ_M, γ_m and γ_M are positive constants. Thus, the posterior distributions of the parameters are given by

$$\mathsf{p}(\mu|\mathbf{x},\mathbf{z},\gamma) \propto \prod_{i=1}^{Q}\left(\gamma^2+\frac{(x_i-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}} 1_{[-\mu_m,\mu_M]}(\mu),$$

$$\mathsf{p}(\gamma|\mathbf{x},\mathbf{z},\mu) \propto \gamma^{Q\nu}\prod_{i=1}^{Q}\left(\gamma^2+\frac{(x_i-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}} 1_{[\gamma_m,\gamma_M]}(\gamma).$$

2) Sampling from the posterior distribution of the signal and the hyperparameters: Φ and Ψ satisfy the properties in Section III-C. We can thus apply the 3MH algorithm to sample from the posterior distribution of $\mathbf{x}$. More specifically, we will test the performance of 3MH using the three different curvature matrices proposed in Section III, namely $\mathbf{Q}_1$, the constant circulant matrix $\mathbf{Q}_2$, and the diagonal matrix $\mathbf{Q}_3$. In our context, these matrices are defined by

$$(\forall \mathbf{x}\in\mathbb{R}^Q)\quad \mathbf{Q}_1(\mathbf{x}) = \frac{1}{\sigma^2}\mathbf{H}^\top\mathbf{H} + \mathrm{Diag}\{\boldsymbol{\omega}(\mathbf{x})\} + \zeta\mathbf{I}_Q, \tag{31}$$

$$\mathbf{Q}_2 = \frac{1}{\sigma^2}\mathbf{H}^\top\mathbf{H} + \frac{\nu+1}{\nu\gamma^2}\,\mathbf{I}_Q, \tag{32}$$

$$(\forall \mathbf{x}\in\mathbb{R}^Q)\quad \mathbf{Q}_3(\mathbf{x}) = \mathrm{Diag}\left(\frac{1}{\sigma^2}\mathbf{L}^\top\mathbf{1}_N + \boldsymbol{\omega}(\mathbf{x})\right), \tag{33}$$

where $\boldsymbol{\omega}(\mathbf{x}) = (\omega_i(\mathbf{x}))_{1\leq i\leq Q}$ is such that

$$(\forall i\in\{1,\dots,Q\})\quad \omega_i(\mathbf{x}) = \frac{\nu+1}{\nu\gamma^2+(x_i-\mu)^2} \tag{34}$$

and $\mathbf{L} \in \mathbb{R}^{N\times Q}$ is given by (21). Hereabove, $\zeta \geqslant 0$ is a constant added to ensure the positive definiteness of the matrix $\mathbf{Q}_1$. Matrix $\mathbf{Q}_2$ is positive definite for every $\mathbf{x}\in\mathbb{R}^Q$. Furthermore, $\mathbf{Q}_3$ is also ensured to be positive definite for all $\mathbf{x}\in\mathbb{R}^Q$ provided that the observation matrix $\mathbf{H}$ contains no column whose elements are all equal to zero. It is worth noting that, the posterior density satisfies the sufficient conditions in Proposition IV.1 when $\mathbf{H}$ is injective. Similarly to MALA [18], the geometric ergodicity of 3MH is not theoretically guaranteed if $\mathbf{H}$ is not injective.

The posterior laws of the $\mathcal{ST}$ prior parameters do not have usual forms. Then, it is not easy to directly generate samples of μ and γ. We propose therefore to estimate them using a RW algorithm whose scale parameter is tuned automatically during

the burn-in period so as to reach an acceptance probability equal to 0.33.

3) Results: The test signal is artificially degraded by a band-pass filter with finite impulse response of length 41 with a frequency band concentrated between 10 and 40 Hz and an additive Gaussian noise of variance $\sigma^2 = 2.5 \times 10^{-3}$ (see Figure 2). The initial signal-to-noise ratio (SNR) is -4.58 dB[4]. We fix $\nu = 1$ which corresponds to the special case of the Cauchy prior. Simulations are performed on an Intel(R) Xeon(R) CPU E5-2630, @ 2.40 GHz, using a Matlab7 implementation. Figure 4 shows the error between the original signal and the degraded one as well as the error between the original signal and the restored one using the Minimum Mean Square Estimator (MMSE) which corresponds to a SNR equal to 8.24 dB.

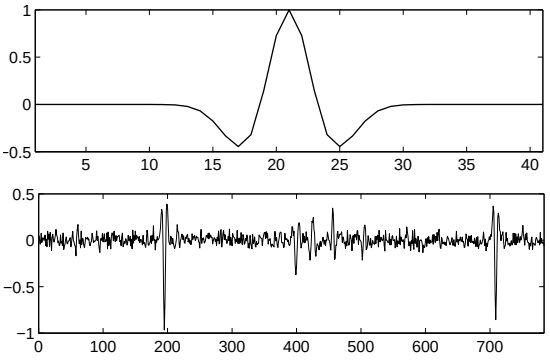

Fig. 3: Blurring kernel (top). Degraded signal (bottom).

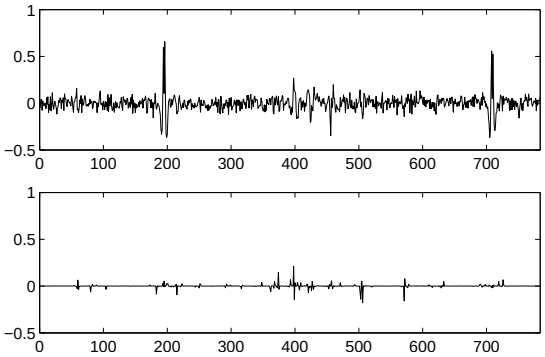

Fig. 4: Initial error $\bar{\mathbf{x}} - \mathbf{z}$ (top). Estimation error $\bar{\mathbf{x}} - \widehat{\mathbf{x}}$ (bottom).

We propose to compare the 3MH algorithm using the different curvatures matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ and the standard MALA algorithm. All tested algorithms have been run until convergence. The discretization stepsize ε is adjusted for all these algorithms during the burn-in period to correspond to an acceptance probability between 0.3 and 0.6. Note that in order to reduce the complexity of each iteration when using $\mathbf{Q} = \mathbf{Q}_1$, the inversion of the curvature matrix is performed in an approximate manner, using inner conjugate gradient iterations and the generation of random variables according to the proposal is then ensured using the sampling method from [64]. Table I summarizes the obtained samples for hyperparameters in terms of mean and standard deviation. Figure 5 shows the evolution of $\mathcal{J}$ with respect to the number of iterations and to the computational time. Following [15], we also compare the different methods in terms of the Mean Square Jump (MSJ) at stationarity which indicates how much the Markov chain is exploring the whole target space after convergence. Note that MSJ has been estimated with an empirical average over $T = 5000$ samples $\mathbf{x}^{(t_0+1)}, \ldots, \mathbf{x}^{(t_0+T)}$ generated after the burn-in period as follows

$$\text{MSJ} = \left(\frac{1}{T-1} \sum_{t=1}^{T-1} \|\mathbf{x}^{(t_0+t)} - \mathbf{x}^{(t_0+t+1)}\|^2 \right)^{1/2}. \quad (35)$$

It is worth noting that maximizing the MSJ is equivalent to minimizing a weighted sum of autocorrelations [65].

