MAJORIZE-MINIMIZE ADAPTED METROPOLIS–HASTINGS ALGORITHM. APPLICATION TO MULTICHANNEL IMAGE RECOVERY

Y. Marnissi[1], *A. Benazza-Benyahia*[2], *E. Chouzenoux*[1], *and J.-C. Pesquet*[1]

[1] Université Paris-Est, LIGM, UMR CNRS 8049, Champs sur Marne, France

[2] COSIM Lab., SUP'COM, Carthage Univ., Cité Technologique des Communications, Tunisia

ABSTRACT

One challenging task in MCMC methods is the choice of the proposal density. It should ideally provide an accurate approximation of the target density with a low computational cost. In this paper, we are interested in Langevin diffusion where the proposal accounts for a directional component. We propose a novel method for tuning the related drift term. This term is preconditioned by an adaptive matrix based on a Majorize-Minimize strategy. This new procedure is shown to exhibit a good performance in a multispectral image restoration example.

Index Terms— MCMC methods, Langevin diffusion, Majorize-Minimize, MMSE, multichannel image restoration.

1. INTRODUCTION

Recovering the signal of interest from degraded observations embedded in an additive noise is a key issue for many applications such as remote sensing imaging. In this respect, a Bayesian framework can be adopted to compute the Minimum Mean Squared Estimator (MMSE). However, it is not always possible to derive a closed expression of the posterior distribution involved in the MMSE. To alleviate this problem, Markov Chain Monte Carlo (MCMC) approaches have been developed. They consist of constructing an irreducible Markov chain whose stationary distribution is the unknown posterior distribution. Building the Markov chain corresponds to a specific way of exploring the state space. For this purpose, Metropolis–Hastings (MH) algorithms have been intensively used [1]. The key issue however is the choice of a proposal density. Recent methods such as those based on Langevin diffusion have incorporated a directional component for the proposal [2]. More precisely, two parameters (a stepsize and a scale matrix) are introduced to guide the directional component. The problem of setting the scale matrix must be carefully addressed, especially for high dimensional problems. Several solutions have been considered [2–4]. In this work, we propose a novel approach for choosing the scale matrix based on a Majorize-Minimize (MM) strategy.

The paper is organized as follows: In Section 2, we formulate the problem and we give a brief overview of the Langevin diffusion process. In Section 3, we describe the new MM adapted MH algorithm. Section 4 is devoted to experimental results for multicomponent image restoration. Finally, some concluding remarks are drawn in Section 5

2. RELATED WORKS

2.1. Problem statement

In this paper, we address a wide array of problems where the vector $\mathbf{x}$ in $\mathbb{R}^Q$ of samples of an unknown signal $\mathbf{X}$ is estimated from the observation vector $\mathbf{z}$ in $\mathbb{R}^N$ given the following observation model:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{1}$$

where the matrix $\mathbf{H} \in \mathbb{R}^{N\times Q}$ corresponds to a linear degradation operator, and $\mathbf{w}$ in $\mathbb{R}^N$ is the vector of a Gaussian additive noise samples. In this work, we adopt a Bayesian framework and we aim at computing the MMSE $\widehat{\mathbf{x}}_{\text{MMSE}} = \mathsf{E}_{\pi_{\mathbf{X}}}[\mathbf{x}]$ where the posterior distribution $\pi_{\mathbf{X}}$ is related to the prior distribution $\mathsf{p}_{\mathbf{X}}$ of the unknown vector $\mathbf{X}$:

$$\pi_{\mathbf{X}}(\mathbf{x}) \propto \mathsf{p}_{\mathbf{X}}(\mathbf{x})\mathsf{p}(\mathbf{z}|\mathbf{x}). \tag{2}$$

Generally, the computation of $\widehat{\mathbf{x}}_{\text{MMSE}}$ involves integrals that are both analytically and numerically intractable. The Monte Carlo approach is a classical alternative solution which consists of simulating a sufficient number of i.i.d. random variables from the posterior distribution $\pi_{\mathbf{X}}$ and approximating the MMSE estimator by the empirical average over all these samples. However, the target posterior is often complex and does not present a closed form, so that direct sampling is not always possible. To alleviate this difficulty, MCMC methods have been developed. They consist of building an irreducible Markov chain whose stationary distribution is $\pi_{\mathbf{X}}$. The asymptotic state of the chain is then considered as a sample of the target distribution. From an initial state $\mathbf{x}^0$, the problem reduces to exploring the state space according to the transition probabilities that characterize the Markov chain. Different ways of moving from a state to another have been reported. Among them, much attention was paid to MH algorithms that generate a random walk according to a proposal density and implement a method for rejecting proposed moves [1].

