An Auxiliary Variable Method for MCMC Algorithms in High Dimension

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

An Auxiliary Variable Method for MCMC Algorithms in High Dimension

Yosra Marnissi[1], Emilie Chouzenoux[2,3], Amel Benazza-Benyahia[4], and Jean-Christophe Pesquet[3] *

Abstract

When the parameter space is high dimensional, the performance of stochastic sampling algorithms is very sensitive to existing dependencies between parameters. For instance, this problem arises when one aims to sample from a high dimensional Gaussian distribution whose covariance matrix does not present a simple structure. Then, one often resorts to sampling algorithms based on a perturbation-optimization technique that requires to minimize a cost function using an iterative algorithm. This makes the sampling process time consuming, especially when used within a Gibbs sampler. Another challenge is the design of Metropolis-Hastings proposals that make use of information about the local geometry of the target density in order to speed up the convergence and improve mixing properties in the parameter space, while being not too computationally expensive. These two contexts are mainly related to the presence of two heterogeneous sources of dependencies stemming either from the prior or the likelihood in the sense that the related covariances matrices cannot be diagonalized in the same basis. In this paper, we are interested in inverse problems where either the data fidelity term or the prior distribution is Gaussian or driven from a hierarchical Gaussian model. We propose to add auxiliary variables to the model in order to dissociate the two sources of dependencies. In the new augmented space, only one source of correlation remains directly related to the target parameters, the other sources of correlations being captured by the auxiliary variables. Experiments conducted on two image restoration problems show the good performance of the proposed strategy.

1 Introduction

In a wide range of applicative areas, we do not have access to the signal of interest $\overline{\mathbf{x}} \in \mathbb{R}^Q$ but only to some observations $\mathbf{z} \in \mathbb{R}^N$ related to $\overline{\mathbf{x}}$ through the following model:

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

where $\mathbf{H} \in \mathbb{R}^{N\times Q}$ is the observation matrix that may express a blur or a projection and $\mathcal{D}$ is the noise model representing measurement errors. Our objective is to find an estimator $\hat{\mathbf{x}}$ of $\overline{\mathbf{x}}$ from the observations $\mathbf{z}$. Such inverse problem arises in several signal processing applications such as denoising, deblurring, and tomography reconstruction [1, 2].

The common Bayesian procedure for signal estimation consists in deriving estimators from the posterior distribution that captures all information inferred about the target signal from the

*1 - SAFRAN TECH, Groupe Safran, Rue des Jeunes Bois, Châteaufort, 78772 Magny Les Hameaux Cedex, France (email: yosra.marnissi@gmail.com). 2 - University Paris-East, LIGM, UMR CNRS 8049, 2 Bvd Blaise Pascal - Cité Descartes, 93162 Noisy-Le-Grand, France (e-mail: emilie.chouzenoux@u-pem.fr). 3 - Center for Visual Computing, INRIA, CentraleSupélec, University Paris-Saclay, 9 rue Joliot Curie, 91190 Gif sur Yvette, France (e-mail: jean-christophe@pesquet.eu). 4 - University of Carthage, SUP'COM, COSIM Lab., Route de Raoued Km 3.5 2083 El Ghazala Ariana, Tunisia (e-mail: benazza.amel@supcom.rnu.tn).

collected data. Given the observation model (1), the minus logarithm of the density posterior distribution reads:

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathscr{J}(\mathbf{x}) = -\log \mathsf{p}(\mathbf{x}|\mathbf{z}) = \Phi(\mathbf{H}\mathbf{x};\mathbf{z}) + \Psi(\mathbf{V}\mathbf{x}). \tag{2}$$

Hereabove, Φ is the neg-log likelihood that may take various forms depending on the noise statistical model $\mathcal{D}$. In particular, if $\mathcal{D}$ models an additive Gaussian noise with covariance $\mathbf{\Lambda}^{-1}$, it reduces (up to an additive constant) to the least squares function $\Phi(\mathbf{H}\mathbf{x};\mathbf{z}) = \frac{1}{2}\|\mathbf{H}\mathbf{x} - \mathbf{z}\|^2_{\mathbf{\Lambda}}$. Other common choices can be found for instance in [3, 4]. Moreover, $\Psi(\mathbf{V}\cdot)$ is related to some prior knowledge one can have about $\mathbf{x}$, and $\mathbf{V} \in \mathbb{R}^{M\times N}$ is a linear transform that can describe, for example, a frame analysis [5] or a discrete gradient operator [6]. Within a Bayesian framework, it is related to a prior distribution of density $\mathsf{p}(\mathbf{x})$ whose logarithm is given by $\log \mathsf{p}(\mathbf{x}) = -\Psi(\mathbf{V}\mathbf{x})$.

Monte Carlo inference approaches allow us to have a good description of the target space from a set of samples drawn from a distribution [7, 8, 9, 10, 11, 12]. In particular, these samples can be used to infer useful statistics such as the mean and the variance. In the context of Bayesian estimation, these techniques appear useful to compute, for example, the Minimum Mean Square Error (MMSE) estimator which is equivalent to the posterior mean. In this case, the MMSE estimator is approximated using the empirical average over the generated samples from the posterior distribution. When the exact expression of the posterior density is intractable, Markov Chain Monte Carlo (MCMC) have been widely used to approximate it [13]. These techniques are random variable generators that allow us to draw samples from complicated distributions. Perhaps, the most commonly used MCMC algorithm is the Metropolis-Hastings (MH) one which operates as follows [14]: from a given proposal distribution, we construct an irreducible Markov chain whose stationary distribution is the sought posterior law i.e., after a sufficient number of iterations, the samples generated by the algorithm follow the desired posterior distribution. At each iteration t, a decision rule is applied to accept or to reject the proposed sample given by the following acceptance probability

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

where $\tilde{\mathbf{x}}^{(t)}$ is the proposed sample at iteration t, generated from a proposal distribution with density $\mathsf{g}(.|\mathbf{x}^{(t)})$ that may depend on the current state $\mathbf{x}^{(t)}$. Note that, when more than one unknown variable has to be estimated (e.g., acquisition parameters or prior hyperparameters), one can draw iteratively samples from the conditional posterior distribution for each variable given the remaining ones using an MH iteration. This is known as the hybrid Gibbs sampler [15]. High dimensional models, often encountered in inverse problems, (e.g., in multispectral remote sensing applications [16]) constitute a challenging task for Bayesian inference problems. While many popular sampling algorithms have been widely used to fit complex multivariable models in small dimensional spaces [17, 18, 19, 20, 21, 22], they generally fail to explore the target distribution efficiently when applied to large scale problems, especially when the variables are highly correlated. This may be due to the poor mixing properties of the Markov chain or to the high computational cost of each iteration [17].

In this work, we propose a novel approach based on a data augmentation strategy [23] which aims at overcoming the limitations of standard Bayesian sampling algorithms when facing large scale problems. The remainder of this paper is organized as follows. In Section 2, we discuss the main difficulties encountered in standard sampling methods for large scale problems. We show how the addition of auxiliary variables to the model can improve their robustness with respect to these issues. The core of our contribution is detailed in Section 3. We first give a complete description of the proposed approach in the case of Gaussian noise and we study its extension to scale mixtures of Gaussian models. Furthermore, we demonstrate how the proposed approach can facilitate sampling from Gaussian distributions in Gibbs algorithms. Then, some computational issues, arising in the proposed Bayesian approach, are discussed. Sections 4 and 5 are devoted to the experimental validation of our method. In Section 4, we show the advantages of the proposed approach in dealing with high dimensional models involving highly correlated variables over a dataset of multispectral images affected by blur and additive Gaussian noise. In Section 5, we test the performance of our method in sampling from large scale Gaussian distributions through an application to image recovery under two-term mixed Gaussian noise. Finally, we give some conclusions and perspectives in Section 6.

2 Motivation

2.1 Sampling issues in high dimensional space

MCMC sampling methods may face two main difficulties when applied to large scale inverse problems. First, except for particular cases (e.g. circulant observation matrix), the structure of the observation model that links the unknown signal to the observations usually makes the estimation of the parameters of the posterior distribution quite involved. Second, even with simple models, the posterior distribution may still be difficult to sample from directly or to explore efficiently using standard sampling algorithms. As a specific case, this problem arises for Gaussian distributions, if the problem dimension is too high [24]. It can also arise in MH algorithms when sophisticated proposal rules are employed with the aim to cope with both the high dimensionality and the strong correlation existing between the target parameters [22]. In what follows, we will give more details about these two contexts.

2.1.1 Sampling from high dimensional Gaussian distribution:

Let us focus on the problem of sampling from a multivariate Gaussian distribution with a given precision matrix $\mathbf{G} \in \mathbb{R}^{Q\times Q}$. This problem emerges in many applications such as linear inverse problems involving Gaussian or hierarchical Gaussian models. More precisely, let us consider the following linear model:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w}, \tag{4}$$

where $\mathbf{w}$ is $\mathbb{R}^N$-valued and let us assume that conditionally to some latent variables, $\mathbf{w}$ and $\mathbf{x}$ are drawn from Gaussian distributions $\mathcal{N}(\mathbf{0}_N, \mathbf{\Lambda}^{-1})$ and $\mathcal{N}(\mathbf{m_x}, \mathbf{G_x}^{-1})$ respectively where $\mathbf{m_x} \in \mathbb{R}^Q$, $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ and $\mathbf{G_x} \in \mathbb{R}^{Q\times Q}$ are positive semi-definite matrices.[1] The parameters of these Gaussian distributions may be either fixed or unknown (i.e., involving some unknown hyperparameters such as regularization or acquisition parameters). It follows that the posterior distribution of $\mathbf{x}$ is Gaussian with mean $\mathbf{m} \in \mathbb{R}^Q$ and precision matrix $\mathbf{G} \in \mathbb{R}^{N\times N}$ defined as follows:

$$\mathbf{G} = \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H} + \mathbf{G_x} \tag{5}$$

$$\mathbf{m} = \mathbf{G}^{-1}\left(\mathbf{H}^\top \mathbf{\Lambda} \mathbf{z} + \mathbf{G_x}\mathbf{m_x}\right). \tag{6}$$

A common solution to sample from $\mathcal{N}(\mathbf{m}, \mathbf{G}^{-1})$ is to use the Cholesky factorization of the covariance or the precision matrix $\mathbf{G}$ [25]. However, when implemented through a Gibbs sampler, this method is of limited interest. First, the precision matrix $\mathbf{G}$ may depend on the unknown parameters of the model and may thus take different values along the algorithm. Thereby,

[1] In the following, when not mentioned, the Gaussian law can be degenerated that is, the precision matrix is semi-definite positive but not with full rank. In this case, $(\cdot)^{-1}$ denotes the generalized inverse.

spending such high computational time at each iteration of the Gibbs sampler to compute the Cholesky decomposition of the updated matrix may be detrimental to the convergence speed of the Gibbs sampler. Another concern is that, when dealing with high dimensional problems, we have generally to face not only computational complexity issues but also memory limitations. Such problems can be alleviated when the matrix presents some specific structures (e.g., circulant [26, 27] or sparse [28]). However, for more complicated structures, the problem remains critical especially when $\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}$ and $\mathbf{G}_\mathbf{x}$ cannot be diagonalized in the same basis. Other recently proposed algorithms for sampling Gaussian distributions in high dimension follow a two-step perturbation-optimization approach [29, 30, 31, 32, 33, 34], which can be summarized as follows

- Perturbation: Draw a Gaussian random vector $\mathbf{n}_1 \sim \mathcal{N}(\mathbf{0}_Q, \mathbf{G})$.
- Optimization: Solve the linear system $\mathbf{G}\mathbf{n}_2 = \mathbf{n}_1 + \mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{G}_\mathbf{x}\mathbf{m}_x$.

