AN AUXILIARY VARIABLE METHOD FOR LANGEVIN BASED MCMC ALGORITHMS

Yosra Marnissi[1]*, Emilie Chouzenoux*[1]*, Jean-Christophe Pesquet*[1]*, and Amel Benazza-Benyahia*[2]

[1] Université Paris-Est, LIGM, UMR CNRS 8049, Marne-la-Vallée, France
[2] COSIM SUP'COM, Université de Carthage, Cité Technologique des Communications, Tunisie

ABSTRACT

Markov Chain Monte Carlo sampling algorithms are efficient Bayesian tools to explore complicated posterior distributions. However, sampling in large scale problems remains a challenging task since the Markov chain is very sensitive to the dependencies between the signal samples. In this paper, we are mainly interested in Langevin based MCMC sampling algorithms that allow us to speed up the convergence by controlling the direction of sampling and/or exploiting the correlation structure of the target signal. However, these techniques may sometimes fail to explore efficiently the target space because of poor mixing properties of the chain or the high cost of each iteration. By adding some auxiliary variables, we show that the resulting conditional distribution of the target signal is much simpler to explore by using these algorithms. Experiments performed in the context of multicomponent image restoration illustrate that the proposed approach can achieve substantial performance improvement compared with standard algorithms.

1. INTRODUCTION

In many applications, because of the imperfection of acquisition systems, the observed data is a degraded version of the original one. Throughout this paper, we will consider the following degradation model, in which the observation $\mathbf{z} \in \mathbb{R}^N$ is related to some unknown random signal of interest $\mathbf{x} \in \mathbb{R}^Q$ by

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

where the matrix $\mathbf{H} \in \mathbb{R}^{N \times Q}$ corresponds to a linear degradation operator, eventually combined with a linear transform (e.g., Fourier transform), and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2)$ models some acquisition noise. In the following, we address the problem of estimating $\mathbf{x}$ from the observations $\mathbf{y}$.

To obtain meaningful solutions, the problem must be regularized using additional informations about the unknown signal $\mathbf{x}$. In the Bayesian setting, the prior knowledge about $\mathbf{x}$ is modeled by a prior distribution $\mathsf{p}(\mathbf{x})$. To infer the unknown signal, we can rely on the posterior distribution $\mathsf{p}(\mathbf{x}|\mathbf{y})$ expressed by

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

Let us denote by $\mathcal{J}$ the minus-log of $\mathsf{p}(\mathbf{x}|\mathbf{z})$. The most commonly used inferences are the Maximum A Posteriori (MAP) and the Minimum Mean Square Error (MMSE) estimators [1]. On the one hand, the MAP estimator is defined as the mode of the posterior distribution, or, equivalently, as the minimizer of $\mathcal{J}$. On the other hand, the MMSE estimator aims at finding a solution that minimizes the expected quadratic estimation error and thus reduces to the computation of the posterior mean. However, the exact posterior mean cannot be calculated since, in most cases, it can only be known up to a multiplicative constant. Moreover, when the problem dimension Q is large, the computation of the involved integral becomes intractable. An alternative is to use Monte Carlo Markov Chain (MCMC) methods [2–5]. The general idea of such methods is as follows: even if it is difficult to draw directly independent samples from a complicated distribution π, one can often find a way of constructing an irreducible Markov chain whose stationary distribution is π so that, after a sufficient number of iterations, the samples drawn by the MCMC algorithm follow the distribution of interest. From these samples, a good summary about the desired distribution is obtained, from which one can compute efficient estimates of useful statistics such as the mean and the variance. MCMC algorithms only need, in general, to know the expression of the target density up to an additive multiplicative constant, which is actually the case of most of posterior distributions encountered in the context of inverse problems. However, many standard MCMC algorithms become inefficient when the problem dimension becomes large [6, 7]. In fact, their convergence may be very slow and they tend to generate samples that are highly correlated and thus fail to explore the full parameter space, leading to an incomplete description of the target density. Some solutions have been devised to alleviate this difficulty. Many works have proposed to use some local information about the target density to explore the parameter space more efficiently in order to speed up the convergence of the algorithms. This information can be included using the first order derivatives of the posterior logarithm which leads to the famous class of the Langevin based MCMC algorithms [8, 9]. Other works have proposed to fight against slow mixing by introducing some auxiliary variables to the model which have shown to exhibit good performance [10–13]. For instance, two well known examples of auxiliary variable MCMC methods, are hybrid Monte Carlo [14] and slice sampling [15, 16].