2.2. Langevin diffusion

The choice of the proposal distribution is crucial as it impacts the statistical properties of the resulting Markov chain especially for complex and high-dimensional target distributions. Advanced MH methods introduce a directional component for the proposal. In this respect, Langevin diffusion strategies adjust the state transition by accounting for the gradient direction of the target density. We have thus, for every $t \in \mathbb{N}$,

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \varepsilon^2\, \mathbf{b}(\mathbf{x}^t) + \varepsilon\, \sigma(\mathbf{x}^t)\mathbf{n}^{t+1}, \tag{3}$$

where $\varepsilon > 0$ is the stepsize resulting from Euler's discretization of the diffusion, $(\mathbf{n}^t)_{t\in\mathbb{N}}$ are realizations of a zero-mean white noise, $\sigma(\mathbf{x}^t)$ is a positive definite matrix and, $\mathbf{b}(\mathbf{x}) = (b_i(\mathbf{x}))_{i=1}^Q$ is a drift term. The latter is defined as follows:

$$\begin{aligned} b_i(\mathbf{x}) &= \tfrac{1}{2}\textstyle\sum_{j=1}^N \mathbf{A}_{ij}(\mathbf{x})\dfrac{\partial \log \pi_{\mathbf{X}}(\mathbf{x})}{\partial \mathbf{x}_j} \\ &+ |\mathbf{A}(\mathbf{x})|^{\frac{1}{2}} \textstyle\sum_{j=1}^N \dfrac{\partial}{\partial \mathbf{x}_j}\left(\mathbf{A}_{ij}(\mathbf{x})|\mathbf{A}(\mathbf{x})|^{-\frac{1}{2}}\right), \end{aligned} \tag{4}$$

where $\mathbf{A}(\mathbf{x}) = \sigma(\mathbf{x})\sigma^\top(\mathbf{x})$ and $|\mathbf{A}(\mathbf{x})|$ denotes the determinant of this matrix. It can be proved that the Langevin process has $\pi_{\mathbf{X}}$ as its stationary distribution and is more likely to accept proposed values than a standard random walk. 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, enables to achieve high acceptance rates. To this end, it is worth noting that the two scale parameters play an important role: ε determines the length of proposed jumps whereas $\mathbf{A}$ controls the direction. Three classes of algorithms have been developed from this diffusion depending on the choice of $\mathbf{A}$.

2.3. Choice of the scale matrix

The standard Metropolis adjusted Langevin algorithm (MALA) is the simplest form of this diffusion when $\mathbf{A}$ equals $\mathbf{I}_Q$, the identity matrix of $\mathbb{R}^Q$ [2]:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{t+1} = \mathbf{x}^t + \frac{\varepsilon^2}{2}\nabla \log \pi_{\mathbf{X}}(\mathbf{x}^t) + \varepsilon\, \mathbf{n}^{t+1}. \tag{5}$$

However, it should be noted that a bad adjustment of ε can significantly affect the convergence rate especially for large scale problems [3]. For this reason, many methods focus on how to choose a suitable stepsize such that the asymptotic average acceptance rate is bounded away from zero for high dimensions [3, 5]. Another approach consists of accelerating the algorithm by preconditioning the proposal density using a given constant scale matrix [4]. However, there is no clear guiding strategies for the selection of such a constant matrix. Recent algorithms [6–10] propose adaptive procedures where $\mathbf{A}$ is tuned according to the past behavior of the Markov chain resorting to some deterministic optimization tools. For example, when setting $\mathbf{A}$ to the inverse of the Hessian matrix of $-\log \pi_{\mathbf{X}}$ and, assuming a locally constant curvature, the term involving the derivatives of the scale matrix in (4) reduces to zero. Consequently, the computation of the drift term $\mathbf{b}$ becomes a scaled Newton step for minimizing $-\log \pi_{\mathbf{X}}$. 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 [6–8]. However, in practice, this method has a high computational load since it requires the computation of the Hessian matrix and its inverse at each iteration. This is particularly critical for large scale problems and/or when the Hessian matrix is not definite positive. One appealing solution is to replace the Hessian by a scale matrix that can efficiently accelerate the algorithm with a lower computational cost. In particular, many methods have proposed the Fisher information matrix as a preconditioning matrix in the Langevin diffusion [9, 10] which can be interpreted as the discretization of the MALA algorithm directly on a natural Riemannian manifold where the parameters live. In this work, we propose a new approach where the scale matrix of the Langevin diffusion is chosen according to a Majorize-Minimize strategy.