The solution to the above linear system can be approximated using iterative methods such as conjugate gradient algorithms, so leading to an approximate sample of the sought distribution [30, 31]. This issue has been considered in [32] by adding a Metropolis step in the sampling algorithm. In [33, 34], the authors propose to reduce the computational cost by sampling along mutually conjugate directions instead of the initial high dimensional space.

2.1.2 Designing efficient proposals in MH algorithms:

Non-Gaussian models arise in numerous applications in inverse problems [35, 36, 37, 38]. In this context, the posterior distribution is non-Gaussian and does not generally follow a standard probability model. In this respect, MH algorithms are good tools for exploring such posteriors and hence for drawing inferences about models and parameters. However, the challenge for MH algorithms is to construct a proposal density that provides a good approximation of the target density while being inexpensive to manipulate. Typically, in large scale problems, the proposal distribution takes the form of a random walk (RW), that is, in each iteration, the proposal density $\mathsf{g}(.|\mathbf{x}^{(t)})$ in (3) is a Gaussian law centered at the current state $\mathbf{x}^{(t)}$ and with covariance matrix $\varepsilon^2\mathbf{Q}(\mathbf{x}^{(t)})$. Moreover, ε is a positive constant whose value is adjusted so that the acceptance probability in (3) is bounded away from zero at convergence [17]. Other sampling algorithms incorporate information about the derivative of the logarithm of the target distribution to guide the Markov chain toward the target space where samples should be mostly concentrated. For instance, when the target density is differentiable, one can use Langevin-based algorithms where the mean of the Gaussian proposal density is replaced with one iteration of a preconditioned gradient descent algorithm as follows [39, 40, 41, 20, 22, 42]:

$$\tilde{\mathbf{x}}^{(t)} \sim \mathcal{N}\left(\mathbf{x}^{(t)} - \frac{\varepsilon^2}{2}\mathbf{Q}(\mathbf{x}^{(t)})^{-1}\nabla\mathcal{J}(\mathbf{x}^{(t)}), \varepsilon^2\mathbf{Q}(\mathbf{x}^{(t)})^{-1}\right). \tag{7}$$

In (7), $\nabla\mathcal{J}$ is the gradient of $\mathcal{J}$, ε is a positive constant and $\mathbf{Q}$ is a symmetric definite positive matrix that captures possible correlations between the coefficients of the signal. Note that some advanced versions of Langevin based algorithms have been proposed to address problems with non smooth laws [43, 44]. It is worth noting that the choice of the scale matrices $\left(\mathbf{Q}(\mathbf{x}^{(t)})\right)_t$ may deeply affect the efficiency of the aforementioned algorithms [22]. In fact, an inappropriate choice of $\mathbf{Q}$ may alter the quality of the Markov chain leading to very correlated samples and thereby biased estimates. Moreover, computationally cheap matrices are also preferable especially in high dimensional spaces. In the case of low dimensional problems and when the coefficients of the signal are not highly correlated, the standard RW and Metropolis adapted Langevin algorithm (MALA) obtained for $\mathbf{Q} \equiv \mathbf{I}_Q$ achieve overall good results. For instance, in the context of denoising problems with uncorrelated Gaussian noise, when the coefficients of the signal are assumed to be statistically independent in the prior, they can be either sampled independently using RW or jointly by resorting to MALA. However, these algorithms may be inaccurate for large scale problems especially when the coefficients of the signal exhibit high correlations [22]. In this case, the design of a good proposal often requires considering the curvature of the target distribution. More sophisticated (and thus more computationally expensive) scale matrices should be chosen to drive the chain in the directions that reflect the dependence structure. Optimally, the curvature matrix should be chosen such that it adequately captures two kinds of dependencies: correlation over the observations specified by the observation model and, correlation between different coefficients of the target signal specified by the prior law. For instance, $\mathbf{Q}$ can be set to the Hessian matrix of the minus logarithm of the posterior density in the current state [20, 21], or to the Fisher matrix especially when the Hessian matrix is not definite positive [22, 42] or to the empirical covariance matrix computed according to the previous states of the Markov chain [45]. When the minus-log of the target density can be expressed as in (2), good candidates of the curvature matrix take the following form:

$$\mathbf{Q} = \mathbf{H}^\top \boldsymbol{\Lambda} \mathbf{H} + \mathbf{V}^\top \boldsymbol{\Omega} \mathbf{V} \tag{8}$$

where $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ are semi-definite positive matrices. Feasible numerical factorization of $\mathbf{Q}$ can be ensured if $\mathbf{H}^\top \boldsymbol{\Lambda} \mathbf{H}$ and $\mathbf{V}^\top \boldsymbol{\Omega} \mathbf{V}$ are diagonalizable in the same basis. Otherwise, the use of the full matrix (8) in the scheme (7) remains generally of limited interest especially for large scale problems where the manipulation of the resulting proposal generally induces a high computational complexity altering the convergence speed. Alternatively, under mild conditions on the posterior density, the Majorize-Minimize strategy offers a high flexibility for building curvatures matrices with a lower computational cost (e.g., diagonal matrices, bloc-diagonal matrices, circulant...) [41]. It should however be pointed out that MH algorithms with too simple preconditioning matrices resulting from rough approximations of the posterior density may fail to explore the target space efficiently. Therefore, the scale matrix $\mathbf{Q}$ should be adjusted to achieve a good tradeoff between the computational complexity induced in the algorithm and the accuracy/closeness of the proposal to the the true distribution.

2.2 Auxiliary variables and data augmentation strategies

It can be noted that the main difficulty arising in the aforementioned sampling problems is related to the presence of heterogeneous types of dependencies between the coefficients of the signal. These dependencies may come either from the likelihood or from the prior information. In fact, the operator $\mathbf{H}$ in the likelihood may cause high dependencies between coefficients in a wide neighborhood even if the coefficients of the signal are assumed to be statistically dependent in the prior law. The problem can be treated in another domain where $\mathbf{H}$ can be easily diagonalized i.e., the coefficients of the signal become uncorrelated in the likelihood. However, when we take into account the prior dependencies, this strategy becomes inefficient especially when the prior covariance matrix cannot be diagonalized in the same basis as $\mathbf{H}$, which is the case of most real problems. One should therefore process these two sources of correlations separately.

To improve the mixing of sampling algorithms, many works have proposed to eliminate one of these sources of correlation directly related to $\mathbf{x}$ by adding some auxiliary variables to the initial model, associated with a given conditional distribution such that simulation can be performed in a simpler way in the new larger space. Instead of simulating directly from the initial distribution, a Markov chain is constructed by alternately drawing samples from the conditional distribution of each variable, which reduces to a Gibbs sampler in the new space. This technique has been used in two different statistical literatures: data augmentation

[46] and, auxiliary variables strategies [47]. It is worthwhile noting that the two methods are equivalent in their general formulation and the main difference is often related to the statistical interpretation of the auxiliary variable (unobserved data or latent variable) [23]. In the following, we will use the term Data Augmentation (DA) to refer to any method that constructs sampling algorithms by introducing auxiliary variables. Some DA algorithms have been proposed in [48, 49, 50, 51, 52, 53, 54]. A specific attention has been turned towards the Hamiltonian MCMC (HMC) approach [55, 22], that defines auxiliary variables based on physically inspired dynamics.

In the following, we propose to alleviate the problem of heterogeneous dependencies by resorting to a DA strategy. More specifically, we propose to add some auxiliary variables $\mathbf{u} \in \mathbb{R}^J$ with predefined conditional distribution of density $\mathsf{p}(\mathbf{u}|\mathbf{x},\mathbf{z}) = \mathsf{p}(\mathbf{u}|\mathbf{x})$ so that the minus logarithm of the joint distribution density $\mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})$ can be written as follows:

$$\mathcal{J}(\mathbf{x},\mathbf{u}) = \mathcal{J}(\mathbf{u}|\mathbf{x}) + \mathcal{J}(\mathbf{x}) \tag{9}$$

where $\mathcal{J}(\mathbf{u}|\mathbf{x}) = -\log \mathsf{p}(\mathbf{u}|\mathbf{x})$ up to an additive constant. Two conditions should be satisfied by $\mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})$ for the DA strategy to be valid:

(C_1) $\int_{\mathbb{R}^J} \mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})\, \mathrm{d}\mathbf{u} = \mathsf{p}(\mathbf{x}|\mathbf{z})$;

(C_2) $\int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})\, \mathrm{d}\mathbf{x} = \mathsf{p}(\mathbf{u}|\mathbf{z})$,

where $\mathsf{p}(\mathbf{u}|\mathbf{z})$ should define a valid probability density function (i.e., nonnegative and with integral with respect to $\mathbf{u}$ equal to 1). In fact, the importance of Condition (C_1) is obvious because the latent variable is only introduced for computational purposes and should not alter the considered initial model. The need for the second requirement (C_2) stems from the fact that $\mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})$ should define the density of a proper distribution. Note that

- the first condition is satisfied thanks to the definition of the joint distribution in (9) provided that $\mathsf{p}(\mathbf{u}|\mathbf{x},\mathbf{z})$ is a density of a proper distribution;
- for the second condition, it can be noticed that if the first condition is met, Fubini-Tonelli's theorem allows us to claim that

$$\int_{\mathbb{R}^J} \left(\int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})\, \mathrm{d}\mathbf{x} \right) \mathrm{d}\mathbf{u} = \int_{\mathbb{R}^Q} \left(\int_{\mathbb{R}^J} \mathsf{p}(\mathbf{x},\mathbf{u}|\mathbf{z})\, \mathrm{d}\mathbf{u} \right) \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{x}|\mathbf{z})\, \mathrm{d}\mathbf{x} = 1. \tag{10}$$

  This shows that $\mathsf{p}(\mathbf{u}|\mathbf{z})$ as defined in (C_2) is a valid probability density function.

Instead of simulating directly from $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$, we now draw alternatively (in an arbitrary order) samples from the conditional distributions of the two variables $\mathbf{x}$ and $\mathbf{u}$ of respective densities $\mathcal{P}_{\mathbf{x}|\mathbf{u},\mathbf{z}}$ and $\mathcal{P}_{\mathbf{u}|\mathbf{x},\mathbf{z}}$. This simply reduces to a special case of a hybrid Gibbs sampler algorithm with two variables where each iteration t is composed of two sampling steps which can be expressed as follows:

- Sample $\mathbf{u}^{(t+1)}$ from $\mathcal{P}_{\mathbf{u}|\mathbf{x}^{(t)},\mathbf{z}}$;
- Sample $\mathbf{x}^{(t+1)}$ from $\mathcal{P}_{\mathbf{x}|\mathbf{u}^{(t+1)},\mathbf{z}}$.

Under mild technical assumptions [56, 9], the constructed chain $\left(\mathbf{x}^{(t)},\mathbf{u}^{(t)}\right)_{t\geqslant 0}$ can be proved to have a stationary distribution $\mathcal{P}_{\mathbf{x},\mathbf{u}|\mathbf{z}}$. The usefulness of the DA strategy is mainly related to the fact that, with an appropriate choice of $\mathsf{p}(\mathbf{u}|\mathbf{x},\mathbf{z})$, drawing samples from the new conditional distributions $\mathcal{P}_{\mathbf{x}|\mathbf{u},\mathbf{z}}$ and $\mathcal{P}_{\mathbf{u}|\mathbf{x},\mathbf{z}}$ is much easier than sampling directly from the initial distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$. Let us emphasize that, for the sake of efficiency, the manipulation of $\mathsf{p}(\mathbf{u}|\mathbf{x},\mathbf{z})$ must not induce a high computation cost in the algorithm. In this work, we propose to add auxiliary variables $\mathbf{u}$ to the model such that the dependencies resulting from the likelihood and the prior are separated, that is, $\mathcal{J}(\mathbf{u}|\mathbf{x})$ is chosen in such a way that only one source of correlations remains related directly to $\mathbf{x}$ in $\mathsf{p}(\mathbf{x}, \mathbf{u}|\mathbf{z})$, the other sources of correlations only intervening through the auxiliary variables $\mathbf{u}$ and $\mathbf{z}$. Note that the advantage of introducing auxiliary variables in optimization or sampling algorithms has also been illustrated in several works in the image processing literature, related to half quadratic approaches [57, 26, 58, 59, 60, 61]. This technique has also been adopted to facilitate sampling in classical MH algorithm and Gibbs sampler in the maximum likelihood estimation approach proposed in [62]. Finally, in [63], a half-quadratic formulation was used to replace the prior distribution, leading to a new posterior distribution from which inference results are deduced.