In Table II, we show estimates of the mean square jump per second at stationarity which is defined as the ratio of the mean square jump and the computational time per iteration. We also compare the statistical efficiency of the different samplers with respect to MALA defined as the mean square jump per second for each sampler over the mean square jump per second of MALA.

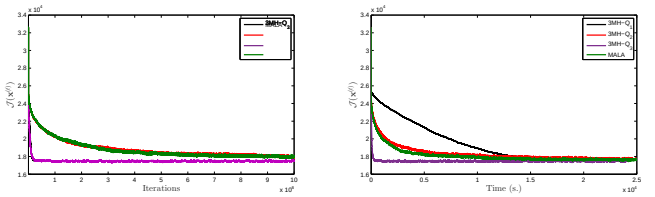

Fig. 5: Convergence speed of MALA, 3MH - $\mathbf{Q}_1$, 3MH - $\mathbf{Q}_2$ and 3MH - $\mathbf{Q}_3$ with respect to the number of iteration (lest) and computational time (right).

TABLE I: Mean and standard deviation of hyperparameter γ.

	MALA	3MH-Q_1	3MH-Q_2	3MH-Q_3
Mean	4.17 e-5	4.17 e-5	4.17 e-5	4.17 e-05
Std.	(4.94 e-8)	(4.92 e-8)	(4.93 e-8)	(4.92 e-8)

TABLE II: Mixing results for the different algorithms. First row: Estimates of the mean square jump at stationarity. Second row: Time per iteration in stationarity. Third row: Estimates of the mean square jump per second in stationarity. Fourth row: Efficiency relatively to MALA.

	MALA	3MH-$\mathbf{Q}_1$	3MH-$\mathbf{Q}_2$	3MH-$\mathbf{Q}_3$
MSJ	1.40 e-5	8.14 e-5	1.39 e-5	2.32 e-5
T(s.)	3.88 e-4	9.40 e-2	1.19 e-3	5.95 e-4
MSJ per s.	3.60 e-2	8.65 e-4	1.17 e-2	3.89 e-2
Efficiency	1	0.02	0.32	1.08

It can be noted that all the compared sample methods provide the same statistics for the estimated hyperparameter. Furthermore, one can notice that the behavior of 3MH

[4]Throughout this paper, the SNR is computed in dB as follows: $\text{SNR}=20\log_{10}\left(\frac{\|\bar{\mathbf{x}}\|}{\|\bar{\mathbf{x}}-\mathbf{u}\|}\right)$ where $\bar{\mathbf{x}}$ is the true signal and $\mathbf{u} = \mathbf{z}$ for the initial SNR while $\mathbf{u} = \hat{\mathbf{x}}$ for the final SNR (SNR of the restored signal).

algorithm using the constant curvature matrix $\mathbf{Q}_2$ is close to MALA in terms of convergence speed. This fact can be explained by the low dispersion of the eigenvalues of $\mathbf{Q}_2$ in this particular example. Nevertheless, the use of the matrix $\mathbf{Q}_1$ at each iteration becomes more expensive as the problem dimension increases which deteriorates the efficiency of the algorithm. The choice of the diagonal adaptive matrix $\mathbf{Q}_3$ appears to outperform the other algorithms due to the low complexity that it induces at each iteration. It allows to reach stability much faster than the other algorithms while achieving mixing properties slightly better than MALA at convergence.

B. Multispectral image denoising with a multivariate prior

In our second example, we are interested in the denoising of a multispectral image comprising B spectral channels with K pixels in each spectral image, corrupted with independent additive white Gaussian noises $\mathcal{N}(0, \sigma^2)$. We assume that the noise variance σ^2 is known. We denote by $\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_B$ the vectors that correspond to the reshaped unknown spectral images into vectors in $\mathbb{R}^K$. The objective is to recover these vectors from observed noisy vectors $\mathbf{z}_1, \ldots, \mathbf{z}_B$. Our recovery procedure will operate in the wavelet transform domain since the wavelet representations of the B unknown spectral images are sparse. To this end, we choose a set of orthogonal wavelet synthesis operators $\mathbf{F}_1^*, \ldots, \mathbf{F}_B^*$ [66]. More precisely, for each spectral position $b \in \{1, \ldots, B\}$, $\mathbf{F}_b^*$ is a linear mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$ that outputs $\bar{\mathbf{y}}_b$ from the vector $\bar{\mathbf{x}}_b \in \mathbb{R}^K$ of coefficients:

$$\bar{\mathbf{y}}_b = \mathbf{F}_b^* \bar{\mathbf{x}}_b. \tag{36}$$

Each spectral component with index b is decomposed into M subbands with sizes K_m, $m \in \{1, \ldots, M\}$ according to different orientations and resolutions. Obviously, we have $\sum_{m=1}^M K_m = K$ and the vector $\bar{\mathbf{x}}_b$ is defined by

$$\begin{aligned} \bar{\mathbf{x}}_b = \; & (\bar{x}_{b,1,1}, \ldots, \bar{x}_{b,1,K_1}, \ldots, \\ & \bar{x}_{b,m,1}, \ldots, \bar{x}_{b,m,K_m}, \ldots, \\ & \bar{x}_{b,M,1}, \ldots, \bar{x}_{b,M,K_M})^\top. \end{aligned} \tag{37}$$

Thus, the problem of recovering the multispectral image can be viewed as a special case of (1) expressed by $\mathbf{z} = \mathbf{H}\overline{\mathbf{x}} + \mathbf{w}$, where $N = Q = KB$, $\mathbf{z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_B^\top]^\top \in \mathbb{R}^N$, $\overline{\mathbf{x}} = [\overline{\mathbf{x}}_1^\top, \ldots, \overline{\mathbf{x}}_B^\top]^\top \in \mathbb{R}^Q$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2 \mathbf{I}_N)$ and, the matrix $\mathbf{H}$ is a block matrix formed by B blocks $\mathbf{F}_1^*, \ldots, \mathbf{F}_B^*$.

Our objective is to build an estimate $\hat{\mathbf{x}}$ of the frame coefficients $\overline{\mathbf{x}}$ based on the available observed frame coefficients $\mathbf{z}$ and the transform domain matrix $\mathbf{H}$.