3. MAJORIZE-MINIMIZE ADAPTED METROPOLIS–HASTINGS

We focus on the case when the minus-log of the target density function $\Theta = -\log \pi_{\mathbf{X}}$ can be expressed as:

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \Theta(\mathbf{x}) = \Phi(\mathbf{H}\mathbf{x} - \mathbf{z}) + \Psi(\mathbf{x}), \tag{6}$$

where $\mathbf{z} \in \mathbb{R}^N$, $\mathbf{H} \in \mathbb{R}^{N\times Q}$, Φ is a continuous coercive differentiable function with an L-Lipschitzian gradient and

$$\Psi(\mathbf{x}) = \sum_{s=1}^S \psi_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|), \tag{7}$$

where $(\forall s \in \{1,...,S\})$ $\mathbf{V}_s \in \mathbb{R}^{P_s\times Q}$, $\mathbf{c}_s \in \mathbb{R}^{P_s}$ and $(\psi_s)_{1\leqslant s\leqslant S}$ is a set of positive continuous functions satisfying the following properties:

- $(\forall s \in \{1,...,S\})$ ψ_s is a differentiable function,
- $(\forall s \in \{1,...,S\})$ $\psi_s(\sqrt{\cdot})$ is concave over $\mathbb{R}^+$,
- $(\forall s \in \{1,...,S\})$ $(\exists \bar{\omega}_s > 0)$ such that $(\forall u > 0)$ $0 \leqslant \dot{\psi}_s(u) \leqslant \bar{\omega}_s u$ and $\lim\limits_{t\to 0} \dot{\psi}_s(u)/u < \infty$.

The minimization of (6) using the MM approach consists of performing successive minimizations of its tangent majorant functions [11]. Let $\mathbf{x}' \in \mathbb{R}^Q$. A function f is said to be a tangent majorant function of Θ at $\mathbf{x}'$ provided that

$$\begin{cases} f(\mathbf{x}', \mathbf{x}') = \Theta(\mathbf{x}'), \\ f(\mathbf{x}, \mathbf{x}') \geqslant \Theta(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^Q). \end{cases} \tag{8}$$

Let us assume 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 $\mathbf{x} \in \mathbb{R}^Q$,

$$f(\mathbf{x},\mathbf{x}') = \Theta(\mathbf{x}')+(\mathbf{x}-\mathbf{x}')^\top \nabla\Theta(\mathbf{x}')+\frac{1}{2}(\mathbf{x}-\mathbf{x}')^\top \mathbf{Q}(\mathbf{x}')(\mathbf{x}-\mathbf{x}') \tag{9}$$

is a tangent majorant of (6) at $\mathbf{x}'$. Then, the MM optimization algorithm reduces to building a sequence $(\mathbf{x}^t)_{t\in\mathbb{N}}$ through the following scheme:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{t+1} = \mathbf{x}^t + \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{x}^t)\nabla \log \pi_{\mathbf{X}}(\mathbf{x}^t), \tag{10}$$

with $\varepsilon \in (0, \sqrt{2}]$. According to the majorization properties (8), the MM update rule (10) will produce a monotically decreasing sequence $(\Theta(\mathbf{x}^t))_{t\in\mathbb{N}}$ that converges to a local minimum of Θ. We take up this idea to speed up the Langevin diffusion by using the inverse of the curvature matrix $\mathbf{Q}(\mathbf{x}^t)$ as a scale matrix in (4). Similarly to Newton-based MCMC methods, the drift term, assuming zero curvature changes, proposes, from a current state $\mathbf{x}^t$, a state with a higher value of $\log \pi_{\mathbf{X}}$, resulting from an iteration of MM algorithm on $-\log \pi_{\mathbf{X}}$. Then, the obtained proposal reduces to a noisy version of a MM iteration for minimizing $-\log \pi_{\mathbf{X}}$. Since the deterministic MM optimization approach can suffer from convergence to a local minimum in the nonconvex case, the addition of the noise component can solve this issue. The resulting 3MH sampler is described by Algorithm 1.