The contribution of our work is to propose an extended formulation of the data augmentation method, that was introduced in [61] in the context of variational image restoration under uncorrelated Gaussian noise. Our proposal leads to a novel acceleration strategy for sampling algorithms in large scale problems.

3 Proposed approach

In this section, we discuss various scenarios typically arising in inverse problems and we explain how our approach applies in these contexts.

3.1 Correlated Gaussian noise

Let us consider the linear observation model (4) when the noise term $\mathbf{w}$ is assumed to be Gaussian, additive and independent from the signal that is $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{\Lambda}^{-1})$ with $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ a symmetric semi-definite positive precision matrix that is assumed to be known. In this context, the minus logarithm of the posterior density takes the following form:

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}) = \frac{1}{2}(\mathbf{H}\mathbf{x} - \mathbf{z})^\top \mathbf{\Lambda}(\mathbf{H}\mathbf{x} - \mathbf{z}) + \Psi(\mathbf{V}\mathbf{x}). \tag{11}$$

Simulating directly from this distribution is generally not possible and standard MCMC methods may fail to explore it efficiently due to the dependencies between signal coefficients [22]. In particular, the coupling induced by the matrix $\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}$ may hinder the construction of suitable proposals when using MH algorithms. For example, when $\mathbf{V} = \mathbf{I}_Q$ and $\Psi(\mathbf{x}) = \sum_{i=1}^{Q} \psi_i(x_i)$, RW and standard MALA algorithms may behave poorly as they do not take into account data fidelity dependencies, while a preconditioned MALA approach with full curvature matrices may exhibit high computational load due the presence of heterogeneous dependencies [40].

In the following, we propose to eliminate the coupling induced by the linear operators $(\mathbf{H}, \mathbf{\Lambda})$ by adding auxiliary variables. Since the data fidelity term is Gaussian, a natural choice is to define $\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z})$ as a Gaussian distribution with mean $\mathbf{A}\mathbf{x}$ and covariance matrix $\mathbf{C}$:

$$\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z}) = \frac{\det(\mathbf{C})^{-1/2}}{(2\pi)^{J/2}} \exp\left(-\frac{1}{2}\|\mathbf{C}^{-1/2}(\mathbf{u} - \mathbf{A}\mathbf{x})\|^2\right) \tag{12}$$

where $\mathbf{C} \in \mathbb{R}^{J\times J}$ is a symmetric positive definite covariance matrix and $\mathbf{A} \in \mathbb{R}^{J\times Q}$. Then, the joint distribution satisfies the two conditions (C_1) and (C_2) defined in Section 2 and its minus logarithm has the following expression:

$$(\forall \mathbf{x} \in \mathbb{R}^Q)(\forall \mathbf{u} \in \mathbb{R}^J) \quad \mathcal{J}(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\left(\mathbf{x}^\top\mathbf{Y}\mathbf{x} + \mathbf{z}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{u}^\top\mathbf{C}^{-1}\mathbf{u} - 2\mathbf{x}^\top\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}\right)\right) + \Psi(\mathbf{V}\mathbf{x}), \tag{13}$$

with

$$\mathbf{Y} = \mathbf{H}^\top \boldsymbol{\Lambda} \mathbf{H} + \mathbf{A}^\top \mathbf{C}^{-1} \mathbf{A}. \tag{14}$$

The expression in (12) yields the sampling scheme:

$$(\forall t \in \mathbb{N}) \quad \mathbf{u}^{(t+1)} = \mathbf{A}\mathbf{x}^{(t)} + \mathbf{C}^{1/2}\mathbf{n}^{(t)}, \tag{15}$$

with $\mathbf{n}^{(t)} \sim \mathcal{N}(\mathbf{0}_J, \mathbf{I}_J)$. The efficiency of the DA strategy is thus highly related to the choice of the matrices $\mathbf{A}$ and $\mathbf{C}$. Under the requirement that $\mathbf{C}$ is positive definite, the choice of $(\mathbf{A}, \mathbf{C})$ is subjective and is related to specifying the source of heterogeneous dependencies that one wants to eliminate in the target distribution based on the properties of $\mathbf{H}$, $\boldsymbol{\Lambda}$, $\mathbf{V}$, and Ψ. More specifically, one should identify if the main difficulty stems from the structure of matrix $\mathbf{H}^\top \boldsymbol{\Lambda} \mathbf{H}$ or only from the non trivial form of the precision matrix $\boldsymbol{\Lambda}$. In what follows, we will elaborate different solutions according to the type of encountered difficulty.

<u>Alternative I:</u> Eliminate the coupling induced by $\boldsymbol{\Lambda}$

Let us first consider the problem of eliminating the coupling induced by matrix $\boldsymbol{\Lambda}$. This problem is encountered for example for Model (5) with circulant matrices $\mathbf{H}$ and $\mathbf{G}_\mathbf{x}$ and with $\boldsymbol{\Lambda} \neq \mathbf{I}_N$, which induces further correlation when passing to the Fourier domain. In this context, we propose to eliminate the correlations induced by $\boldsymbol{\Lambda}$ by setting

$$\mathbf{Y} = \frac{1}{\mu} \mathbf{H}^\top \mathbf{H} \tag{16}$$

where $\mu > 0$ is such that $\mu \|\boldsymbol{\Lambda}\|_\mathrm{S} < 1$, where $\|\cdot\|_\mathrm{S}$ denotes the spectral norm. This is equivalent to choosing $\mathbf{A}$ and $\mathbf{C}$ such that

$$\mathbf{A}^\top \mathbf{C}^{-1} \mathbf{A} = \mathbf{H}^\top \left(\frac{1}{\mu} \mathbf{I}_N - \boldsymbol{\Lambda} \right) \mathbf{H}. \tag{17}$$

Note that the condition over μ allows to guarantee that $\mathbf{C}$ is positive definite. Under (16), the minus logarithm of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ and $\mathbf{u}$ reads, up to an additive constant:

$$(\forall \mathbf{x} \in \mathbb{R}^Q)(\forall \mathbf{u} \in \mathbb{R}^J) \quad \mathcal{J}(\mathbf{x}|\mathbf{u}) = \frac{1}{2\mu} \|\mathbf{H}\mathbf{x}\|^2 - \mathbf{x}^\top \left(\mathbf{H}^\top \boldsymbol{\Lambda} \mathbf{z} + \mathbf{A}^\top \mathbf{C}^{-1} \mathbf{u} \right) + \Psi(\mathbf{V}\mathbf{x}). \tag{18}$$

Let us discuss the application of the hybrid Gibbs sampling algorithm from Section 2 to this particular decomposition. The sampling scheme (15) yields:

$$(\forall t \in \mathbb{N}) \quad \mathbf{A}^\top \mathbf{C}^{-1} \mathbf{u}^{(t+1)} = \mathbf{A}^\top \mathbf{C}^{-1} \mathbf{A} \mathbf{x}^{(t)} + \mathbf{A}^\top \mathbf{C}^{-1/2} \mathbf{n}^{(t)} \tag{19}$$

where $\mathbf{n}^{(t)} \sim \mathcal{N}(\mathbf{0}_J, \mathbf{I}_J)$. Since $\mathbf{A}$ and $\mathbf{C}$ satisfy (17), this leads to:

$$(\forall t \in \mathbb{N}) \quad \mathbf{A}^\top \mathbf{C}^{-1} \mathbf{u}^{(t+1)} = \mathbf{H}^\top \left(\frac{1}{\mu} \mathbf{I}_N - \boldsymbol{\Lambda} \right) \mathbf{H} \mathbf{x}^{(t)} + \mathbf{A}^\top \mathbf{C}^{-1/2} \mathbf{n}^{(t)}. \tag{20}$$

We can remark that, for every $t \in \mathbb{N}$, $\mathbf{A}^\top \mathbf{C}^{-1/2} \mathbf{n}^{(t)}$ follows the centered Gaussian distribution with covariance matrix $\mathbf{H}^\top \left(\frac{1}{\mu} \mathbf{I}_N - \boldsymbol{\Lambda} \right) \mathbf{H}$. It follows that

$$(\forall t \in \mathbb{N}^*) \quad \mathbf{A}^\top \mathbf{C}^{-1} \mathbf{u}^{(t)} = \mathbf{H}^\top \mathbf{v}^{(t)} \tag{21}$$

where

$$(\forall t \in \mathbb{N}) \quad \mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma} \mathbf{H} \mathbf{x}^{(t)}, \boldsymbol{\Gamma} \right), \tag{22}$$

and $\boldsymbol{\Gamma} = \frac{1}{\mu}\mathbf{I}_N - \boldsymbol{\Lambda}$ is definite positive by construction. Then, the resulting algorithm can be viewed as a hybrid Gibbs sampler, associated to the minus logarithm of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ and a new auxiliary variable $\mathbf{v} \sim \mathcal{N}(\boldsymbol{\Gamma}\mathbf{H}\mathbf{x}, \boldsymbol{\Gamma})$:

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}|\mathbf{v}) = \frac{1}{2\mu}\|\mathbf{H}\mathbf{x} - \mu(\boldsymbol{\Lambda}\mathbf{z} + \mathbf{v})\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{23}$$

The main steps of the proposed Gibbs sampling algorithm are given in Algorithm 1. The appealing advantage of this algorithm with respect to a Gibbs sampler which would be applied directly to Model (5) when $\mathbf{H}$ and $\mathbf{G}_{\mathbf{x}}$ are diagonalizable in the same domain, is that it allows to easily handle the case when $\boldsymbol{\Lambda}$ is not equal to a diagonal matrix having identical diagonal elements.

Algorithm 1 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\boldsymbol{\Lambda}$.