In this paper, we propose to combine these two approaches to construct a new sampling algorithm that remains efficient in high dimensional problems. In Section 2, we will do a survey on Langevin based MCMC sampling methods and emphasize the main difficulties encountered with these algorithms. In Section 3, we show how to integrate auxiliary variables methods in Langevin based MCMC sampling algorithms in order to improve their efficiency. In Section 4, we present an application of our method through an example of multispectral

image restoration, before drawing some conclusions.

2. MARKOV CHAIN MONTE CARLO SAMPLING ALGORITHMS

2.1. Metropolis-Hastings algorithm

One the most commonly used MCMC samplers is the Metropolis Hastings (MH) algorithm. In order to draw a sample from a target distribution π, it generates a sample according to some proposal density q and accept or reject it with an acceptance probability related to π [17,18]. Note that the choice of the proposal distribution q has a critical impact on the performance of the MH algorithm. Indeed, it should be both a good approximation of the target density and it should be easy to sample from. Various MCMC algorithms have been proposed as special instances of MH algorithm, based on specific choices for the proposal distribution. One simple proposal is provided by the random walks algorithm for which $q(\mathbf{x},.) = \mathcal{N}(\mathbf{x},\varepsilon^2)$ with $\varepsilon > 0$. This proposal is easy to implement, but it does not tolerate large moves in the parameter space. Hence, the convergence is generally slow especially in large dimensional problems and the resulting samples may be highly correlated. Other sophisticated algorithms use some informations about the target density to construct appropriate candidate moves. Here, we will focus on the particular case of Langevin-based MCMC algorithms, that we describe below.

2.2. Langevin based MCMC

Langevin-based MCMC algorithms construct proposals that account for a directional component pushing the chain towards areas of high probability (i.e. where most samples lie) with the aim to accelerate the convergence of the algorithm [8]. This is achieved by using the gradient direction of $\mathcal{J}$ at the current state, in combination with some symmetric definite positive matrix $\mathbf{Q}$ that may reflect possible correlation structures between coefficients. Hence, using a discrete-time forward Euler approximation of the Langevin diffusion on $\mathbb{R}^Q$, Langevin-based MCMC algorithms propose samples at each iteration t according to the following scheme:

$$\tilde{\mathbf{x}} \sim \mathcal{N}\left(\mathbf{x}^t - \frac{\varepsilon^2}{2}\mathbf{Q}(\mathbf{x}^t)^{-1}\nabla\mathcal{J}(\mathbf{x}^t), \varepsilon^2\mathbf{Q}(\mathbf{x}^t)^{-1}\right), \quad (3)$$

where $\mathbf{x}^t$ is the current state and ε is a positive constant. In the simple case when $\mathbf{Q}(\cdot) \equiv \mathbf{I}_Q$, we recover the standard MALA algorithm [8,9,19]. It has been shown, in recent works [20–24], that the use of varying preconditioning matrices allows to accelerate MALA algorithm. These matrices are usually tuned according to the past behavior of the Markov chain, using some deterministic optimization strategies. For example, when $\mathbf{Q}(\mathbf{x}^t)$ is chosen at each iteration to be the Hessian matrix of $\mathcal{J}$ at $\mathbf{x}^t$, the drift term of the algorithm reduces to a scaled Newton step for minimizing $\mathcal{J}$ [23,24]. However, in practice, such strategy has a very high computational load since it requires the computation of the Hessian matrix and its inverse at each iteration. Moreover, in many scenarios, the Hessian matrix is not positive definite or well conditioned which may cause numerical problems. One appealing alternative has been introduced in [25]. It proposes preconditioning matrices tuned according to a Majorize-Minimize approach. This method offers a great flexibility since, under some mild hypotheses, it is always possible to define a quadratic function majorizing $\mathcal{J}$ that approximate well the target density and whose curvature matrix can be computed with a lower computational cost [25, 26].