Algorithm 1: Majorize–Minimize adapted Metropolis–Hastings algorithm

```
0. Initialize $\mathbf{x}^0$, $t = 0$, $\varepsilon \in (0, \sqrt{2}]$
1. Compute $\mathbf{A}(\mathbf{x}^t) = \mathbf{Q}^{-1}(\mathbf{x}^t)$ and
   $\mathbf{g}(\mathbf{x}^t) = \nabla \log \pi_{\mathbf{X}}(\mathbf{x}^t)$
2. Generate $\mathbf{x}^* \sim q(\mathbf{x}^t, \cdot)$, where
   $q(\mathbf{x}^t, \cdot) = \mathcal{N}\left(\mathbf{x}^t + \frac{\varepsilon^2}{2}\mathbf{A}(\mathbf{x}^t)\mathbf{g}(\mathbf{x}^t), \varepsilon^2\mathbf{A}(\mathbf{x}^t)\right)$
3. Accept with probability
   $\alpha(\mathbf{x}^t, \mathbf{x}^*) = \min\left(1, \frac{\pi_{\mathbf{X}}(\mathbf{x}^*)q(\mathbf{x}^*, \mathbf{x}^t)}{\pi_{\mathbf{X}}(\mathbf{x}^t)q(\mathbf{x}^t, \mathbf{x}^*)}\right)$
3. Set $t \leftarrow t + 1$ and go to 1 until convergence.
```

There remains to define a set of suitable preconditioning matrices $\{\mathbf{Q}(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^Q}$. According to [12], convex quadratic tangent majorants of (6) can be obtained by using

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}(\mathbf{x}) = \mu\mathbf{H}^\top\mathbf{H} + \mathbf{V}^\top \mathrm{diag}\{\boldsymbol{\omega}(\mathbf{x})\}\mathbf{V} + \zeta\, \mathbf{I}_Q, \tag{11}$$

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

$$\omega_{P_1+P_2+\ldots+P_{s-1}+p}(\mathbf{x}) = \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|}. \tag{12}$$

Moreover, $\zeta \geqslant 0$ is a constant that ensures the invertibility of $\mathbf{Q}(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^Q$. In the context of large scale problems, the inversion of the curvature matrix (11) at each iteration may become intractable. We thus also propose to resort to the following alternative choice described in [13], which can be understood as a diagonal approximation of (11):

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}(\mathbf{x}) = (\mu\|\mathbf{H}\|^2 + \zeta)\mathbf{I}_Q + \mathrm{Diag}\left(\mathbf{P}^\top\boldsymbol{\omega}(\mathbf{x})\right), \tag{13}$$

where $\mathbf{1}_Q$ is the unit vector of $\mathbb{R}^Q$ and $\mathbf{P} \in \mathbb{R}^{P\times Q}$, with $P = \sum_s P_s$, is the matrix whose elements are given by

$$(\forall i \in \{1, \ldots, P\})(\forall j \in \{1, \ldots, Q\})\mathbf{P}_{i,j} = |\mathbf{V}_{i,j}| \sum_{k=1}^{Q} |\mathbf{V}_{i,k}|.$$

4. EXPERIMENTAL RESULTS

Objective The experiments we carried out deal with the recovery of a multicomponent image with B components degraded by a blur modelled by a linear operator $\mathbf{D}$ and an additive Gaussian noise $\mathbf{w}$ with covariance matrix $\boldsymbol{\Lambda}$. The restoration is performed in the wavelet transform domain: our objective is to compute the MMSE of $\mathbf{x} \in \mathbb{R}^Q$ through the linear model defined in (1) with $\mathbf{H} = \mathbf{D}F^*$ where F^* denotes a linear synthesis wavelet operator. Note that the wavelets coefficients are grouped into M subbands of size Q_m, $m \in \{1, \ldots, M\}$ and, for each subband m, we can extract the set of vectors $(\mathbf{x}_{m,q})_{q=1}^{Q_m}$ containing the wavelet coefficients located at the same spatial position q through all the B channels using a $B \times Q$ permutation matrix $\mathbf{P}_{m,q}$ such as $\mathbf{x}_{m,q} = \mathbf{P}_{m,q}\mathbf{x}$.