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^N$, $\mu > 0$ such that $\mu\|\boldsymbol{\Lambda}\|_{\mathrm{S}} < 1$

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

2: Generate $\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}\mathbf{H}\mathbf{x}^{(t)}, \boldsymbol{\Gamma}\right)$ where
$\boldsymbol{\Gamma} = \frac{1}{\mu}\mathbf{I}_N - \boldsymbol{\Lambda}$

3: Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\mathbf{z}}$

4: **end for**

Note that, minimizing (23) can be seen as a restoration problem with an uncorrelated noise of variance μ. It can be expected that Step 3 in Algorithm 1 can be more easily implemented in the transform domain where $\mathbf{H}$ and $\mathbf{V}$ are diagonalized, when this is possible (see Section 5 for an example)

<u>Alternative II</u>: Eliminate the coupling induced by $\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}$

In a large class of regularized models, $\mathbf{H}$ and $\mathbf{V}$ have different properties. While $\mathbf{H}$ almost reflects a blur, a projection, or a decimation matrix, $\mathbf{V}$ may model a wavelet transform or a discrete gradient operator. Such difference in their properties induces a complicated structure of the posterior covariance matrix. To address such cases, we propose to eliminate the source of correlations related to $\mathbf{x}$ through $\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H} + \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}$, by setting $\mathbf{Y} = \frac{1}{\mu}\mathbf{I}_Q$, so that $\mathbf{A}$ and $\mathbf{C}$ satisfy

$$\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A} = \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}, \tag{24}$$

where $\mu > 0$ is such that $\mu\|\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}\|_{\mathrm{S}} < 1$, so that $\mathbf{C}$ is positive definite. It follows that the minus logarithm of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ and $\mathbf{u}$ is defined up to an additive constant as

$$(\forall \mathbf{x} \in \mathbb{R}^Q)(\forall \mathbf{u} \in \mathbb{R}^J) \quad \mathcal{J}(\mathbf{x}|\mathbf{u}) = \frac{1}{2\mu}\|\mathbf{x}\|^2 - \mathbf{x}^\top\left(\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{z} + \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}\right) + \Psi(\mathbf{V}\mathbf{x}). \tag{25}$$

Let us make the following change of variables within the Gibbs sampling method:

$$(\forall t \in \mathbb{N}^*) \quad \mathbf{v}^{(t)} = \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}^{(t)}.$$

According to (15) and (24), we obtain

$$(\forall t \in \mathbb{N}) \quad \mathbf{v}^{(t+1)} = \left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}\right) \mathbf{x}^{(t)} + \mathbf{A}^\top \mathbf{C}^{-1/2} \mathbf{n}^{(t)} \tag{26}$$

where $\mathbf{n}^{(t)} \sim \mathcal{N}(\mathbf{0}_J, \mathbf{I}_J)$. Let us define $\mathbf{\Gamma} = \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}$, which is positive definite. Since $\mathbf{A}^\top \mathbf{C}^{-1/2} \mathbf{n}^{(t)}$ follows a zero-mean Gaussian distribution with covariance matrix $\mathbf{\Gamma}$, then

$$(\forall t \in \mathbb{N}) \quad \mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\mathbf{\Gamma}\mathbf{x}^{(t)}, \mathbf{\Gamma}\right) \tag{27}$$

and the new target conditional distribution reads

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}|\mathbf{v}) = \frac{1}{2\mu}\|\mathbf{x} - \mu(\mathbf{v} + \mathbf{H}^\top \mathbf{\Lambda} \mathbf{z})\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{28}$$

The proposed Gibbs sampling algorithm is then summarized by Algorithm 2.

Algorithm 2 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}$.

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^Q$, $\mu > 0$ such that $\mu\|\mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}\| < 1$

1: **for** $t = 0, 1, \ldots$ **do**
2: Generate $\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\mathbf{\Gamma}\mathbf{x}^{(t)}, \mathbf{\Gamma}\right)$ where
$\mathbf{\Gamma} = \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}$
3: Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\mathbf{z}}$
4: **end for**

Note that in (28), the two operators reflecting the correlation between the coefficients of the target signal induced from the likelihood and the prior are now dissociated. Correlations from the likelihood are no longer related directly to the target signal but to the auxiliary variable $\mathbf{v}$ and the observation $\mathbf{z}$. In other words, the original problem reduces to solving a denoising problem where the variance of the Gaussian noise is μ. Thereby, the new target distribution (28) is generally simpler to sample from compared to the initial one. In particular, one can sample them independently when the coefficients of the signal are independent in the prior. Otherwise, if Ψ is a smooth function, one can use a Langevin-based MCMC algorithm. For instance, it may be possible to construct an efficient curvature matrix that takes into account the prior correlation and that can be easily manipulated.

Table 1 summarizes the two different cases we have presented here. We would like to emphasize that the approach we propose for adding auxiliary variables according to the structure of the matrix $\mathbf{H}$ and $\mathbf{\Lambda}$ is sufficiently generic so that it covers a wide diversity of applications.

It is worth noting that the auxiliary variable could be introduced in the data fidelity term as well as in the prior information. The derivation of the proposed method in (13) allows us to identify classes of models for which our approach can be extended. Obviously, the key

Table 1: Different alternatives for adding auxiliary variables

Problem source	Proposed auxiliary variable	Resulting conditional density $\mathsf{p}(\mathbf{x}\vert\mathbf{z},\mathbf{v}) \propto \exp(-\mathcal{J}(\mathbf{x}\vert\mathbf{v}))$
$\boldsymbol{\Lambda}$	$\mathbf{v} \sim \mathcal{N}\left(\left(\frac{1}{\mu}\mathbf{I}_N - \boldsymbol{\Lambda}\right)\mathbf{H}\mathbf{x}, \frac{1}{\mu}\mathbf{I}_N - \boldsymbol{\Lambda}\right)$	$\mathcal{J}(\mathbf{x}\vert\mathbf{v}) = \frac{1}{2\mu}\\|\mathbf{H}\mathbf{x} - \mu\left(\boldsymbol{\Lambda}\mathbf{z} + \mathbf{v}\right)\\|^2 + \Psi(\mathbf{V}\mathbf{x})$
$\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}$	$\mathbf{v} \sim \mathcal{N}\left(\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}\right)\mathbf{x}, \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}\right)$	$\mathcal{J}(\mathbf{x}\vert\mathbf{v}) = \frac{1}{2\mu}\\|\mathbf{x} - \mu(\mathbf{v} + \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{z})\\|^2 + \Psi(\mathbf{V}\mathbf{x})$

requirement is that the term which should be simplified can be written as a quadratic function with respect to some variables. Hence, without completely relaxing the Gaussian requirement, we can extend the proposed method to Gaussian models in which some hidden variables control the mean and/or the variance. This includes for example scale mixture of Gaussian models [64, 65] such as the alpha-stable family (including the Cauchy distribution), the Bernoulli Gaussian model and the generalized Gaussian distributions, and also Gaussian Markov random fields [56]. In Section 3.2, we will investigate the case of scale mixture of Gaussian models. When both the likelihood and the prior distribution are Gaussian conditionally to some parameters, the proposed method can be applied to each term as explained in Section 3.3.

Another point to pay attention to is the sampling of the auxiliary variable $\mathbf{v}$. In particular, in Algorithm 2, we should be able to sample from the Gaussian distribution whose covariance matrix is of the form $\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}\right)$, which is possible for a large class of observation models as discussed in Section 3.4.

3.2 Scale mixture of Gaussian noise

3.2.1 Problem formulation

Let us consider the following observation model:

$$(\forall i \in \{1,\dots,N\}) \quad z_i = [\mathbf{H}\mathbf{x}]_i + w_i \tag{29}$$

such that, for every $i \in \{1,\dots,N\}$,

$$\begin{cases} w_i = 0 & \text{if } \sigma_i = 0 \\ w_i \sim \mathcal{N}(0,\sigma_i^2) & \text{if } \sigma_i > 0 \end{cases} \tag{30}$$

where $(\sigma_1,\dots,\sigma_N)$ are independent random variables distributed on $\mathbb{R}^+$ according to $\mathcal{P}_\sigma$. Different forms of the mixing distribution $\mathcal{P}_\sigma$ lead to different noise statistics. In particular, the Cauchy noise is obtained when $\sigma_1^2,\dots,\sigma_N^2$ are random variable following an inverse Gamma distribution. Let $\boldsymbol{\sigma} = [\sigma_1,\dots,\sigma_N]^\top$. By assuming that $\mathbf{x}$ and $\boldsymbol{\sigma}$ are independent, the joint posterior distribution of $\mathbf{x}$ and $\boldsymbol{\sigma}$ is given by:

$$\mathsf{p}(\mathbf{x},\boldsymbol{\sigma} \mid \mathbf{z}) = \mathsf{p}(\mathbf{x} \mid \mathbf{z})\mathsf{p}(\boldsymbol{\sigma} \mid \mathbf{z}). \tag{31}$$

In such a Bayesian estimation context, a Gibbs sampling algorithm is generally adopted to sample alternatively from the distributions $\mathcal{P}_{\mathbf{x}|\boldsymbol{\sigma},\mathbf{z}}$ and $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x},\mathbf{z}}$.

In the following, we assume that the set $\mathcal{S}_0 = \{\sigma_1 = \sigma_2 = \ldots = \sigma_N = 0\}$ has a zero probability given the vector of observations $\mathbf{z}$. Note that by imposing such rule, we ensure that, at each iteration t of the Gibbs algorithm, $\boldsymbol{\sigma}^{(t)} \neq \mathbf{0}_N$ almost surely.

Since sampling from $\mathcal{P}_{\mathbf{x}|\boldsymbol{\sigma},\mathbf{z}}$ is supposed to be intractable, we propose to add auxiliary variables $\mathbf{v} \in \mathbb{R}^J$ that may depend on the variables of interest $\mathbf{x}$ and $\boldsymbol{\sigma}$ according to a given conditional distribution density $\mathsf{p}(\mathbf{v}|\mathbf{x},\boldsymbol{\sigma},\mathbf{z}) = \mathsf{p}(\mathbf{v}|\mathbf{x},\boldsymbol{\sigma})$ which satisfies the following conditions:

1- $\int_{\mathbb{R}^J} \mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z})\mathrm{d}\mathbf{v} = \mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}|\mathbf{z})$.

2- $\int_{\mathbb{R}^Q} \int_{\mathbb{R}^N} \mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z})\ \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\sigma} = \mathsf{p}(\mathbf{v}|\mathbf{z})$,

where $\mathsf{p}(\mathbf{v}|\mathbf{z})$ should be a valid probability density function.

Using the same arguments as in Section 2.2, these two properties are satisfied provided that $\mathsf{p}(\mathbf{v}|\mathbf{x}, \boldsymbol{\sigma}, \mathbf{z})$ defines a proper probability density function. It follows that the initial two step-Gibbs iteration is replaced by the following three sampling steps. First, sample $\mathbf{v}^{(t+1)}$ from $\mathcal{P}_{\mathbf{v}|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},\mathbf{z}}$ then sample $\mathbf{x}^{(t+1)}$ from $\mathcal{P}_{\mathbf{x}|\boldsymbol{\sigma}^{(t)},\mathbf{v}^{(t+1)},\mathbf{z}}$ and finally sample $\boldsymbol{\sigma}^{(t+1)}$ from $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)},\mathbf{z}}$.

3.2.2 Proposed algorithms

Let $\mathbf{D}(\boldsymbol{\sigma})$ be the diagonal matrix whose diagonal elements are given by

$$(\forall i \in \{1, \dots, N\}) \quad D(\boldsymbol{\sigma})_{i,i} = \begin{cases} 0 & \text{if } \sigma_i = 0 \\ (\sigma_i)^{-2} & \text{if } \sigma_i > 0. \end{cases} \tag{32}$$

Note that, since $\mathcal{S}_0$ has zero probability, we have almost surely

$$\|\mathbf{D}(\boldsymbol{\sigma})\|_{\mathrm{S}} > 0. \tag{33}$$

- Suppose first that there exists a constant $\nu > 0$ such that

  $$(\forall t \geqslant 0)\,(\forall i \in \{1, \dots, N\}) \quad \nu \leqslant \sigma_i^{(t)}. \tag{34}$$

  Then, results in Section 3.1 with a Gaussian noise can be extended to scale mixture of Gaussian noise by substituting, at each iteration t, $\mathbf{D}^{(t)}$ for $\boldsymbol{\Lambda}$, and by choosing $\mu < \nu^2$ in Algorithm 1 and $\mu\|\mathbf{H}\|_{\mathrm{S}}^2 < \nu^2$ in Algorithm 2. The only difference is that an additional step must be added to the Gibbs algorithm to draw samples of the mixing variables $\sigma_1, \dots, \sigma_N$ from their conditional distributions given $\mathbf{x}$, $\mathbf{v}$ and $\mathbf{z}$.