2.3. Discussion

As already pointed out, the numerical efficiency of Langevin-based MCMC sampling algorithms relies both on the use of a proposal providing tight approximations of the target density and that are simple to sample from. Let us focus on the resolution of the inverse problem (1). If the coefficients of the unknown signal $\mathbf{x}$ are not correlated, and the dimension of the problem is low, the best strategy is probably to employ the standard MALA algorithm [8]. However, in large scale problems, most of posterior distributions arising in real world applications exhibit strong dependencies between the coefficients of the signal. In this case, standard MALA algorithm may fail to provide a well mixed chain. One can instead expect better numerical performance with more sophisticated scaling matrices that make possible large moves in the directions that reflect the dependence structure. However, in this case, sampling from the proposal becomes generally complicated in practice as the problem dimension increases due to the high cost of each iteration since the curvature matrix becomes difficult to handle. Consequently, the main difficulty for all these algorithms is related to the presence of two different sources of correlations which may come from the likelihood or from the prior information. The operator $\mathbf{H}$ in the likelihood may induce high interactions between coefficients on a very wide neighborhood even when the coefficients are supposed independent in the prior law. One can pass to another domain where $\mathbf{H}$ can be diagonalised, for example the Fourier domain when $\mathbf{H}$ is circulant. However, the problem remains when we take into account the prior dependencies between the coefficients as the prior covariance matrix cannot usually be diagonalised in the same domain as $\mathbf{H}$. One should therefore handle these two sources of correlations separately. In the following, we propose to alleviate this problem by adding some auxiliary variables to the model.

3. INTRODUCING AUXILIARY VARIABLES

An important class of MCMC methods is based on a remarkable trick: replace an initial difficult problem by a higher dimensional one (having more variables) but that is easier to solve than the original problem. Thus, to generate samples from π, an auxiliary variable $\mathbf{u}$ is added with a given conditional distribution $\mathsf{p}(\mathbf{x}|\mathbf{u})$. A Markov chain is then constructed by alternating the update of $\mathbf{u}$ and $\mathbf{x}$ i.e drawing samples from their conditional distribution. Such an MCMC scheme may lead to conditional distributions that are easier to simulate and may achieve substantial gain in terms of efficiency and mixing properties compared with standard methods.

Let us consider the problem model defined in (1) and define the auxiliary variable $\mathbf{u} \in \mathbb{R}^Q$ such that the conditional distribution of $\mathbf{u}$ given $\mathbf{x} \in \mathbb{R}^Q$ is

$$\mathsf{p}(\mathbf{u}|\mathbf{x}) = \mathcal{N}\left((\mathbf{I}_Q + \mathbf{C})^{-1}\mathbf{x}, \sigma^2\alpha(\mathbf{I}_Q + \mathbf{C})^{-1}\right), \quad (4)$$

where $\mathbf{C} = \mathbf{B}(\mathbf{I}_Q - \mathbf{B})^{-1}$, $\mathbf{B} = \alpha \mathbf{H}^* \mathbf{H}$ and $\alpha > 0$ is a chosen constant such that $\alpha \|\mathbf{H}^* \mathbf{H}\| < 1$. Note that $(\mathbf{I}_Q + \mathbf{C})^{-1} = \mathbf{I}_Q - \alpha \mathbf{H}^* \mathbf{H}$. Then, the minus-log of the joint distribution of $\mathbf{x}$ and $\mathbf{u}$ is given by

$$f(\mathbf{x}, \mathbf{u}) = \frac{1}{2\sigma^2 \alpha} \left(\|\mathbf{u} - \mathbf{x}\|^2 + \mathbf{u}^\top \mathbf{C}^\top \mathbf{u} - 2\alpha \, \mathbf{z}^\top \mathbf{H} \mathbf{x} \right) - \log \mathsf{p}(\mathbf{x}). \quad (5)$$