1) Prior and posterior distributions: It is worth noting that the mutual similarities between the spectral images also propagate to their corresponding frame coefficients. Our idea is to capture such dependencies by resorting to a joint estimation of the frame coefficients of all the B components at a given orientation and resolution. To this end, for each subband $m \in \{1, \ldots, M\}$, we stack the coefficients of all the B channels at the same spatial position $k \in \{1, \ldots, K_m\}$ so as to build the vectors $\overline{\mathbf{x}}_{m,k} = (\overline{x}_{b,m,k})_{1 \leqslant b \leqslant B} \in \mathbb{R}^B$. Mathematically, it is easy to show that these vectors result from a linear transform of $\mathbf{x}$: $\overline{\mathbf{x}}_{m,k} = \mathbf{P}_{m,k}\overline{\mathbf{x}}$, where $\mathbf{P}_{m,k} \in \mathbb{R}^{B \times Q}$ denotes a sparse matrix containing B lines of an appropriate permutation matrix.

The sparsity of the frame coefficients and the spectral redundancies are captured by assuming that for each orientation and scale associated to index m, the vectors $\overline{\mathbf{x}}_{m,1}$, …, $\overline{\mathbf{x}}_{m,K_m}$ correspond to K_m realizations of a random vector whose distribution is a generalized multivariate exponential power distribution ($\mathcal{GMEP}$) characterized by its scale matrix $\mathbf{\Sigma}_m$, its shape parameter β_m and its smoothing parameter δ_m [67]. Consequently, the likelihood function Φ can be defined similarly to (30). Furthermore, the minus-log of the prior density is easily derived (up to an additive constant):

$$\Psi(\mathbf{x}) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \psi_m(\|\mathbf{\Sigma}_m^{-1/2}(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|) \tag{38}$$

where, for every $m \in \{1, \ldots, M\}$, $\mathbf{a}_m$ is a vector of $\mathbb{R}^B$ and for every real t, $\psi_m(t) = \frac{1}{2}(t^2 + \delta_m)^{\beta_m}$. It should be pointed out that for every $m \in \{1, \ldots, M\}$, the shape of the $\mathcal{GMEP}$ is controlled by β_m. A Gaussian distribution is associated to $\beta_m = 1$, while distributions with heavier tails correspond to $\beta_m < 1$. In particular, $\beta_m = 0.5$ corresponds to the multivariate Laplace distribution. In our work, for the sake of simplicity, the values of every β_m and δ_m are assumed to be known. These values change from a subband to another. Typically, small β_m values are assigned at the first scales in order to promote the frame coefficients sparsity while relatively higher values are chosen at higher scales. Finally, a normal distribution is often retained for the approximation subband at the coarsest scale. The value of δ_m is adjusted to a positive small value to guarantee the differentiability of Ψ given in (38). Furthermore, we decompose the scale matrix, for every $m \in \{1, \ldots, M\}$ as follows:

$$\mathbf{\Sigma}_m = \gamma_m^{-1/\beta_m} \mathrm{Diag}(\mathbf{n}_m)^{-1} \mathbf{R}_m \mathrm{Diag}(\mathbf{n}_m)^{-1}, \tag{39}$$

where $\mathbf{R}_m$ is the normalized correlation matrix of size $B \times B$ (with diagonal elements equal to 1 and the remaining ones correspond to the correlation factors between the coefficients), $\mathbf{n}_m$ is a B-dimensional vector of positive elements whose sum is equal to 1 and γ_m is a positive real. It is worth noting that $\gamma_m^{1/(2\beta_m)} \mathbf{n}_m$ can be seen as the vector containing the square root of the scale parameters for all the B components in subband m. For the sake of simplicity, we assume without loss of generality that the different spectral components of the image have the same correlation and weights in all subbands i.e., $\mathbf{R} = \mathbf{R}_m$ and $\mathbf{n}_m = \mathbf{n}$ for all m. Moreover, $\mathbf{n}$ and $\mathbf{R}$ are assumed to be known. Hence, the unknown hyperparameters form the set:

$$\mathbf{\Theta} = \{\gamma_1, \ldots, \gamma_M\}. \tag{40}$$

For every $m \in \{1, \ldots, M\}$, a gamma prior for γ_m is selected: $\gamma_m \sim \mathcal{G}(a_{\gamma_m}, b_{\gamma_m})$ where $a_{\gamma_m} > 0$ and $b_{\gamma_m} > 0$ [68].

Consequently, γ_m has the following posterior distribution:

$$\mathsf{p}(\gamma_m|\mathbf{x},\mathbf{R}) \propto \gamma_m^{a_{\gamma_m}+\frac{K_m}{2\beta_m}-1}\exp\left(-b_{\gamma_m}\gamma_m\right) \times \exp\left(-\frac{1}{2}\sum_{k=1}^{K_m}\left(\gamma_m^{\frac{1}{\beta_m}}\|\mathbf{R}^{-\frac{1}{2}}\mathrm{Diag}(\mathbf{n})(\mathbf{P}_{m,k}\mathbf{x}-\mathbf{a}_m)\|^2 +\delta_m\right)^{\beta_m}\right). \quad (41)$$

Our goal is to compute the posterior mean estimates of the target frame coefficients $\mathbf{x}$ as well as of $\boldsymbol{\Theta}$ thanks to MCMC sampling algorithms.

2) Sampling from the posterior distribution of the image and hyperparameters: Samples of vectors $\overline{\mathbf{x}}_{m,k}$ can be drawn in an independent manner for every $m \in \{1,\ldots,M\}$ and $k \in \{1,\ldots,K_m\}$. Indeed, as the posterior law is differentiable, we propose to apply Langevin based MCMC algorithms to produce samples according to the posterior law of $\overline{\mathbf{x}}_{m,k}$. Furthermore, adding a curvature matrix that accounts for the cross-spectral dependencies can improve the sampling performance. Note that its Hessian and Fisher matrices are equal because of the Gaussianity of the fidelity term. Nevertheless, the convexity of ψ_m only holds when $\beta_m \geqslant 0.5$ and, hence there is no guarantee that the Hessian and the Fisher matrices are definite positive if $\beta_m < 0.5$. For every $m \in \{1,\ldots,M\}$, ψ_m is differentiable and, the concavity on $\mathbb{R}_+$ of the function $t \mapsto \psi_m(\sqrt{t})$ is valid when $\beta_m \leqslant 1$. Consequently, we propose to employ the curvature matrices built by the MM strategy described in Section III-C. More precisely, we make use of the curvature matrix introduced in (17). Its expression for each subband m, is given by

$$(\forall \mathbf{c} \in \mathbb{R}^B)\ \ \mathbf{Q}_1^{(m)}(\mathbf{c}) = \frac{1}{\sigma^2}\mathbf{I}_B + \boldsymbol{\Sigma}_m^{-1}\frac{\dot{\psi}_m\left(\|\boldsymbol{\Sigma}_m^{-1/2}(\mathbf{c}-\mathbf{a}_m)\|\right)}{\|\boldsymbol{\Sigma}_m^{-1/2}(\mathbf{c}-\mathbf{a}_m)\|}. \quad (42)$$

Note that, the geometric ergodicity of the 3MH algorithm is fulfilled as $\mathbf{H}$ is injective. It is also worth pointing out that in the case when a normal distribution (i.e., $\beta_m = 1$) is assigned to the low frequency subband, it can be proved that the 3MH algorithm is still geometrically ergodic for a deconvolution problem (non necessarily injective $\mathbf{H}$) as $(\psi_m)_{1\leqslant m\leqslant M}$ satisfy the assumptions of Proposition IV.1.