- Otherwise, when $\nu > 0$ satisfying (34) does not exist, results in Section 3.1 remain also valid when, at each iteration t, for a given value of $\boldsymbol{\sigma}^{(t)}$, we replace $\boldsymbol{\Lambda}$ by $\mathbf{D}(\boldsymbol{\sigma}^{(t)})$. There is however a main difference with respect to the case when $\nu > 0$ which is that μ depends on the value of the mixing variable $\boldsymbol{\sigma}^{(t)}$ and hence can take different values along the iterations. Subsequently, $\mu(\boldsymbol{\sigma})$ will denote the chosen value of μ for a given value of $\boldsymbol{\sigma}$. Here again, two strategies can be distinguished for setting $\big(\mu(\boldsymbol{\sigma}^{(t)})\big)_{t\in\mathbb{N}}$, depending on the dependencies one want to eliminate through the DA strategy.

Alternative I: Eliminate the coupling induced by $\mathbf{D}(\boldsymbol{\sigma}^{(t)})$

A first option is to choose, at each iteration t, $\mu(\boldsymbol{\sigma}^{(t)})$ positive such that

$$\mu(\boldsymbol{\sigma}^{(t)}) = \frac{\epsilon}{\|\mathbf{D}(\boldsymbol{\sigma}^{(t)})\|_{\mathrm{S}}} = \epsilon \left(\min(\sigma_i^{(t)})_{i\in\mathbb{I}^{(t)}}\right)^2 \tag{35}$$

with $\epsilon \in]0,1[$ and

$$\mathbb{I}^{(t)} = \big\{i \in \{1, \dots, N\} \mid \sigma_i^{(t)} > 0\big\}. \tag{36}$$

The auxiliary variable is then drawn as follows:

$$\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\mathbf{H}\mathbf{x}^{(t)}, \boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\right) \tag{37}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)}) = \dfrac{1}{\mu(\boldsymbol{\sigma}^{(t)})}\mathbf{I}_N - \mathbf{D}(\boldsymbol{\sigma}^{(t)})$ is positive definite by construction. The minus logarithm of the posterior density $\mathsf{p}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}, \mathbf{z})$ is given by

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}) = \frac{1}{2\mu(\boldsymbol{\sigma})}\|\mathbf{H}\mathbf{x} - \mu(\boldsymbol{\sigma})\big(\mathbf{v} + \mathbf{D}(\boldsymbol{\sigma})\mathbf{z}\big)\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{38}$$

Alternative II: Eliminate the coupling induced by $\mathbf{H}^\top\mathbf{D}(\boldsymbol{\sigma}^{(t)})\mathbf{H}$

Similarly, in order to eliminate the coupling induced by the full matrix $\mathbf{H}^\top\mathbf{D}(\boldsymbol{\sigma}^{(t)})\mathbf{H}$, $\mu(\boldsymbol{\sigma}^{(t)})$ can be chosen at each iteration $t \in \mathbb{N}$ so as to satisfy

$$\mu(\boldsymbol{\sigma}) = \frac{\epsilon}{\|\mathbf{H}\|_{\mathrm{S}}^2 \|\mathbf{D}(\boldsymbol{\sigma})\|_{\mathrm{S}}} = \frac{\epsilon}{\|\mathbf{H}\|_{\mathrm{S}}^2} \left(\min(\sigma_i^{(t)})_{i\in\mathbb{I}^{(t)}}\right)^2 \tag{39}$$

with $\epsilon \in]0,1[$ and $\mathbb{I}^{(t)}$ is given by (36). Then, the auxiliary variable is drawn as

$$\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\mathbf{x}^{(t)}, \boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\right), \tag{40}$$

where $\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)}) = \dfrac{1}{\mu(\boldsymbol{\sigma}^{(t)})}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{D}(\boldsymbol{\sigma}^{(t)})\mathbf{H}$ is positive definite. The minus logarithm of the posterior density $\mathsf{p}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}, \mathbf{z})$ then reads

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}) = \frac{1}{2\mu(\boldsymbol{\sigma})}\|\mathbf{x} - \mu(\boldsymbol{\sigma})\big(\mathbf{v} + \mathbf{H}^\top\mathbf{D}(\boldsymbol{\sigma})\mathbf{z}\big)\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{41}$$

It is worth noting that $\boldsymbol{\sigma}$ and $\mathbf{v}$ are two dependent random variables conditionally to both $\mathbf{x}$ and $\mathbf{z}$. The resulting Gibbs samplers, corresponding to Alternatives I and II, respectively, are summarized in Algorithms 3 and 4.

Algorithm 3 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\mathbf{D}(\boldsymbol{\sigma})$ in the case of a scale mixture of Gaussian noise.

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^N$, $\boldsymbol{\sigma}^{(0)} \in \mathbb{R}_+^N$, $0 < \epsilon < 1$, $\mu(\boldsymbol{\sigma}^{(0)}) = \epsilon\big(\min(\sigma_i^{(0)})_{i\in\mathbb{I}^{(0)}}\big)^2$

1: **for** $t = 0, 1, \dots$ **do**

2: Generate

$\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\mathbf{H}\mathbf{x}^{(t)}, \boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\right)$ where $\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)}) = \dfrac{1}{\mu(\boldsymbol{\sigma}^{(t)})}\mathbf{I}_N - \mathbf{D}(\boldsymbol{\sigma}^{(t)})$

3: Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\boldsymbol{\sigma}^{(t)},\mathbf{z}}$

4: Generate $\boldsymbol{\sigma}^{(t+1)} \sim \mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)},\mathbf{z}}$

5: Set $\mu(\boldsymbol{\sigma}^{(t+1)}) = \epsilon\big(\min(\sigma_i^{(t+1)})_{i\in\mathbb{I}^{(t+1)}}\big)^2$

6: **end for**

3.2.3 Partially collapsed Gibbs sampling

It can be noted that it is generally complicated to sample from $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x},\mathbf{v},\mathbf{z}}$ due to the presence of $\mu(\boldsymbol{\sigma})$ and $\mathbf{D}(\boldsymbol{\sigma})$ in the conditional distribution of $\mathbf{v}$. One can replace this step by sampling from $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x},\mathbf{z}}$, that is directly sampling $\boldsymbol{\sigma}$ from its marginal posterior distribution with respect to $\mathbf{v}$ and conditionally to $\mathbf{x}$ and $\mathbf{z}$. In this case, we say that we are partially collapsing $\mathbf{v}$ in the Gibbs sampler. One of the main benefits of doing so is that, conditionally to $\mathbf{x}$ and $\mathbf{z}$, $\boldsymbol{\sigma}$ has independent components. However, as $\boldsymbol{\sigma}$ is sampled independently from $\mathbf{v}$, the constructed Markov chain $\big(\mathbf{x}^{(t)}, \boldsymbol{\sigma}^{(t)}, \mathbf{v}^{(t)}\big)_{t\geqslant 0}$ may have a transition kernel with an unknown stationary distribution [66]. This problem can also be encountered when the auxiliary variable $\mathbf{v}$ depends on other unknown hyperparameters changing along the algorithm such as prior covariance matrix or regularization

Algorithm 4 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\mathbf{H}^\top \mathbf{D}(\boldsymbol{\sigma})\mathbf{H}$ in the case of a scale mixture of Gaussian noise.

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^Q$, $\boldsymbol{\sigma}^{(0)} \in \mathbb{R}_+^N$, $0 < \epsilon < 1$, $\mu(\boldsymbol{\sigma}^{(0)}) = \epsilon \, \|\mathbf{H}\|_{\mathrm{S}}^{-2} \, \big(\min(\sigma_i^{(0)})_{i \in \mathbb{I}^{(0)}}\big)^2$

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

Generate

$$\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\mathbf{x}^{(t)}, \boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})\right) \text{ where } \boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)}) = \frac{1}{\mu(\boldsymbol{\sigma}^{(t)})}\mathbf{I}_Q - \mathbf{H}^\top \mathbf{D}(\boldsymbol{\sigma}^{(t)})\mathbf{H}$$

Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)}, \boldsymbol{\sigma}^{(t)}, \mathbf{z}}$

Generate $\boldsymbol{\sigma}^{(t+1)} \sim \mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x}^{(t+1)}, \mathbf{v}^{(t+1)}, \mathbf{z}}$

Set $\mu(\boldsymbol{\sigma}^{(t+1)}) = \epsilon \|\mathbf{H}\|_{\mathrm{S}}^{-2} \big(\min(\sigma_i^{(t+1)})_{i \in \mathbb{I}^{(t+1)}}\big)^2$

end for

parameter when the auxiliary variable is added to the prior instead of the likelihood. However, there exist some rules based on marginalization, permutation and trimming, that allow to replace the conditional distributions in the standard Gibbs sampler with conditional distributions marginalized according to some variables while ensuring that the target stationary distribution of the Markov chain is maintained. The resulting algorithm is known as the Partially Collapsed Gibbs Sampler (PCGS) [66]. Although this strategy can significantly decrease the complexity of the sampling process, it must be implemented with care to guarantee that the desired stationary distribution is preserved. Applications of PCGS algorithms can be found in [67, 68, 69].

Assume that, in addition to $\mathbf{x}$, $\boldsymbol{\sigma}$, $\mathbf{v}$, we have a vector $\boldsymbol{\Theta} \in \mathbb{R}^P$ of unknown parameters to be sampled. Note that, $\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \boldsymbol{\Theta}, \mathbf{v}|\mathbf{z})$ should be integrable with respect to all the variables. Following [66], we propose to use a PCGS algorithm that allows us to replace the full conditional distribution $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x},\mathbf{v},\boldsymbol{\Theta},\mathbf{z}}$ with its conditional distribution $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x},\boldsymbol{\Theta},\mathbf{z}}$ without affecting the convergence of the algorithm to the target stationary law. Algorithm 5 shows the main steps of the proposed sampler. It should be noted that, unlike the standard Gibbs algorithm, permuting the steps of this sampler may result in a Markov chain with an unknown stationary distribution.

Algorithm 5 PCGS in the case of a scale mixture of Gaussian noise

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^Q$, $\boldsymbol{\sigma}^{(0)} \in \mathbb{R}_+^N$, $\boldsymbol{\Theta}^{(0)} \in \mathbb{R}^P$

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

For all $i \in \{1, \ldots, N\}$, generate $\sigma_i^{(t+1)} \sim \mathcal{P}_{\sigma_i|\mathbf{x}^{(t)}, \boldsymbol{\Theta}^{(t)}, \mathbf{z}}$

Generate $\boldsymbol{\Theta}^{(t+1)} \sim \mathcal{P}_{\boldsymbol{\Theta}|\mathbf{x}^{(t)}, \boldsymbol{\sigma}^{(t+1)}, \mathbf{z}}$

Set $\mu(\boldsymbol{\sigma}^{(t)})$ and $\boldsymbol{\Gamma}(\boldsymbol{\sigma}^{(t)})$

Generate $\mathbf{v}^{(t+1)} \sim \mathcal{P}_{\mathbf{v}|\mathbf{x}^{(t)}, \boldsymbol{\sigma}^{(t+1)}, \boldsymbol{\Theta}^{(t+1)}, \mathbf{z}}$

Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)}, \boldsymbol{\sigma}^{(t+1)}, \boldsymbol{\Theta}^{(t+1)}, \mathbf{z}}$

end for

3.3 High dimensional Gaussian distribution

The proposed DA approach can also be applied to the problem of drawing random variables from a high dimensional Gaussian distribution with parameters $\mathbf{m}$ and $\mathbf{G}$ as defined in (5) and (6). The introduction of auxiliary variables can be especially useful to facilitate the sampling process in a number of problems that we discuss below. In order to make our presentation clearer, an additional index will be added to the variables $\mathbf{v}$ and μ introduced in Section 2.