It can be noticed that computing the global minimizer of $\mathcal{J}$ reduces to minimize each of the partial functions $f(\mathbf{x}, .)$ and $f(., \mathbf{u})$. For this reason, this method has been used in some variational applications for image restoration in [27]. Moreover, this technique has been adopted to facilitate sampling using classical Metropolis-Hastings algorithm and Gibbs sampler in the maximum likelihood estimation approach proposed in [28]. From (5), it follows that the minus-log of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ and $\mathbf{u}$ is defined up to an additive constant as follows

$$\mathcal{J}(\mathbf{x}|\mathbf{u}) = \frac{1}{2\sigma^2 \alpha} \left(\|\mathbf{u} - \mathbf{x}\|^2 - 2\alpha \, \mathbf{z}^\top \mathbf{H} \mathbf{x} \right) - \log \mathsf{p}(\mathbf{x}). \quad (6)$$

Hence, in (6), the original problem now reduces to solve a denoising problem where the variance of the noise is $\sigma^2 \alpha$. Another step is added in the MCMC algorithm to sample the auxiliary variable $\mathbf{u}$ from (4). The main steps of the proposed algorithm can be summarized as follows:

For $t = 0, 1, \ldots$

Step 1: Generate

$\tilde{\mathbf{x}} \sim \mathcal{N}\left(\mathbf{x}^t - \frac{\varepsilon^2}{2} \mathbf{Q}(\mathbf{x}^t)^{-1} \nabla \mathcal{J}(\mathbf{x}^t | \mathbf{u}^t), \varepsilon^2 \mathbf{Q}(\mathbf{x}^t)^{-1} \right)$

Step 2: Accept ($\mathbf{x}^{t+1} = \tilde{\mathbf{x}}$) with probablity

$\alpha(\mathbf{x}^t, \tilde{\mathbf{x}}) = \min\left(1, \frac{\pi(\tilde{\mathbf{x}}) q(\tilde{\mathbf{x}}, \mathbf{x}^t)}{\pi(\mathbf{x}^t) q(\mathbf{x}^t, \tilde{\mathbf{x}})}\right)$

Step3: Generate

$\mathbf{u}^{t+1} \sim \mathcal{N}\left((\mathbf{I}_Q - \alpha \mathbf{H}^* \mathbf{H}) \mathbf{x}^{t+1}, \sigma^2 \alpha (\mathbf{I}_Q - \alpha \mathbf{H}^* \mathbf{H}) \right).$

Note that, when $\mathbf{H}$ is circulant, sampling from (4) can be efficiently handled in the Fourier domain. The new proposal (6) is generally simpler to sample from, since the two operators reflecting the correlation between the coefficients of the target signal induced from the likelihood and the prior are dissociated. Correlations from the likelihood are no longer related directly to the target signal but only through the auxiliary variable $\mathbf{u}$. In the particular case when the coefficient of the signal are supposed to be uncorrelated, one can sample the coefficients of the signal independently. Otherwise, we propose to use Langevin based MCMC algorithms. In particular, it is possible to construct an efficient curvature matrix that takes into account the prior correlation and that can be easily handled.