Because of the unusual form of the posterior law of $\boldsymbol{\Theta}$, sampling from (41) is carried out by an independent MH algorithm with a gamma proposal of parameters $\tilde{a}_{\gamma_m} = a_{\gamma_m} + K_m/(2\beta_m)$, and

$$\tilde{b}_{\gamma_m} = b_{\gamma_m} + \sum_{k}^{K_m}\|\mathbf{R}^{-\frac{1}{2}}\mathrm{Diag}(\mathbf{n})(\mathbf{P}_{m,k}\mathbf{x}-\mathbf{a}_m)\|^{2\beta_m}. \quad (43)$$

3) Results: In our experiments, we select the Hydice[5] hyperspectral dataset containing 191 spectral components in the range [0.4,2.4]μm of the visible and infrared spectrum. From this dataset, we extract a portion of size $R = 256 \times 256$ over $B = 10$ spectral channels that we consider as our test image. Thus, the problem dimension amounts to $N = 655\,360$. We add a zero-mean white Gaussian noise with variance $\sigma^2 = 225$ to this test image. The initial SNR is 9.83 dB. We apply to the noisy image a four-stage orthonormal wavelet decomposition using a Symlet wavelet of order 3. Hence, we have $M = 13$ and $Q = N$. Regarding the approximation coefficients ($m = M$), we retain a Gaussian prior ($\beta_M = 1$, $\delta_M = 0$). For all the remaining subbands ($m \in \{1,\ldots,M-1\}$, we choose $\delta_m = 10^{-6}$. Moreover, in these experiments, in order to be able run comparisons with Newton MCMC algorithm, we will constrain the shape parameters β_m to be greater or equal than 0.5. Hence, the posterior distribution is strongly log-concave and the Hessian of the neg-log-likelihood is positive definite. More specifically, we set $\beta_m = 0.5$ for the wavelet coefficients at the two first lowest resolution levels, $\beta_m = 0.6$ at the third level of decomposition, and $\beta_m = 0.7$ at the coarsest level of decomposition. The Gibbs sampling algorithm is run with enough iterations to reach the stability state. The empirical MMSE estimator for the original image is computed with the generated samples of the wavelet coefficients after the burn-in period. Figure 6 displays the results achieved for the various components in terms of SNR and SSIM (Structural SIMilarity [69]). It appears that our method leads to a dramatic improvement of the values of the objective metrics and the perceptual ones for all the spectral components. For example, the average increase of the SNR (resp. SSIM) values approximately amounts to 10 dB (resp. 0.3). The resulting gains tend to indicate that the MMSE estimator leads to good numerical results. This is also corroborated by a visual inspection of the recovered components. Indeed, the reduction of the noise degradation in the different components is clearly noticeable in Figure 7. Besides, small details have been enhanced in a satisfactory manner.

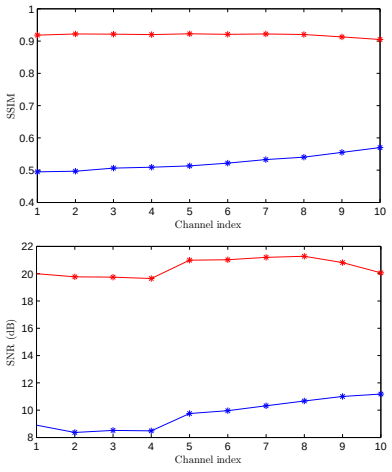

Fig. 6: SNR and SSIM values per spectral band for the $B = 10$ channels of the degraded (blue) and restored (red) images.

As benchmarking, we compare the performance of the Gibbs sampler when the posterior law of the wavelet coefficients is explored using either RW, MALA, Newton MCMC or 3MH algorithms. Simulations were performed on an Intel(R) Core(TM) i5-6300U CPU, @ 2.40 GHz, using a Matlab7 implementation. Figure 8 illustrates the evolution of the scale parameter γ_1 in the horizontal subband at the first level of

[5]https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html

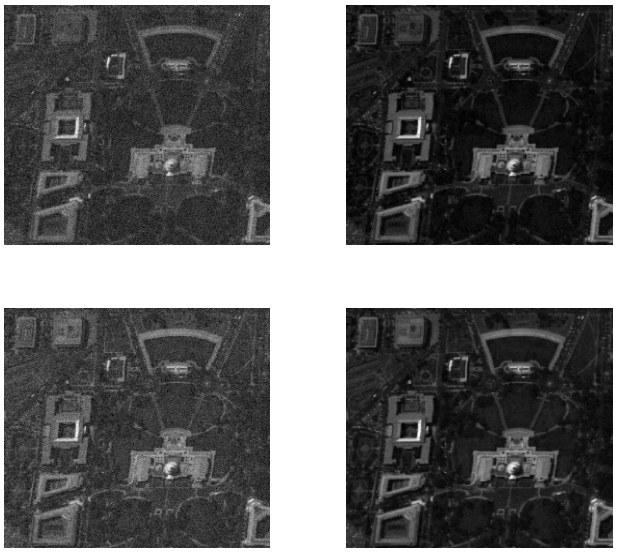

Fig. 7: From top to bottom: Components 1 and 10 of the degraded (left) and restored (right) images. SNR= (8.91 dB, 20 dB) (11.18 dB, 20.06 dB). SSIM=(0.3909 0.7734) (0.4881 0.7956).