Prior distribution Similarly to [14], we assume that the vectors $(\mathbf{x}_{m,q})_{q=1}^{Q_m}$ are realizations of a random vector with a Generalized Multivariate Exponential Power (GMEP) distribution whose multivariate probability density function $\mathsf{p}_{\mathrm{GMEP}}$ is defined, for every $\mathbf{u}$ in $\mathbb{R}^B$, by

$$\mathsf{p}_{\mathrm{GMEP}}(\mathbf{u}; \theta_m) = C_m|\boldsymbol{\Sigma}_m|^{-1/2} g\left(\mathbf{u}^\top\boldsymbol{\Sigma}_m^{-1}\mathbf{u}; \beta_m, \delta_m\right), \tag{14}$$

where $\theta_m = \{\beta_m > 0, \delta_m > 0, \boldsymbol{\Sigma}_m\}$, for every $t \in \mathbb{R}_+$, $g(t; \beta_m, \delta_m) = \exp\left(-\frac{1}{2}(t + \delta_m)^{\beta_m}\right)$, and C_m is the associated normalization constant. $\boldsymbol{\Sigma}_m$ is related to the covariance matrix $\boldsymbol{\Gamma}_m$ through $\boldsymbol{\Sigma}_m = K^2_{\beta_m,\delta_m}\boldsymbol{\Gamma}_m$ with

$$K^2_{\beta_m,\delta_m} = B\frac{\int_0^\infty t^{\frac{B}{2}-1}e^{-\frac{1}{2}(t+\delta_m)^{\beta_m}}\,dt}{\int_0^\infty t^{\frac{B}{2}}e^{-\frac{1}{2}(t+\delta_m)^{\beta_m}}\,dt}. \tag{15}$$

The GMEP is an elliptical distribution that reflects the sparsity of the coefficients and, following [15] it can be proved that it is a scale mixture of Gaussian distributions when $\beta_m < 1$. Its mixing density is expressed as follows:

$$h_{\beta_m,\delta_m}(\nu) = \frac{2^{\frac{B}{2}}\Gamma(\frac{B}{2})\nu^{B-3}S_{\beta_m,\delta_m}(\frac{1}{2}\nu^{-2}, 2^{-\frac{1}{\beta_m}})}{\int_0^\infty t^{\frac{B}{2}-1}e^{-\frac{1}{2}(t+\delta_m)^{\beta_m}}\,dt} \tag{16}$$

where, for every $\alpha_1 \in (0,1)$, $\alpha_2 > 0$, $\sigma > 0$, $S_{\alpha_1,\alpha_2}(\cdot,\sigma) = e^{-\frac{1}{2}\alpha_2} S_{\alpha_1}(\cdot,\sigma)$, and $S_{\alpha_1}(\cdot,\sigma)$ is the alpha-stable density whose characteristic function is $\exp(-\sigma^{\alpha_1}|\cdot|^{\alpha_1} e^{-i\frac{\pi}{2}\alpha_1 \operatorname{sign}(\cdot)})$ [16].

Proposed priors for the hyperparameters In the following, we suppose that $(\delta_m)_{1\leqslant m\leqslant M}$ is known. We also assume that, for every $m \in \{1,\dots,M\}$, parameters β_m and $\boldsymbol{\Gamma}_m$ are independent and we denote by p_{β_m} and $\mathsf{p}_{\boldsymbol{\Gamma}_m}$ their respective prior density functions. Since wavelet coefficient images have leptokurtic histograms [17], we set $\mathsf{p}_{\beta_m} = \mathcal{U}(0,1)$. We use an inverse Wishart prior for $\boldsymbol{\Gamma}_m$, with parameters fixed according to a prior knowledge about $\boldsymbol{\Gamma}_m$.