4. SIMULATION RESULTS

4.1. Problem formulation

We address the problem of recovering a multicomponent image $\mathbf{x} \in \mathbb{R}^Q$ with B components degraded by an operator $\mathbf{H} \in \mathbb{R}^{N \times Q}$ modeling spatially invariant blur and a zero-mean additive Gaussian noise with variance σ^2. To tackle this problem, we propose to make use of a sparse representation of the image. Let $\mathbf{F} \in \mathbb{R}^{Q \times Q}$ denotes a discrete wavelet transform (DWT) and $\mathbf{c} \in \mathbb{R}^Q$ the vector of wavelet coefficients i.e $\mathbf{c} = \mathbf{F}\mathbf{x}$. Note that the wavelets coefficients are grouped into M subbands of size Q_m corresponding to a specific orientation and scale. For all $m \in \{1, \ldots, M\}$, $q \in \{1, \ldots, Q_m\}$ let $\mathbf{P}_{m,q}$ be the permutation matrix of size $B \times Q$ that allows us to access to the vector of coefficients $\mathbf{c}_{m,q} \in \mathbb{R}^B$ in a given location q through all the B channel i.e $\mathbf{c}_{m,q} = \mathbf{P}_{m,q} \mathbf{c}$. Following [25, 29], we assume that for every $m \in \{1, \ldots, M\}$, $q \in \{1, \ldots, Q_m\}$, the vectors $(\mathbf{c}_{m,q})_{q=1}^{Q_m}$ are realizations of a random vector following a Generalized Multivariate Exponential Power (GMEP) distribution with scale matrix $\mathbf{\Sigma}_m$ and shape parameter β_m. Thus, the minus-log of the prior likelihood is given up to an additive constant by

$$-\log \mathsf{p}(\mathbf{x}) = \sum_{m=1}^{M} \sum_{q=1}^{Q_m} \psi_m(\|\mathbf{\Sigma}_m^{-1/2} \mathbf{P}_{mq} \mathbf{F}\mathbf{x} - \mathbf{a}_m\|) \quad (7)$$

where, for every $m \in \{1, \ldots, M\}$, $\mathbf{a}_m \in \mathbb{R}^B$ and for all $t \in \mathbb{R}$, $\psi_m(t) = \frac{1}{2}(t^2 + \delta_m)^{\beta_m}$ where $\delta_m > 0$ is an additional parameter added to ensure the differentiability of the function for all $\beta_m > 0$.

4.2. Results

The test image is a Landsat Thematic Mapper image of size 256×256 having $B = 6$ channels. Hence, the problem dimension is $Q = 393216$. We consider the resolution of the inverse problem defined in (1) where the original image is artificially degraded by a uniform blur of size 3×3 and an additive zero-mean white Gaussian noise with variance $\sigma^2 = 25$ so that the initial signal-to-noise ratio (SNR) is 16.92 dB. We adopt an orthonormal wavelet decomposition using the Beylkin wavelet, and three resolution levels, hence $M = 10$. For the subband corresponding to the approximation coefficients, we choose a Gaussian prior i.e $\beta_m = 2$. For the remaining subbands, we fix $\delta_m = 0.0001$ and $\beta_m = 0.5$, which yields a smoothed version of the Laplace distribution. We aim to compute the MMSE estimator of the original image $\mathbf{x}$ from the degraded version $\mathbf{z}$ using Langevin based sampling algorithms with auxiliary variables. Since, we have adopted an orthonormal basis, we can draw samples of the vector regrouping the wavelet coefficients of the B components, at a given location and scale, in an independent manner. Thus, the resolution of the initial high dimensional problem reduces to the resolution of 256×256 small subproblems of size B. To sample from the signal coefficients in each subproblem, we propose to use either standard MALA algorithm, or a preconditioned version (3MH) [25, 30] with the following curvature matrix resulting from a Majorize-Minimize strategy :

$$(\forall \mathbf{z} \in \mathbb{R}^B) \quad \mathbf{Q}_{m,q}(\mathbf{z}) = \mu \mathbf{I}_B + \mathbf{\Sigma}_m^{-1} \omega_{m,q}(\mathbf{z}), \quad (8)$$

where $\omega_{m,q}(\mathbf{z}) = \beta_m \left(\mathbf{z}^\top \mathbf{\Sigma}_m^{-1} \mathbf{z} + \delta_m \right)^{\beta_m - 1}$ and we choose the stepsize ε to achieve an acceptance probability of approximately 0.45. Figure 1 shows the degraded image as well as the MMSE estimator computed over 5000 samples after convergence which corresponds to a final SNR equals to 20.03 dB. Since the subproblems dimension is small, MALA and 3MH behave similarly in this plot. However, one can note that 3MH algorithm is slightly better in term of mixing properties as depicted in Figure 2. We also compare the speed of our proposed approach with standard MALA and 3MH algorithms without use of auxiliary variables. Figure 3 shows the