decomposition with respect to the number of iterations and the computational time by employing the aforementioned algorithms. Table III provides the mixing results of the algorithms in terms of mean square jump per second in stationarity. Figure 9 shows the evolution of the SNR of the posterior mean with respect to the number of samples used to compute the empirical average. It can be noticed that the stationary state is reached by our proposed algorithm much faster than by RW, MALA and Newton MCMC algorithms. More precisely, about 1000 iterations and equivalently 500 seconds are enough for 3MH algorithm to reach stability which is fourfold less than the time required by MALA algorithm. It can be also noticed that Newton MCMC needs approximately the same number of iterations as MALA to reach stability. However, each iteration of Newton MCMC is more costly so that it turns out to be slower than MALA. The RW algorithm appears as the slowest algorithm since it needs more than 10 000 seconds to converge. Moreover, it is worth emphasizing that 3MH algorithm presents the best mixing properties at stability in terms of MSJ. The computational load of a single iteration of the 3MH algorithm is around twice higher than that of RW. Nonetheless, 3MH still appears as the most efficient choice after reaching stability compared with Newton, MALA and RW. This can also be deduced from Figure 9 since 3MH algorithm requires fewer samples to provide an accurate estimator compared to state-of-the-art algorithms. In particular, Newton MCMC presents poor mixing results compared to 3MH. This may be due to the bad conditioning of the Hessian matrix in this example. We further summarize the obtained samples by showing in Table IV the marginal means and standard deviations of the hyperparameters in the horizontal subbands over all the decomposition levels. It can be noted that all algorithms provide similar estimation results except

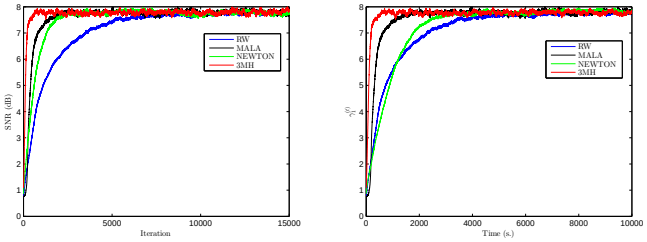

Fig. 8: Convergence speed of RW, MALA, Newton MCMC and 3MH with respect to the number of iteration (left) and computational time (right).

RW which gives slightly different variance estimation for the first level of decomposition. This can be related to the slow convergence of this algorithm which may not have reached yet the convergence as well as its poor mixing properties.

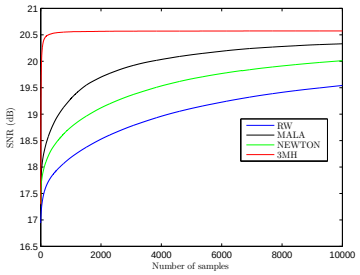

Fig. 9: Evolution of the SNR of the computed MMSE with respect to the number of samples.

TABLE III: Results for the different proposed algorithms. First row: Estimates of the mean square jump in stationarity. Second row: Time per iteration. Third row: Estimates of the mean square jump per second in stationarity. Fourth row: Relative efficiency compared to RW.

	RW	MALA	Newton	3MH
MSJ	1.40	2.28	1.90	4.49
T(s.)	0.69	0.73	1.29	0.91
MSJ per s.	2.02	3.10	1.46	4.93
Efficiency	1	1.53	0.72	2.43

TABLE IV: Mean and standard deviation of hyperparameters of the horizontal subbands over all the scales.

		RW	MALA	Newton	3MH
γ_1	Mean	2.49	2.53	2.52	2.53
	Std.	(2.61 e-2)	(2.04 e-2)	(2.21 e-2)	(2.04 e-2)
γ_2	Mean	0.97	0.96	0.97	0.97
	Std.	(1.13 e-2)	(1.18 e-2)	(1.16 e-2)	(1.22 e-2)
γ_3	Mean	0.31	0.31	0.31	0.31
	Std.	(6.64 e-3)	(6.76 e-3)	(6.78 e-3)	(6.87 e-3)
γ_4	Mean	0.10	0.10	0.10	0.10
	Std.	(3.50 e-3)	(3.45 e-3)	(3.53 e-3)	(3.39 e-3)

Another appealing property of the proposed Gibbs sampler concerns its straightforward extension to the case of

a deconvolution problem corresponding to $\mathbf{H} = \mathbf{D}\mathbf{F}^*$, the matrix $\mathbf{D}$ being a blurring operator. This extension can be realized by inserting an additional step in the Gibbs algorithm to draw samples of auxiliary variables [70]. Therefore, the deconvolution problem reduces a denoising type problem in the new augmented space. Using the 3MH algorithm to sample from the conditional distribution of $\mathbf{x}$ given the observations and the auxiliary variables, the preconditioning matrix would have the same form (42) as the considered denoising problem and would not be affected by the potential conditioning issues of the blur matrix.

VI. Conclusion

In this work, we have proposed a new MCMC algorithm that can be considered as a scaled MALA where the scale matrix is adapted at each iteration by following an MM strategy. We have shown that the geometric ergodicity property of the standard Langevin MH algorithms is maintained by introducing this scale matrix for the class of super-exponential distributions. We have then applied this algorithm to compute the MMSE estimator in signal and multicomponent image recovery problems. Experimental results emphasize the good performance of this new MCMC method compared to the standard MALA algorithm.

For future work, it would be interesting to derive more explicit expressions to assess the convergence speed of the proposed algorithm. This would likely depend on the conditioning number of the majorant matrix, controlled by $\nu_{\min}$ and $\nu_{\max}$ as well as on the computational complexity of each iteration. Another possible extension of our proposed proof is the adjustment of the improved convergence analysis proposed in [71] for the Riemann Hamiltonian MCMC sampling algorithms.

Appendix A
Proof of Proposition IV.1

Let $\mathbf{x} \in \mathbb{R}^Q \setminus \{0\}$. According to Assumption (ii), we have

$$\nabla \mathcal{J}(\mathbf{x}) = \mathbf{H}^\top (\mathbf{H}\mathbf{x} - \mathbf{z}) + \sum_{s=1}^{S} \mathbf{V}_s^\top (\mathbf{V}_s\mathbf{x} - \mathbf{c}_s) \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|}. \tag{44}$$

We deduce that[6]

$$\|\nabla \mathcal{J}(\mathbf{x})\| \leqslant \|\mathbf{H}\|\|\mathbf{H}\mathbf{x} - \mathbf{z}\| + \sum_{s=1}^{S} \|\mathbf{V}_s\| \dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|) \tag{45}$$

$$\leqslant \|\mathbf{H}\|(\|\mathbf{H}\mathbf{x}\| + \|\mathbf{z}\|) + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\|(\|\mathbf{V}_s\mathbf{x}\| + \|\mathbf{c}_s\|), \tag{46}$$

where the second inequality stems from Assumption III.1(iv). It follows from (44) that

$$\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x}) = \|\mathbf{H}\mathbf{x}\|^2 + \sum_{s=1}^{S} \|\mathbf{V}_s\mathbf{x}\|^2 \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|} + h(\mathbf{x}) \tag{47}$$

where

$$h(\mathbf{x}) = -\mathbf{x}^\top \left(\mathbf{H}^\top \mathbf{z} + \sum_{s=1}^{S} \mathbf{V}_s^\top \mathbf{c}_s \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|} \right). \tag{48}$$