- If the prior precision matrix $\mathbf{G_x}$ and the observation matrix $\mathbf{H}$ can be diagonalized in the same basis, it can be of interest of adding the auxiliary variable $\mathbf{v}_1$ in the data fidelity term. Following Algorithm 1, let $\mu_1 > 0$ such that $\mu_1\|\mathbf{\Lambda}\|_{\mathrm{S}} < 1$ and

$$\mathbf{v}_1 \sim \mathcal{N}\left(\left(\frac{1}{\mu_1}\mathbf{I}_N - \mathbf{\Lambda}\right)\mathbf{Hx}, \frac{1}{\mu_1}\mathbf{I}_N - \mathbf{\Lambda}\right). \tag{42}$$

  The resulting conditional distribution of the target signal $\mathbf{x}$ given the auxiliary variable $\mathbf{v}_1$ and the vector of observation $\mathbf{z}$ is a Gaussian distribution with the following parameters:

$$\widetilde{\mathbf{G}} = \frac{1}{\mu_1}\mathbf{H}^\top\mathbf{H} + \mathbf{G_x}. \tag{43}$$

$$\widetilde{\mathbf{m}} = \widetilde{\mathbf{G}}^{-1}\left(\mathbf{H}^\top\mathbf{\Lambda z} + \mathbf{G_x m_x} + \mathbf{H}^\top\mathbf{v}_1\right). \tag{44}$$

  Then, sampling from the target signal can be performed by passing to the transform domain where $\mathbf{H}$ and $\mathbf{G_x}$ are diagonalizable (e.g., Fourier domain when $\mathbf{H}$ and $\mathbf{G_x}$ are circulant) .

  Similarly, if it is possible to write $\mathbf{G_x} = \mathbf{V}^\top\mathbf{\Omega V}$, so as $\mathbf{H}$ and $\mathbf{V}$ can be diagonalized in the same basis, we suggest to introduce an extra auxiliary variable $\mathbf{v}_2$ independent of $\mathbf{v}_1$ in the prior term to eliminate the coupling introduced by $\mathbf{\Omega}$ when passing to the transform domain. Let $\mu_2 > 0$ be such that $\mu_2\|\mathbf{\Omega}\|_{\mathrm{S}} < 1$ and let the distribution of $\mathbf{v}_2$ conditionally to $\mathbf{x}$ be given by

$$\mathbf{v}_2 \sim \mathcal{N}\left(\left(\frac{1}{\mu_2}\mathbf{I}_N - \mathbf{\Omega}\right)\mathbf{Vx}, \frac{1}{\mu_2}\mathbf{I}_N - \mathbf{\Omega}\right). \tag{45}$$

  The joint distribution of the unknown parameters is given by

$$\mathsf{p}(\mathbf{x}, \mathbf{v}_1, \mathbf{v}_2|\mathbf{z}) = \mathsf{p}(\mathbf{x}|\mathbf{z})\mathsf{p}(\mathbf{v}_1|\mathbf{x}, \mathbf{z})\mathsf{p}(\mathbf{v}_2|\mathbf{x}, \mathbf{z}). \tag{46}$$

  It follows that the minus logarithm of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$, $\mathbf{v}_1$ and $\mathbf{v}_2$ is Gaussian with parameters:

$$\widetilde{\mathbf{G}} = \frac{1}{\mu_1}\mathbf{H}^\top\mathbf{H} + \frac{1}{\mu_2}\mathbf{V}^\top\mathbf{V} \tag{47}$$

  and

$$\widetilde{\mathbf{m}} = \widetilde{\mathbf{G}}^{-1}\left(\mathbf{H}^\top\mathbf{\Lambda z} + \mathbf{G_x m_x} + \mathbf{H}^\top\mathbf{v}_1 + \mathbf{V}^\top\mathbf{v}_2\right). \tag{48}$$

- If $\mathbf{G_x}$ and $\mathbf{H}$ are not diagonalizable in the same basis, the introduction of an auxiliary variable either in the data fidelity term or the prior allows us to eliminate the coupling between these two heterogeneous operators. Let $\mu_1 > 0$ such that $\mu_1\|\mathbf{H}^\top\mathbf{\Lambda H}\|_{\mathrm{S}} < 1$ and

$$\mathbf{v}_1 \sim \mathcal{N}\left(\left(\frac{1}{\mu_1}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda H}\right)\mathbf{x}, \frac{1}{\mu_1}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda H}\right). \tag{49}$$

Then, the parameters of the Gaussian posterior distribution of $\mathbf{x}$ given $\mathbf{v}_1$ read:

$$\widetilde{\mathbf{G}} = \frac{1}{\mu_1}\mathbf{I}_Q + \mathbf{G}_{\mathbf{x}} \tag{50}$$

$$\widetilde{\mathbf{m}} = \widetilde{\mathbf{G}}^{-1}\left(\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{z} + \mathbf{G}_{\mathbf{x}}\mathbf{m}_{\mathbf{x}} + \mathbf{v}_1\right). \tag{51}$$

Note that if $\mathbf{G}_{\mathbf{x}}$ has some simple structure (e.g,. diagonal, block diagonal, sparse, circulant,...), the precision matrix (50) will inherit this simple structure.

Otherwise, if $\mathbf{G}_{\mathbf{x}}$ does not present any specific structure, one could apply the proposed DA method to both data fidelity and prior terms. It suffices to introduce an extra auxiliary variable $\mathbf{v}_2$ in the prior law, additionally to the auxiliary variable $\mathbf{v}_1$ in (49). Let $\mu_2 > 0$ be such that $\mu_2\|\mathbf{G}_{\mathbf{x}}\|_{\mathrm{S}} < 1$ and

$$\mathbf{v}_2 \sim \mathcal{N}\left(\left(\frac{1}{\mu_2}\mathbf{I}_Q - \mathbf{G}_{\mathbf{x}}\right)\mathbf{x}, \frac{1}{\mu_2}\mathbf{I}_Q - \mathbf{G}_{\mathbf{x}}\right). \tag{52}$$

Then, the posterior distribution of $\mathbf{x}$ given $\mathbf{v}_1$ and $\mathbf{v}_2$ is Gaussian with the following parameters:

$$\widetilde{\mathbf{G}} = \frac{1}{\mu}\mathbf{I}_Q \tag{53}$$

and

$$\widetilde{\mathbf{m}} = \mu\left(\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{z} + \mathbf{G}_{\mathbf{x}}\mathbf{m}_{\mathbf{x}}\right) \tag{54}$$

where

$$\mu = \frac{\mu_1\mu_2}{\mu_1 + \mu_2}. \tag{55}$$

3.4 Sampling the auxiliary variable

It is clear that the main issue in the implementation of all the proposed Gibbs algorithms arises in the sampling of the auxiliary variable $\mathbf{v}$. The aim of this section is to propose efficient strategies for implementing this step at a limited computational cost, in the context of large scale problems.

For the sake of generality, we will consider that $\mathbf{v}$ follows a multivariate Gaussian distribution with covariance matrix of the form $\boldsymbol{\Gamma} = \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}$ where $\mu > 0$ satisfies $\mu\|\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}\|_{\mathrm{S}} < 1$. Our first suggestion is to set μ such that

$$\mu\|\mathbf{H}\|_{\mathrm{S}}^2 < \beta < \frac{1}{\|\boldsymbol{\Lambda}\|_{\mathrm{S}}}, \tag{56}$$

with $\beta > 0$. For example, one can set $\mu \leqslant \frac{\epsilon}{\|\mathbf{H}\|_{\mathrm{S}}^2\|\boldsymbol{\Lambda}\|_{\mathrm{S}}}$ and $\beta = \frac{\sqrt{\epsilon}}{\|\boldsymbol{\Lambda}\|_{\mathrm{S}}}$ where $0 < \epsilon < 1$. This allows to verify the requirement $\mu\|\mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H}\|_{\mathrm{S}} < 1$. Moreover, it leads to

$$\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\boldsymbol{\Lambda}\mathbf{H} = \frac{1}{\beta}\left(\frac{\beta}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right) + \mathbf{H}^\top\left(\frac{1}{\beta}\mathbf{I}_N - \boldsymbol{\Lambda}\right)\mathbf{H}. \tag{57}$$

Thus, the sampling step of the auxiliary variable at iteration $t \in \mathbb{N}$ can be replaced by the three following steps:

1) Generate $\mathbf{n}^{(t+1)} \sim \mathcal{N}\left(\mathbf{0}_N, \frac{1}{\beta}\mathbf{I}_N - \boldsymbol{\Lambda}\right)$.

2) Generate $\mathbf{y}^{(t+1)} \sim \mathcal{N}\left(\mathbf{0}_Q, \frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right)$ with $\lambda = \frac{\mu}{\beta} \leqslant \frac{\sqrt{\epsilon}}{\|\mathbf{H}\|_\mathrm{S}^2}$.

3) Compute
$\mathbf{v}^{(t+1)} = \left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\right)\mathbf{x}^{(t+1)} + \frac{1}{\sqrt{\beta}}\mathbf{y}^{(t+1)} + \mathbf{H}^\top\mathbf{n}^{(t+1)}$.

Hereabove, $\mathbf{y}^{(t+1)}$ and $\mathbf{n}^{(t+1)}$ are independent random variables. One can notice that the sampling problem of the auxiliary variables is now separated into two independent subproblems of sampling from large scale Gaussian distributions. The first sampling step can usually be performed efficiently. For instance, if $\mathbf{\Lambda}$ is diagonal (e.g., when the model is a scale mixture of Gaussian variables), coefficients $n_i^{(t+1)}$, $i \in \{1,\dots,N\}$, can be drawn separately. Let us now discuss the implementation of the second sampling step, requiring to sample from the zero mean Gaussian distribution with covariance matrix $\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}$.

- In the particular case when $\mathbf{H}$ is circulant, sampling can be performed in the Fourier domain. More generally, since $\mathbf{H}^\top\mathbf{H}$ is symmetric, there exists an orthogonal matrix $\mathbf{N}$ such that $\mathbf{N}\mathbf{H}^\top\mathbf{H}\mathbf{N}^\top$ is diagonal with positive diagonal entries. It follows that sampling from the Gaussian distribution with covariance matrix $\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}$ can be fulfilled easily within the basis defined by the matrix $\mathbf{N}$.

- Suppose that $\mathbf{H}$ satisfies $\mathbf{H}\mathbf{H}^\top = \nu\mathbf{I}_N$ with $\nu > 0$, which is the case for example of tight frame synthesis operators or decimation matrices. Note that $\nu\lambda \leqslant \sqrt{\epsilon} < 1$. We have then:

$$\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H} = \left(\frac{1}{\sqrt{\lambda}}\mathbf{I}_Q - \sqrt{\lambda}\mathbf{H}^\top\mathbf{H}\right)^2 + (1-\lambda\nu)\,\mathbf{H}^\top\mathbf{H}. \tag{58}$$

  It follows that a sample from the Gaussian distribution with covariance matrix $\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}$ can be obtained as follows:

$$\mathbf{y}^{(t+1)} = \left(\frac{1}{\sqrt{\lambda}}\mathbf{I}_Q - \sqrt{\lambda}\mathbf{H}^\top\mathbf{H}\right)\mathbf{y}_1^{(t+1)} + \sqrt{1-\lambda\nu}\mathbf{H}^\top\mathbf{y}_2^{(t+1)} \tag{59}$$

  where $\mathbf{y}_1^{(t+1)} \in \mathbb{R}^Q$ and $\mathbf{y}_2^{(t+1)} \in \mathbb{R}^N$ are independent Gaussian random vectors with covariance matrices equal to $\mathbf{I}_Q$ and $\mathbf{I}_N$ respectively.