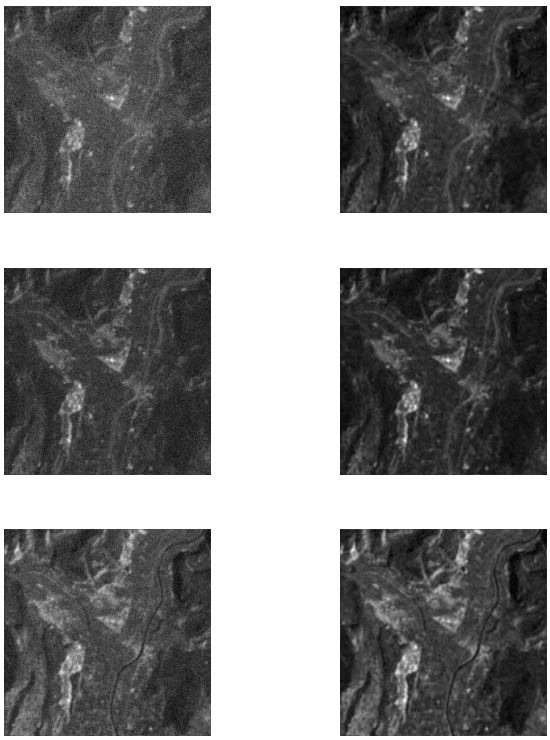

Fig. 1. From top to bottom: Components 2, 3 and 6 of the degraded image (left) and restored image (right). SNR= (14.74 dB, 20.86 dB) (13.52 dB, 17.23 dB) (12.99 dB, 15.90 dB).

evolution of SNR with respect to the computational time for the proposed algorithms (in solid lines) and standard algorithms without use of auxiliary variables (in dashed lines). It can be observed that the proposed algorithms reach stability much faster than the standard methods. Indeed, since the problem dimension is reduced, one can expect that the step-size ε takes larger values compared with standard algorithms so that the chain makes larger moves and explores the target space fast and efficiently. In addition, note that for larger dimensional problem, one could further improve the efficiency of the proposed algorithms by exploiting the parallel structure of the sampling tasks.

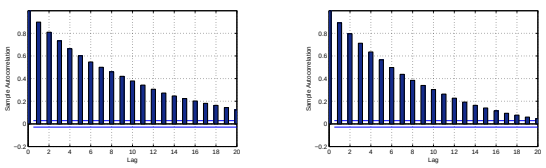

Fig. 2. Autocorrelation plot using mean($\mathbf{x}^t$) as scalar summary.

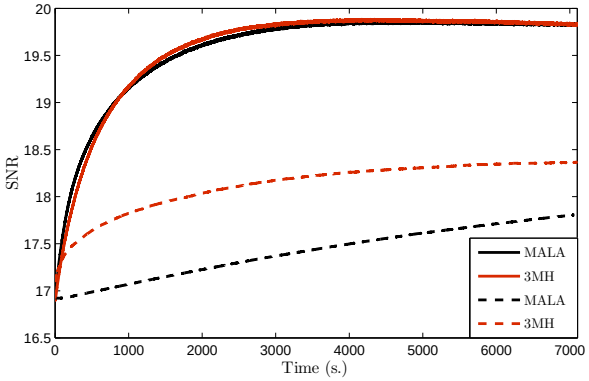

Fig. 3. Convergence speed of standard algorithms compared with the proposed sampling methods.