Assume that $\mathbf{H}$ is injective. According to (47) and using Assumption III.1(iv), we obtain

$$\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|} \geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2}{\|\mathbf{x}\|} + \frac{h(\mathbf{x})}{\|\mathbf{x}\|} = \frac{\|\mathbf{H}\mathbf{x}\|^2}{\|\mathbf{x}\|} + O(1). \tag{49}$$

Then, (25) is satisfied. Moreover, using (46), we have

$$\begin{aligned} &\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|\|\nabla \mathcal{J}(\mathbf{x})\|} \\ &\geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2 + h(\mathbf{x})}{\|\mathbf{x}\|(\|\mathbf{H}\|(\|\mathbf{H}\mathbf{x}\| + \|\mathbf{z}\|) + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\|(\|\mathbf{V}_s\mathbf{x}\| + \|\mathbf{c}_s\|))} \\ &\geq \frac{\|\mathbf{H}\mathbf{x}\|^2}{(\|\mathbf{H}\|^2 + \|\mathbf{V}_s\|^2)\|\mathbf{x}\|^2}(1 + o(1)) + o(1). \end{aligned} \tag{50}$$

Thus, (26) also holds, and so does Assumption IV.1.

Let us now consider the case when $\mathbf{H}$ is not injective and our second set of assumptions applies. According to (47), we have

$$\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|} \geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2}{\|\mathbf{x}\|} + \frac{\|\mathbf{V}_{s_0}\mathbf{x}\|^2 \dot{\psi}_{s_0}(\|\mathbf{V}_{s_0}\mathbf{x} - \mathbf{c}_{s_0}\|)}{\|\mathbf{x}\|\|\mathbf{V}_{s_0}\mathbf{x} - \mathbf{c}_{s_0}\|} + O(1). \tag{51}$$

According to Assumption III.1(iii), $u \mapsto \dot{\psi}_{s_0}(u)/u$ is decreasing on $]0, +\infty[$ and, by using Assumption (ii), we deduce that there exists $\alpha_{s_0} > 0$ such that $(\forall u \in]0, +\infty[)\ \dot{\psi}_{s_0}(u)/u \geq \alpha_{s_0}$. Then, (51) yields

$$\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|} \geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2 + \alpha_{s_0}\|\mathbf{V}_{s_0}\mathbf{x}\|^2}{\|\mathbf{x}\|} + O(1). \tag{52}$$

It then follows from Assumption (i) that (25) is satisfied. Moreover, using (45) and (46), we have

$$\begin{aligned} &\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|\|\nabla \mathcal{J}(\mathbf{x})\|} \\ &\geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2 + h(\mathbf{x})}{\|\mathbf{x}\| \left(\|\mathbf{H}\|(\|\mathbf{H}\mathbf{x}\| + \|\mathbf{z}\|) + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\|(\|\mathbf{V}_s\mathbf{x}\| + \|\mathbf{c}_s\|) \right)} \\ &+ \frac{\frac{\|\mathbf{V}_{s_0}\mathbf{x}\|^2 \dot{\psi}_{s_0}(\|\mathbf{V}_{s_0}\mathbf{x} - \mathbf{c}_{s_0}\|)}{\|\mathbf{V}_{s_0}\mathbf{x} - \mathbf{c}_{s_0}\|}}{\|\mathbf{x}\| \left(\|\mathbf{H}\|\|\mathbf{H}\mathbf{x} - \mathbf{z}\| + \sum_{s=1}^{S} \|\mathbf{V}_s\| \dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|) \right)}. \end{aligned} \tag{53}$$

If $\mathbf{x} \in (\mathrm{Ker}(\mathbf{H}))^\perp$, we have

$$\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|\|\nabla \mathcal{J}(\mathbf{x})\|} \geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2}{a\|\mathbf{x}\|^2 + b\|\mathbf{x}\|} + o(1) \tag{54}$$

where

$$a = \|\mathbf{H}\|^2 + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\|^2 \tag{55}$$

$$b = \|\mathbf{H}\|\|\mathbf{z}\| + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\|\|\mathbf{c}_s\|. \tag{56}$$

[6] The spectral norm is employed for matrices.

Suppose now that $\mathbf{x} \in \mathrm{Ker}(\mathbf{H})$, then $\mathbf{x} \in (\mathrm{Ker}(\mathbf{V}_{s_0}))^\perp$. First note that since, for every $s \in \{1,\dots,S\}$, $u \mapsto \psi_s(u)/u$ is a nonnegative decreasing function on $]0,+\infty[$, there exists $\alpha_s \in [0,+\infty[$ such that $\lim_{u\to+\infty} \dot\psi_s(u)/u = \alpha_s$. According to Assumption (ii),

$$\lim_{u\to+\infty} \frac{\dot\psi_s(u)}{\dot\psi_{s_0}(u)} = \frac{\alpha_s}{\alpha_{s_0}} < +\infty. \tag{57}$$

In addition,

$$\begin{aligned}&\frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|\|\nabla\mathcal{J}(\mathbf{x})\|}\\ &\geqslant \frac{\|\mathbf{V}_{s_0}\mathbf{x}\|^2}{\|\mathbf{x}\|(\|\mathbf{V}_{s_0}\|\|\mathbf{x}\| + \|\mathbf{c}_{s_0}\|)}\ \frac{1}{\|\mathbf{V}_{s_0}\| + \epsilon(\mathbf{x})} + o(1),\end{aligned} \tag{58}$$

where

$$\epsilon(\mathbf{x}) = \frac{\|\mathbf{H}\|\|\mathbf{z}\| + \sum_{s\neq s_0} \|\mathbf{V}_s\| \dot\psi_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\dot\psi_{s_0}(\|\mathbf{V}_{s_0}\mathbf{x} - \mathbf{c}_{s_0}\|)} \geq 0. \tag{59}$$

Moreover, according to (57) and Assumption (ii), $\epsilon(\mathbf{x}) = O(1)$. Then, by using (54) and (58), we conclude that Condition (26) holds. Hence the result.