- Suppose that $\mathbf{H} = \mathbf{M}\mathbf{P}$ with $\mathbf{M} \in \mathbb{R}^{N\times K}$ and $\mathbf{P} \in \mathbb{R}^{K\times Q}$. Hence, one can set $\lambda > 0$ and $\widetilde{\lambda} > 0$ such that

$$\lambda\|\mathbf{P}\|^2 < \widetilde{\lambda} < \frac{1}{\|\mathbf{M}\|^2}. \tag{60}$$

  For example, for $\mu = \frac{\epsilon}{\|\mathbf{P}\|_\mathrm{S}^2\|\mathbf{M}\|_\mathrm{S}^2\|\mathbf{\Lambda}\|_\mathrm{S}}$, we have $\lambda = \frac{\sqrt{\epsilon}}{\|\mathbf{P}\|_\mathrm{S}^2\|\mathbf{M}\|_\mathrm{S}^2}$. Then, we can set $\widetilde{\lambda} = \frac{\epsilon^{1/4}}{\|\mathbf{M}\|_\mathrm{S}^2}$. It follows that

$$\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H} = \frac{1}{\widetilde{\lambda}}\left(\frac{\widetilde{\lambda}}{\lambda}\mathbf{I}_Q - \mathbf{P}^\top\mathbf{P}\right) + \mathbf{P}^\top\left(\frac{1}{\widetilde{\lambda}}\mathbf{I}_K - \mathbf{M}^\top\mathbf{M}\right)\mathbf{P}. \tag{61}$$

It appears that, if it is possible to draw merely random vectors $\mathbf{y}_1^{(t+1)}$ and $\mathbf{y}_2^{(t+1)}$ from the Gaussian distributions with covariance matrices $\frac{\widetilde{\lambda}}{\lambda}\mathbf{I}_Q - \mathbf{P}^\top\mathbf{P}$ and $\frac{1}{\widetilde{\lambda}}\mathbf{I}_K - \mathbf{M}^\top\mathbf{M}$ respectively (for example when $\mathbf{P}$ is a tight frame analysis operator and $\mathbf{M}$ is a convolution matrix with periodic boundary condition), a sample from the Gaussian distribution with a covariance matrix $\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}$ can be obtained as follows:

$$\mathbf{y}^{(t+1)} = \frac{1}{\sqrt{\widetilde{\lambda}}}\mathbf{y}_1^{(t+1)} + \mathbf{P}^\top\mathbf{y}_2^{(t+1)}. \tag{62}$$

4 Application to multichannel image recovery in the presence of Gaussian noise

We now discuss the performance of the proposed DA strategies in the context of restoration of multichannel images (MCI). Such images are widely used in many application areas such as medical imaging and remote sensing [70, 71, 72]. Multiple channel components typically result from imaging a single scene by sensors operating in different spectral ranges. For instance, about a dozen of radiometers may be on-board remote sensing satellites. Most of the time, MCI are corrupted with noise and blur arising from the acquisition process and transmission steps. Therefore, restoring MCI is of primary importance as a preliminary step before addressing analysis tasks such as classification, segmentation or object recognition [73]. Several works dedicated to MCI processing rely on wavelet-based approaches [71, 74]. In this section, we propose to adopt a Bayesian framework for recovering the wavelet coefficients of deteriorated MCI, with the aim to analyze the performance of the aforementioned hybrid Gibbs samplers.

4.1 Problem formulation

Let us consider the problem of recovering a multicomponent image with B components $\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_B$ in $\mathbb{R}^R$ (the images being columnwise reshaped) from some observations $\mathbf{z}_1, \ldots, \mathbf{z}_B$ which have been degraded by spatially invariant blurring operators $\mathbf{B}_1, \ldots, \mathbf{B}_B$ and corrupted by independent zero-mean additive white Gaussian noises having the same known variance σ^2. As already stated, we propose here to address the restoration problem in a transform domain where the target images are assumed to have a sparse representation. Let us introduce a set of tight frame synthesis operators $\mathbf{F}_1^*, \ldots, \mathbf{F}_B^*$ [75] such that

$$(\forall b \in \{1, \ldots, B\}) \quad \bar{\mathbf{y}}_b = \mathbf{F}_b^*\bar{\mathbf{x}}_b \tag{63}$$

where, for every $b \in \{1, \ldots, B\}$, $\mathbf{F}_b^*$ is a linear operator from $\mathbb{R}^K$ to $\mathbb{R}^R$ with $K \geqslant R$ and $\bar{\mathbf{x}}_b$ is the vector of frame coefficients of the image $\bar{\mathbf{y}}_b$. Each frame transform operator decomposes the image into M oriented subbands at multiple scales with sizes K_m, $m \in \{1, \ldots, M\}$, such that $\sum_{m=1}^M K_m = K$:

$$(\forall b \in \{1, \ldots, B\}) \quad \bar{\mathbf{x}}_b = \begin{array}{l}(\bar{x}_{b,1,1}, \ldots, \bar{x}_{b,1,K_1}, \ldots, \\ \bar{x}_{b,m,1}, \ldots, \bar{x}_{b,m,K_m}, \ldots, \\ \bar{x}_{b,M,1}, \ldots, \bar{x}_{b,M,K_M})^\top.\end{array} \tag{64}$$

Then, the problem can be formulated as (4), that is:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{65}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2 \mathbf{I}_N)$, $\mathbf{x} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_B^\top]^\top \in \mathbb{R}^Q$, $\mathbf{z} = [\mathbf{z}_1^\top, \ldots, \mathbf{z}_B^\top]^\top \in \mathbb{R}^N$, $\mathbf{H} = \mathbf{B}\mathbf{F}^* \in \mathbb{R}^{N\times Q}$ with $N = BR$, $Q = KB$,

$$\mathbf{F}^* = \begin{pmatrix} \mathbf{F}_1^* & \mathbf{0} & \ldots & \mathbf{0} \\ 0 & \mathbf{F}_2^* & \mathbf{0} & \mathbf{0} \\ \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{F}_B^* \end{pmatrix} \tag{66}$$

and

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ 0 & \mathbf{B}_2 & \mathbf{0} & \mathbf{0} \\ \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_B \end{pmatrix}. \tag{67}$$

We propose to exploit the cross-component similarities by estimating jointly the frame coefficients at a specific orientation and scale through all the B components. In this respect, for every $m \in \{1, \ldots, M\}$, for every $k \in \{1, \ldots, K_m\}$, let $\mathbf{x}_{m,k} = (x_{b,m,k})_{1\leqslant b\leqslant B} \in \mathbb{R}^B$ be the vector of frame coefficients for a given wavelet subband m at a spatial position k through all the B components. Note that such vector can be easily obtained through $\mathbf{x}_{m,k} = \mathbf{P}_{m,k}\mathbf{x}$ where $\mathbf{P}_{m,k} \in \mathbb{R}^{B\times Q}$ is a sparse matrix containing B lines of a suitable permutation matrix. To promote the sparsity of the wavelet coefficients and the inter-component dependency, following [71], we assume that for every $m \in \{1, \ldots, M\}$, the vectors $\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,K_m}$ are realizations of a random vector following a generalized multivariate exponential power ($\mathcal{GMEP}$) distribution with scale matrix $\mathbf{\Sigma}_m$, shape parameter β_m and smoothing parameter δ_m. Thus, the minus-log of the prior likelihood is given up to an additive constant by

$$-\log \mathsf{p}(\mathbf{x}|\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_M) = \sum_{m=1}^{M}\sum_{k=1}^{K_m} \psi_m(\|\mathbf{\Sigma}_m^{-1/2}(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|) \tag{68}$$

where, for every $m \in \{1, \ldots, M\}$, $\mathbf{a}_m \in \mathbb{R}^B$ and for all $t \in \mathbb{R}$, $\psi_m(t) = \frac{1}{2}\,(t^2 + \delta_m)^{\beta_m}$.

Our goal is to compute the posterior mean estimate of the target image as well as the unknown regularization parameters using MCMC sampling algorithms accelerated thanks to our proposed DA strategies. In the following, we will denote by $\mathbf{\Theta}$ the vector of unknown regularization parameters to be estimated jointly with $\mathbf{x}$ in the Gibbs sampling algorithm.

4.2 Sampling from the posterior distribution of the wavelet coefficients

One can expect that the standard sampling algorithms fail to explore efficiently the target posterior not only because of the high dimensionality of the problem but also because of the anisotropic nature of the wavelet coefficients. In fact, the coefficients belonging to different scales are assumed to follow $\mathcal{GMEP}$ priors with different shapes β_m, $m \in \{1, \ldots, M\}$. For instance, coefficients belonging to the low resolution subband are generally assumed to be driven from a Gaussian distribution (i.e., $\beta_m = 1$) while $\mathcal{GMEP}$ priors with very small shape parameter (i.e., $\beta_m \leqslant \dfrac{1}{2}$) are generally assigned to high resolution subbands at the first level of decomposition in order to promote sparsity. Therein, one can better explore the directions of interest separately by using different amplitudes than sampling them jointly. However, the observation matrix causes high spatial dependencies between the coefficients and thus hinders processing the different wavelet subbands independently.

The DA approaches we introduced in Section 3 allow to tackle this preconditioning problem by adding auxiliary variables to the data fidelity term. More specifically, following Algorithm 2, we propose to introduce an auxiliary variable $\mathbf{v} \in \mathbb{R}^Q$ such that:

$$\mathbf{v} \sim \mathcal{N}\left(\frac{1}{\sigma^2}\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right)\mathbf{x}, \frac{1}{\sigma^2}\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right)\right) \tag{69}$$

where $\mu\|\mathbf{B}\|_{\mathrm{S}}^2\|\mathbf{F}\|_{\mathrm{S}}^2 < 1$.

Since the set of hyperparameters $\boldsymbol{\Theta}$ is independent of the auxiliary variable $\mathbf{v}$ when conditioned to $\mathbf{x}$, each iteration $t \in \mathbb{N}$ of the proposed Gibbs sampling algorithm contains the following steps:

1) Sample $\mathbf{v}^{(t+1)}$ from $\mathcal{P}_{\mathbf{v}|\mathbf{x}^{(t)},\mathbf{z}}$.
2) Sample $\mathbf{x}^{(t+1)}$ from $\mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\boldsymbol{\Theta}^{(t)},\mathbf{z}}$.
3) Sample $\boldsymbol{\Theta}^{(t+1)}$ from $\mathcal{P}_{\boldsymbol{\Theta}|\mathbf{x}^{(t+1)},\mathbf{z}}$.

If $\mathbf{B}$ is circulant (by assuming periodic boundary conditions of the blur kernel), the first sampling step can be easily done by passing to the Fourier domain. In particular, if $\mathbf{F}$ is orthonormal that is $\mathbf{F}\mathbf{F}^* = \mathbf{F}^*\mathbf{F} = \mathbf{I}_Q$, samples of the auxiliary variables can be obtained by first drawing Gaussian random variables in the Fourier domain and then passing to the wavelet domain. Otherwise, if $\mathbf{F}$ is a non orthonormal transform, sampling can be performed using our results stated in (59) and (62).