5. CONCLUSION

In this paper, we have proposed a method for improving the efficiency and the mixing properties of Langevin based MCMC sampling algorithms for high dimensional problems. By adding some auxiliary variables to the model, we succeeded in addressing separately the different sources of correlations in the target posterior density. Hence, the resulting model makes sampling much easier. We have applied the proposed algorithms to compute the MMSE estimator of a multicomponent image from its blurred version. The coefficients of the target image are no longer updated jointly but in parallel. Experimental results have shown the good performance of this new approach compared with state-of-the-art methods.

6. REFERENCES

[1] M. Pereyra, P. Schniter, E. Chouzenoux, J-C Pesquet, J-Y Tourneret, A. Hero, and S. McLaughlin, "A Survey of Stochastic Simulation and Optimization Methods in Signal Processing," *IEEE J. Sel. Topics Signal Process. (to appear)*, 2015, `http://arxiv.org/abs/1505.00273`.

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

[3] J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer Series in Statistics. Springer-Verlag, New-York, 2001.

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

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

[6] A. E. Raftery and S. M. Lewis, "How many iterations for the Gibbs sampler?," *Bayesian Statistics*, vol. 4, pp. 763–773, 1992.

[7] S. P. Brooks and A. Gelman, "General methods for monitoring convergence of iterative simulations," *J. Comput. Graph. Stat.*, vol. 7, no. 4, pp. 434–455, 1998.

[8] 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.

[9] 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.

[10] H. M. David, "Auxiliary variable methods for Markov Chain Monte Carlo with applications," *J. Am. Stat. Assoc.*, vol. 93, pp. 585–595, 1997.

[11] M. Hurn, "Difficulties in the use of auxiliary variables in Markov chain Monte Carlo methods," *Stat. Comput.*, vol. 7, pp. 35–44, 1997.

[12] P. Damlen, J. Wakefield, and S. Walker, "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 61, no. 2, pp. 331–344, 1999.

[13] A. Mira and L. Tierney, "On the use of auxiliary variables in Markov chain Monte Carlo sampling," Tech. Rep., 1997, `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7814`.

[14] S. Duane, A.D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B*, vol. 195, no. 2, pp. 216 – 222, 1987.

[15] R. M. Neal, "Slice sampling," *Ann. Stat.*, vol. 31, pp. 705–767, 2003.

[16] G. O. Roberts and J. S Rosenthal, "Convergence of slice sampler Markov chains," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 61, no. 3, pp. 643–660, 1999.

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

[18] 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.

[19] 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, pp. 255–268, 1997.

[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] 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.

[22] 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.

[23] Y. Zhang and C. A. Sutton, "Quasi-Newton methods for Markov chain Monte Carlo," in *Advances in Neural Information Processing Systems*, Granada, Spain, 12-17 Dec. 2011, number 24, pp. 2393–2401.

[24] 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.

[25] Y. Marnissi, A. Benazza-Benyahia, E. Chouzenoux, and J.-C. Pesquet, "Majorize-Minimize adapted Metropolis-Hastings algorithm. Application to multichannel image recovery," in *Proc. Eur. Sig. Image Proc. Conf. (EUSIPCO 2014)*, Lisbon, Portugal, 1-5 Sept. 2014, pp. 1332–1336.

[26] 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, Jul. 2014.

[27] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle, "A l1-unified variational framework for image restoration," in *8th European Conference on Computer Vision (ECCV 2004)*, Prague, Czech Republic, 11-14 May 2004, vol. 3024 of *Lecture Notes in Computer Science*, pp. 1–13, Springer.

[28] R. Cavicchioli, C. Chaux, L. Blanc-Féraud, and L. Zanni, "ML estimation of wavelet regularization hyperparameters in inverse problems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2013)*, Vancouver, Canada, 26-31 May 2013, pp. 1553–1557.

[29] 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.

[30] Y. Marnissi, E. Chouzenoux, A. Benazza-Benyahia, J.-C. Pesquet, and L. Duval, "Reconstruction de signaux parcimonieux à l'aide d'un algorithme rapide d'échantillonnage stochastique," in *Actes du 25e colloque GRETSI*, Lyon, France, 8-11 Sep. 2015.