APPENDIX B
PROOF OF PROPOSITION IV.2

Let $\mathbf{x} \in \mathbb{R}^Q$ and $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{x} + \mathbf{b}(\mathbf{x})$. On the one hand,

$$\begin{aligned}-\log \mathbf{g}(\mathbf{x}|\mathbf{y}) = \frac{1}{2\varepsilon^2}\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2_{\mathbf{Q}(\mathbf{x})} - \frac{1}{2}\log|\mathbf{Q}(\mathbf{x})|\\ + \frac{Q}{2}\log(2\pi\varepsilon^2).\end{aligned} \tag{60}$$

From Assumption IV.2, we obtain

$$\nu_{\min}\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2 \leqslant \|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2_{\mathbf{Q}(\mathbf{x})} \leqslant \nu_{\max}\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2, \tag{61}$$

and

$$\nu^Q_{\min} \leqslant \|\mathbf{Q}(\mathbf{x})\| \leqslant \nu^Q_{\max}. \tag{62}$$

On the other hand, by using (28) and the triangle inequality, we have

$$\begin{aligned}\|\mathbf{y} - \mathbf{x}\| &\leqslant \|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\| + \|\boldsymbol{\mu}(\mathbf{x}) - \mathbf{x}\|,\\ &\leqslant \|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\| + \frac{\varepsilon^2}{2}\nu^{-1}_{\min} d.\end{aligned} \tag{63}$$

By using Jensen's inequality, it follows that

$$\|\mathbf{y} - \mathbf{x}\|^2 \leqslant 2\left(\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2 + \frac{\varepsilon^4}{4}\nu^{-2}_{\min} d^2\right). \tag{64}$$

Similarly, we have

$$\begin{aligned}\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\| &\leqslant \|\mathbf{y} - \mathbf{x}\| + \|\boldsymbol{\mu}(\mathbf{x}) - \mathbf{x}\|,\\ &\leqslant \|\mathbf{y} - \mathbf{x}\| + \frac{\epsilon^2}{2}\nu^{-1}_{\min} d\end{aligned} \tag{65}$$

and

$$\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2 \leqslant 2\left(\|\mathbf{y} - \mathbf{x}\|^2 + \frac{\varepsilon^4}{4}\nu^{-2}_{\min} d^2\right). \tag{66}$$

It follows from (61), (64) and (66) that

$$\begin{aligned}\frac{\nu_{\min}}{4\varepsilon^2}\|\mathbf{y} - \mathbf{x}\|^2 - \frac{\varepsilon^2 d^2}{8\nu_{\min}} &\leqslant \frac{1}{2\varepsilon^2}\|\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})\|^2_{\mathbf{Q}(\mathbf{x})}\\ &\leqslant \frac{\nu_{\max}}{\varepsilon^2}\|\mathbf{y} - \mathbf{x}\|^2 + \frac{\varepsilon^2\nu_{\max} d^2}{4\nu^2_{\min}}.\end{aligned} \tag{67}$$

Then, by using (60), (62) and (67), the lower bound in (29) holds for

$$k_1 = \left(\frac{\nu_{\min}}{2\nu_{\max}}\right)^{\frac{Q}{2}} \exp\left(-\frac{\varepsilon^2\nu_{\max} d^2}{4\nu^2_{\min}}\right), \quad \sigma_1^2 = \frac{\varepsilon^2}{2\nu_{\max}},$$

and Inequality (b) in (29) is satisfied for

$$k_2 = \left(\frac{2\nu_{\max}}{\nu_{\min}}\right)^{\frac{Q}{2}} \exp\left(\frac{\varepsilon^2 d^2}{8\nu_{\min}}\right), \quad \sigma_2^2 = \frac{2\varepsilon^2}{\nu_{\min}}.$$

REFERENCES

[1] A. C. Bovik, *Handbook of image and video processing*. Academic Press, 2010.

[2] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[3] M. Yan, "Restoration of images corrupted by impulse noise and mixed gaussian impulse noise using blind inpainting," *SIAM J. Imaging Sci.*, vol. 6, no. 3, pp. 1227–1245, 2013.

[4] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, 2003.

[5] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC Press Boca Raton, FL, 2014, vol. 2.

[6] M. Pereyra, P. Schniter, E. Chouzenoux, J.-C. Pesquet, J.-Y. Tourneret, A. O. Hero, and S. McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 224–241, Mar. 2016.

[7] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[8] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, ser. Springer Series in Statistics. Springer-Verlag, New-York, 2001.

[9] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, ser. Interdisciplinary Statistics. Chapman and Hall/CRC, 1999.

[10] D. Gamerman and H. F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, ser. Texts in Statistical Science. Chapman and Hall/CRC, 2006.

[11] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[12] A. Kong, "A note on importance sampling using standardized weights," *University of Chicago, Dept. of Statistics, Tech. Rep*, vol. 348, 1992.

[13] W. R. Gilks and P. Wild, "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, vol. 41, no. 2, pp. 337–348, 1992.

[14] G. O. Roberts and J. S. Rosenthal, "Optimal scaling for various Metropolis-Hastings algorithms," *Statistical Science*, vol. 16, no. 4, pp. 351–367, 2001.

[15] Y. F. Atchadé, "An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift," *Methodol. Comput. Appl. Probab.*, vol. 8, no. 2, pp. 235–254, 2006.

[16] G. O. Roberts, A. Gelman, and W. R. Gilks, "Weak convergence and optimal scaling or random walk Metropolis algorithms," *Ann. Appl. Probab.*, pp. 110–120, 1997.

[17] A. Beskos, G. Roberts, and A. Stuart, "Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions," *Ann. Appl. Probab.*, pp. 863–898, 2009.

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

[19] G. O. Roberts and J. S. Rosenthal, "Examples of adaptive MCMC," *J. Comput. Graph. Stat.*, vol. 18, no. 2, pp. 349–367, 2009.

[20] M. A. Stuart, J. Voss, and P. Wiberg, "Conditional path sampling of SDEs and the Langevin MCMC method," *Commu. Math. Sci.*, vol. 2, no. 4, pp. 685–697, 2004.

[21] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin diffusions," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 60, no. 1, pp. 255–268, 1998.

[22] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu, "Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2011)*, Prague, Czech Republic, 22-27 May 2011, pp. 3964–3967.

[23] M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 73, no. 91, pp. 123–214, Mar. 2011.

[24] Y. Zhang and C. A. Sutton, "Quasi-Newton methods for Markov chain Monte Carlo," in *Proc. Neural Information Processing Systems (NIPS 2011)*, no. 24, Granada, Spain, 12-17 Dec. 2011, pp. 2393–2401.

[25] J. Martin, C. L. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. 1460–1487, Jan. 2012.

[26] N. S. Pillai, A. M. Stuart, and A. H. Thierry, "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions," *Ann. Probab.*, vol. 22, no. 6, pp. 2165–2616, 2012.

[27] G. O. Roberts and O. Stramer, "Langevin diffusions and Metropolis-Hastings algorithms," *Methodol. Comput. Appl. Probab.*, vol. 4, no. 4, pp. 337–357, 2002.

[28] S. Gottlieb, C. W. Shu, and E. Tadmor, "Strong stability-preserving high-order time discretization methods," *SIAM Review*, vol. 43, no. 1, pp. 89–112, 2001.