Posterior distributions The posterior distributions of the GMEP hyperparameters have complicated form and there is no practical way for designing algorithms to simulate samples from them especially for the scale matrix. Since for every $m \in \{1,\dots,M\}$, $\beta_m \in (0,1)$, we propose to exploit the fact that GMEP is a scale mixture of normal distributions. Then, there exists a vector $\mathbf{v}_m = (v_{m,q})_{q=1}^{Q_m}$ of random variables $v_{m,q}$ such that $K_{\beta_m,\delta_m} v_{m,q} \sim h_{\beta_m,\delta_m}$ and, for all $q \in \{1,\dots,Q_m\}$, $\mathbf{x}_{m,q}$ is drawn independently from a zero mean Gaussian distribution with covariance matrix $v_{m,q}^2 \boldsymbol{\Gamma}_m$. Hence, the posterior distributions of $\boldsymbol{\Gamma}_m$ reduces to an inverse Wishart distribution and the posterior distribution of β_m is related to a product of densities of stable distributions [15].

Let $\boldsymbol{\Omega} = \mathbf{H}^\top \boldsymbol{\Lambda}^{-1} \mathbf{H} + \sum\limits_{m=1}^{M} \sum\limits_{q=1}^{Q_m} v_{m,q}^{-2} \mathbf{P}_{m,q}^\top \boldsymbol{\Gamma}_m^{-1} \mathbf{P}_{m,q}$ then the posterior distribution of $\mathbf{x}$ reduces to a normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{\Omega}^{-1}\mathbf{H}^\top \boldsymbol{\Lambda}^{-1}\mathbf{z}$ and covariance matrix $\boldsymbol{\Omega}^{-1}$. Note that sampling from high-dimensional Gaussian distributions is often very difficult since matrix factorization (Cholesky, QR, square root) is not always possible because of its high computation cost and/or memory requirements. Some solutions have been proposed for some special structures of the covariance matrix (circular, sparse,…) [18, 19]. In this work, we propose to use a step of the 3MH algorithm with the diagonal curvature matrix. The interest of this alternative solution is that it does not require any assumption on the structure of the covariance matrix.

Note that we follow the method proposed in [15] for the sampling of $v_{m,q}$. Moreover, MH steps are used to generate samples from the posterior distribution β_m.

Results The test image is a remote sensing multispectral SPOT image of size 128 × 128 with three components ($B = 3$) and, corresponding to a scene depicting the city of Tunis. This image is artificially blurred with a cosine blur FTM and corrupted with a zero-mean white Gaussian noise with a variance adjusted so as to correspond to an initial averaged Blurred Signal-to-Noise Ratio (BSNR) equal to 21.64 dB. We apply a 2-resolution wavelet orthonormal decomposition using a Symmlet wavelet transform of order 8. We run the hybrid Gibbs sampler for 8,000 iterations, reject the 6,000 first ones as a burn-in, and take the last 2,000 results as samples for the target data. Fig. 1 provides the evolution of improved SNR (ISNR) with respect to the computational time using different algorithms to generate $\pi_{\mathbf{x}}$, when performing tests on an Intel Core i7 CPU, @ 3.00 GHz and using a Matlab 7.12 implementation. It can be observed that the 3MH algorithm reaches stability faster than MALA. In fact, MALA algorithm requires less time per iteration but our algorithm converges in a significantly smaller number of iterations. The obtained samples of the wavelet coefficients are then used to compute the empirical MMSE estimator for the original image. We have used the hyperparameters estimated by the Gibbs sampler to run the method described in [14] which computes the MAP estimate with a GMEP prior using the MM Memory Gradient algorithm. Table 1 reports the results obtained for the different components in terms of SNR, BSNR and Structural Similarity Index (SSIM). It can be observed that the MMSE estimator shows better performance than the MAP. This can also be observed on Fig. 2 showing the visual improvement for the first component of the image.

Table 1. Restoration results

		$b=1$	$b=2$	$b=3$	Average
Initial	BSNR	22.32	20.29	22.30	**21.64**
	SNR	21.72	19.56	21.96	21.08
	SSIM	0.729	0.761	0.720	0.737
MAP	BSNR	26.76	24.16	26.48	**25.80**
	SNR	24.95	22.33	25.19	24.16
	SSIM	0.860	0.863	0.843	0.855
MMSE	BSNR	27.34	24.75	27.06	**26.38**
	SNR	25.20	22.59	25.51	24.43
	SSIM	0.872	0.874	0.855	0.867