[29] A. Durmus and E. Moulines, "Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm," *Ann. Appl. Probab.*, vol. 27, no. 3, pp. 1551–1587, 2017.

[30] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau," *SIAM J. Imaging Sci.*, vol. 11, no. 1, pp. 473–506, 2018.

[31] L. A. Breyer, M. Piccioni, and S. Scarlatti, "Optimal scaling of MALA for nonlinear regression," *Ann. Appl. Probab.*, vol. 14, no. 3, pp. 1479–1505, 2004.

[32] O. F. Christensen, G. O. Roberts, and J. S. Rosenthal, "Scaling limits for the transient phase of local Metropolis–Hastings algorithms," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 67, no. 2, pp. 253–268, 2005.

[33] T. Bui-Thanh and O. Ghatas, "A scaled stochastic Newton algorithm for Markov chain Monte Carlo simulations," Tech. Rep., 2012, http://users.ices.utexas.edu/ tanbui/PublishedPapers/SNanalysis.pdf.

[34] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.

[35] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 563–591, 2013.

[36] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794 – 816, Aug. 2016.

[37] M. W. Jacobson and J. A. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2411–2422, Oct. 2007.

[38] N. Pustelnik, A. Benazza-Benhayia, Y. Zheng, and J.-C. Pesquet, "Wavelet-based image deconvolution and reconstruction," *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2016.

[39] L. Chaâri, J.-Y. Tourneret, and H. Batatia, "Sparse bayesian regularization using bernoulli-laplacian priors," in *Proc. European Signal Processing Conf. (EUSIPCO 2013)*, Marrakech, Morocco, 9-13 Sep. 2013, pp. 1–5.

[40] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 5, pp. 1130–1142, 2006.

[41] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 298–311, 1997.

[42] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, no. 4, pp. 439–446, 1990.

[43] M. Zibulevsky and M. Elad, "ℓ_1– ℓ_2 optimization in signal and image processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 76–88, 2010.

[44] P. J. Huber, *Robust statistics*. Springer, 2011.

[45] A. Antoniadis, D. Leporini, and J.-C. Pesquet, "Wavelet thresholding for some classes of non–Gaussian noise," *Statistica Neerlandica*, vol. 56, no. 4, pp. 434–453, 2002.

[46] A. Repetti, E. Chouzenoux, and J.-C. Pesquet, "A penalized weighted least squares approach for restoring data corrupted with signal-dependent noise," in *Proc. Eur. Sig. Proc. Conf. (EUSIPCO 2012)*, Bucharest, Roumania, 27-31 Aug. 2012, pp. 1553–1557.

[47] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A convex approach for image restoration with exact Poisson-Gaussian likelihood," *SIAM J. Imag. Sci.*, vol. 8, no. 4, pp. 2662–2682, 2015.

[48] A. Halimi, Y. Altmann, A. McCarthy, X. Ren, R. Tobin, G. S. Buller, and S. McLaughlin, "Restoration of intensity and depth images constructed using sparse single-photon data," in *Proc. European Signal Processing Conf. (EUSIPCO 2016)*, Budapest, Hungary, 29 Aug.-2 Sep. 2016, pp. 86–90.

[49] S. Ganan and D. McClure, "Bayesian image analysis: An application to single photon emission tomography," *Amer. Statist. Assoc*, pp. 12–18, 1985.

[50] J. E. Dennis Jr and R. E. Welsch, "Techniques for nonlinear least squares and robust regression," *Commun. Stat. Simul. Comput.*, vol. 7, no. 4, pp. 345–359, 1978.

[51] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function," *J. Optim. Theory Appl.*, vol. 162, no. 1, pp. 107–132, 2014.

[52] M. Pereyra, "Proximal Markov chain Monte Carlo algorithms," *Statistics and Computing*, vol. 26, no. 4, pp. 745–760, 2016.

[53] S. F. Jarner and E. Hansen, "Geometric ergodicity of metropolis algorithms," *Stoch. Process. Their Appl.*, vol. 85, no. 2, pp. 341 – 361, 2000.

[54] A. Schreck, G. Fort, S. Le Corff, and E. Moulines, "A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 366–375, 2016.

[55] S. Allassonniere and E. Kuhn, "Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation," *Computational Statistics & Data Analysis*, vol. 91, pp. 4–19, 2015.

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

[57] A. Walden and J. Hosken, "The nature of the non-Gaussianity of primary reflection coefficients and its significance for deconvolution," *Geophysical Prospecting*, vol. 34, no. 7, pp. 1038–1066, 1986.

[58] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet, "Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1-ℓ_2 regularization," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 539–543, 2015.

[59] B. J. Winer, D. R. Brown, and K. M. Michels, *Statistical principles in experimental design*. McGraw-Hill New York, 1971, vol. 2.

[60] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders, "Variational Bayesian image restoration based on a product of t-distributions image prior," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1795–1805, 2008.

[61] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Med. Imag.*, vol. 8, no. 2, pp. 194–202, 1989.

[62] S. A. Geman and D. E. McClure, "Statistical methods for tomographic image reconstruction," *MIT Industrial Liaison Program Report*, vol. 8, pp. 5–21, 1987.

[63] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. R. Stat. Soc. Series B Stat. Methodol.*, pp. 99–102, 1974.

[64] F. Orieux, O. Féron, and J.-F. Giovannelli, "Sampling high-dimensional Gaussian distributions for general linear inverse problems," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 251–254, 2012.

[65] C. Sherlock, P. Fearnhead, and G. O. Roberts, "The random walk Metropolis: linking theory and practice through a case study," *Statistical Science*, pp. 172–190, 2010.

[66] C. Chaux, J.-C. Pesquet, and L. Duval, "Noise covariance properties in dual-tree wavelet decompositions," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4680–4700, Dec. 2007.

[67] Y. Marnissi, A. Benazza-Benyahia, E. Chouzenoux, and J.-C. Pesquet, "Generalized multivariate exponential power prior for wavelet-based multichannel image restoration," in *Proc. IEEE Int. Conf. on Image Process. (ICIP 2013)*, Melbourne, Australia, 15-18 Sep. 2013, pp. 2402–2406.

[68] D. Fink, "A compendium of conjugate priors," *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, p. 46, 1997.

[69] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[70] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, and J.-C. Pesquet, "An auxiliary variable method for MCMC algorithms in high dimension," *Entropy*, vol. 20, no. 110, pp. x–x+35, 2018.

[71] Y. T. Lee and S. S. Vempala, "Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation," in *Proc. Annual ACM Symp. on Theory of Comput. (STOC 2018)*, Los Angeles, CA, 25 - 29 Juin 2018, pp. 1115–1121.