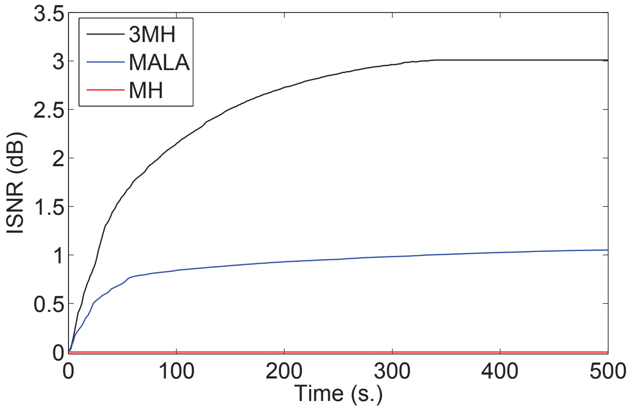
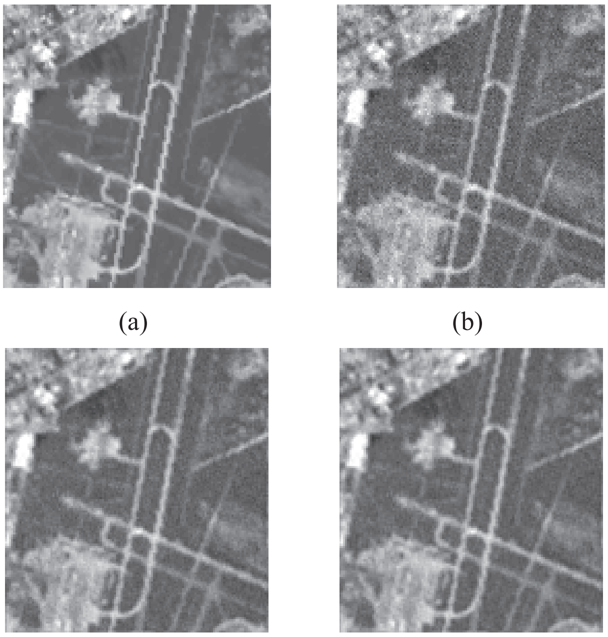


Fig. 1. Convergence speed of 3MH, MALA and MH.



(a) (b)

(c) (d)

Fig. 2. (a) Original image, (b) degraded version of the first component (SNR = 21.72 dB, SSIM= 0.729), (c) restored version using the MAP estimator (SNR =24.95 dB, SSIM = 0.860), (d) restored version using the MMSE estimator (SNR =25.20 dB, SSIM = 0.872).

5. CONCLUSION

In this paper, we have proposed a new MCMC algorithm that can be considered as a scaled MALA where the scale matrix is adapted at each iteration with a MM strategy. We have applied this algorithm to compute the MMSE estimator of a multicomponent image from its blurred version. Experimental results indicate the good performance of this new MCMC method. Note that the proposed approach can be applied to general models where the posterior distribution is non-Gaussian.

REFERENCES

[1] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

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

[3] 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.

[4] 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.

[5] G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin diffusions," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 60, pp. 255–268, 1997.

[6] 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, 2012.

[7] Y. Zhang and C. A. Sutton, "Quasi-Newton methods for Markov chain Monte Carlo," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, Eds., 2011, pp. 2393–2401.

[8] 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`.

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

[10] 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.

[11] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statis.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.

[12] 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.

[13] 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 and Appl.*, 2013, to appear.

[14] 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. Image Process. (ICIP 2013)*, Melbourne, Australia, 15-18 Sep. 2013.

[15] E. G. Sanchez Manzano, M. A. G. Villegas, and J. M. Marin, "Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications," *Comm. Statist. Theory Methods*, vol. 37, no. 6, pp. 972–982, Mar. 2008.

[16] E. G. Sanchez Manzano, M. A. G. Villegas, and J. M. Marin, "Sequences of elliptical distributions and mixtures of normal distributions," *J. Multivariate Analysis*, vol. 97, no. 2, pp. 295–310, Jan. 2006.

[17] E. P. Simoncelli, "Bayesian denoising of visual images in the wavelet domain," in *Bayesian Inference in Wavelet Based Models*, P. Müller and B. Vidakovic, Eds., vol. 141 of *Lectures Notes in Statistics*, pp. 291–308. Springer, 1999.

[18] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 5, pp. 959–963, 1985.

[19] 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.