Note that, in the new augmented space, the restoration problem reduces to a denoising problem with zero-mean Gaussian noise of variance μ and the posterior density reads:

$$\mathsf{p}(\mathbf{x}|\mathbf{z},\mathbf{v},\boldsymbol{\Theta}) \propto \prod_{m=1}^{M}\prod_{k=1}^{K_m} \exp\left(-\mathcal{J}_{m,k}(\mathbf{P}_{m,k}\mathbf{x}|\mathbf{v})\right) \tag{70}$$

where

$$(\forall \mathbf{c} \in \mathbb{R}^B) \quad \mathcal{J}_{m,k}(\mathbf{c}|\mathbf{v}) = \frac{1}{2\mu\sigma^2}\|\mathbf{c} - \mu\mathbf{P}_{m,k}\mathbf{v} - \frac{\mu}{\sigma^2}\mathbf{P}_{m,k}\mathbf{H}^\top\mathbf{z}\|^2 + \psi_m(\|\boldsymbol{\Sigma}_m^{-1/2}(\mathbf{c} - \mathbf{a}_m)\|). \tag{71}$$

It follows that we can draw samples of vectors $\mathbf{x}_{m,k}$, $m \in \{1,\ldots,M\}$, $k \in \{1,\ldots,K_m\}$, in an independent manner. Thus, the resolution of the initial high dimensional problem of size $Q = KB$ reduces to the resolution of K parallel subproblems of size B. This is particularly interesting in the case of MCI where we have generally $K \gg B$.

Instead of processing all the different wavelet coefficients at the same time, the proposed method allows to deal independently with each subproblem. This avoids sampling problems related to the heterogeneous prior distribution. Different sampling algorithms may be chosen according to the properties of the target distribution in each subproblem. Specifically, for each sampling subproblem, we propose to use either RW or MALA algorithms [17, 76].

In the following, we will discuss the practical implementation of the third step of the Gibbs algorithm namely sampling from the posterior distribution of $\boldsymbol{\Theta}$.

4.3 Hyperparameters estimation

Separation strategy: For every $m \in \{1,\ldots,M\}$, β_m controls the shape of the $\mathcal{GMEP}$ distribution allowing for heavier tails than the Laplace distribution ($\beta_m < 0.5$) and approaching the normal distribution when β_m tends to 1. In this work, we assume that, for every $m \in \{1,\ldots,M\}$,

β_m and δ_m are fixed. Actually, the shape parameter is set to different values with respect to the resolution level, spanning from very small values ($\beta_m < 0.5$) in order to enforce sparsity in the detail subbands at the first levels of decomposition to relatively higher values ($0.5 < \beta_m < 1$) for detail subband at higher resolution levels, whereas a Gaussian distribution is generally assigned to the low frequency subband. Furthermore, we set δ_m to a positive small value ensuring that (78) is differentiable [71]. As already mentioned, the scale matrices $(\mathbf{\Sigma}_m)_{1\leq m\leq M}$ will be estimated. Let us define $\mathcal{P}_{\mathbf{\Sigma}_m}$ the prior distribution of the scale matrix for each subband $m \in \{1,\ldots,M\}$ and $\mathsf{p}(\mathbf{\Sigma}_m)$ its related density. The associated posterior density reads

$$\mathsf{p}(\mathbf{\Sigma}_m|\mathbf{x}) \propto \mathsf{p}(\mathbf{\Sigma}_m)\det(\mathbf{\Sigma}_m)^{-K_m/2}\exp\left(-\sum_{k=1}^{K_m}\psi_m(\|\mathbf{\Sigma}_m^{-1/2}(\mathbf{P}_{m,k}\mathbf{x}-\mathbf{a}_m)\|)\right). \tag{72}$$

When $\beta_m = 1$, the $\mathcal{GMEP}$ prior reduces to a Gaussian distribution. In such case, a common choice of $\mathcal{P}_{\mathbf{\Sigma}_m}$ is an inverse Wishart distribution and (72) is also an inverse Wishart distribution [77]. However, when $0 < \beta_m < 1$, (72) does not belong to classical families of matrix distributions. In that respect, rather than estimating the scale matrices directly, we resort to a separation strategy. More specifically, we propose to estimate the standard deviations and the correlation terms independently. Let us decompose the scale matrix for each subband $m \in \{1,\ldots,M\}$ as follows [78]:

$$\mathbf{\Sigma}_m = C_{\beta_m,\delta_m}\mathrm{Diag}(\mathbf{s}_m)^{-1}\mathbf{R}_m\mathrm{Diag}(\mathbf{s}_m)^{-1} \tag{73}$$

where $\mathbf{R}_m \in \mathbb{R}^{B\times B}$ is the correlation matrix (whose diagonal elements are equal to 1 and the remaining ones define the correlation between the coefficients and have absolute value smaller than 1), $\mathbf{s}_m \in \mathbb{R}^B$ is a vector formed by the square root of the precision parameters (the inverse of standard deviations) and C_{β_m,δ_m} is a multiplicative constant that depends on β_m and δ_m [71]. The advantage of such factorization can be explained by the fact that the estimation of the correlation matrix will not alter the estimation of the variances. For every $m \in \{1,\ldots,M\}$, we decompose the precision vector as follows:

$$\mathbf{s}_m = (C_{\beta_m,\delta_m})^{1/2}\gamma_m^{1/(2\beta_m)}\mathbf{n}_m \tag{74}$$

where γ_m is positive and $\mathbf{n}_m \in \mathbb{R}^B$ is a vector of positive coefficients whose sum is equal to 1. Then, $\mathbf{n}_m$ can be seen as the vector containing positive normalized weights of all the B components in the subband m.

For simplicity, let us assume that the different 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 every $m \in \{1,\ldots,M\}$. Furthermore, let us suppose that $\mathbf{n}$ is known. We have then

$$\mathbf{\Theta} = \{\mathbf{R},\gamma_1,\ldots,\gamma_M\}. \tag{75}$$

Prior and posterior distribution for the hyperparameters: One can construct the correlation matrix $\mathbf{R}$ by sampling from an inverse Wishart distribution. Specifically, let $\mathbf{C} \sim \mathcal{IW}(\mathbf{A},c)$ where $\mathbf{A}$ is an appropriate positive definite matrix of $\mathbb{R}^{B\times B}$ and $c > 0$. Then, we can write $\mathbf{R} = \mathbf{\Delta}\mathbf{C}\mathbf{\Delta}$ where $\mathbf{\Delta}$ is the diagonal matrix whose elements are given by $\mathbf{\Delta}_{i,i} = C_{i,i}^{-1/2}$, for every $i \in \{1,\ldots,B\}$. Following [78], we can show that the prior density of $\mathbf{R}$ reads:

$$\mathsf{p}(\mathbf{R}) \propto \det(\mathbf{R})^{-\frac{B+1+c}{2}}\prod_{i=1}^{B}(\mathbf{R}^{-1}\mathbf{A})_{i,i}^{-\frac{\nu}{2}}. \tag{76}$$

In the following, we will use the notation $\mathbf{R} \sim \mathcal{SS}(\mathbf{A},c)$ to denote this prior. In particular, when $\mathbf{A} = \mathbf{I}_B$, individual correlations have the marginal density $\mathsf{p}(\rho_{i,j}) = (1-\rho_{i,j}^2)^{\frac{c-B-1}{2}}$ for every

$(i, j) \in \{1, \ldots, B\}^2$ such that $i \neq j$, which can be seen as a rectangular Beta distribution on the interval $[-1, 1]$ with both parameters equal to $(c-B+1)/2$. For $c = B+1$, we obtain marginally uniformly distributed correlations, whereas, by setting $B \leqslant c < B+1$ (or $B+1 < c$), we get marginal priors with heavier (or lighter) tails than the uniform distribution that is, distributions that promote high correlation values around the extremity of the intervals (or near zero values), respectively [78]. Thus, the posterior distribution of $\mathbf{R}$ is given by

$$\mathsf{p}(\mathbf{R}|\mathbf{x}, \gamma_1, \ldots, \gamma_M) \propto \det(\mathbf{R})^{-\frac{B+1+c+Q}{2}} \exp\left(-\Psi(\mathbf{x})\right) \prod_{i=1}^{B} (\mathbf{R}^{-1}\mathbf{A})_{i,i}^{-\frac{c}{2}} \tag{77}$$

where

$$\Psi(\mathbf{x}) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \psi_m\big(\gamma_m^{1/(2\beta_m)} \|\mathbf{R}^{-\frac{1}{2}} \mathrm{Diag}(\mathbf{n})(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|\big). \tag{78}$$

We propose here to sample from (77) at each iteration $t \in \mathbb{N}$ using a MH algorithm with proposal $\mathcal{SS}(\tilde{\mathbf{A}}, \tilde{c})$ where $\tilde{\mathbf{A}}$ is set to the current value of $\mathbf{R}$ at iteration t and $\tilde{c}$ is chosen to achieve reasonable acceptance probabilities.

For every $m \in \{1, \ldots, M\}$, we assume a Gamma prior for γ_m that is $\gamma_m \sim \mathcal{G}(a_{\gamma_m}, b_{\gamma_m})$ where $a_{\gamma_m} > 0$ and $b_{\gamma_m} > 0$ [79]. Then, the posterior distribution of γ_m is given by:

$$\mathsf{p}(\gamma_m|\mathbf{x}, \mathbf{R}) \propto \gamma_m^{a_{\gamma_m} + \frac{K_m}{2\beta_m} - 1} \exp\left(-b_{\gamma_m}\gamma_m\right) \exp\left(-\frac{1}{2} \sum_{k=1}^{K_m} \left(\gamma_m^{\frac{1}{\beta_m}} \|\mathbf{R}^{-\frac{1}{2}} \mathrm{Diag}(\mathbf{n})(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|^2 + \delta_m\right)^{\beta_m}\right). \tag{79}$$

Note that, if $\delta_m = 0$, then (79) reduces to a Gamma distribution with parameters:

$$\tilde{a}_{\gamma_m} = a_{\gamma_m} + \frac{K_m}{2\beta_m}, \tag{80}$$

$$\tilde{a}_{\gamma_m} = b_{\gamma_m} + \sum_{k}^{K_m} \|\mathbf{R}^{-\frac{1}{2}} \mathbf{N}(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|^{2\beta_m}. \tag{81}$$

When $\delta_m > 0$, sampling from (79) will be performed using an independent MH algorithm with a Gamma proposal of parameters (80) and (81).

Initialization: We propose to set the prior distributions of $\mathbf{R}$, $\gamma_1, \ldots, \gamma_M$, using empirical estimators from the degraded image. In particular, a rough estimator of $\mathbf{R}$ can be computed from the subband containing the low resolution wavelet coefficients at the highest level of decomposition. In the case when $\mathbf{F}$ is orthonormal, the variance of wavelet coefficients of the original image are approximately related to those of the degraded image through:

$$(\forall b \in \{1, \ldots, B\})(\forall m \in \{1, \ldots, M\}) \quad \mathrm{var}([\mathbf{F}_b\mathbf{z}_b]_m) = \alpha_m \mathrm{var}([\mathbf{x}_b]_m) + \sigma^2, \tag{82}$$

where $[.]_m$ designates the wavelet coefficients belonging to the subband m and α_m is a positive constant which depends on the subband index m and on the blur matrix. Expression (82) is derived from the considered observation model (65) by assuming a constant approximation of the impulse response of the blur filter in each wavelet subband. Note that α_m can be calculated beforehand as follows. Given noisy-free data, we compute the original empirical variance for each wavelet subband. Then, we calculate again the new variances of the subbands when the data is blurred using matrix $\mathbf{B}$. The coefficients α_m are finally estimated for each wavelet subband by computing the ratio of the two variances by a linear regression. When α_m is not too small with respect to 1, estimators of $\mathrm{var}([\mathbf{x}_b]_m)$ can be reliably computed from α_m and $\mathrm{var}([\mathbf{F}_b\mathbf{z}_b]_m)$ using (82). We propose to use this method to compute estimators of the variances in subbands at the highest levels of decomposition and then to deduce the variances of the remaining subbands by using some properties of multiresolution wavelet decompositions. Note that each detail subband m, corresponds to a given orientation l (horizontal, vertical, diagonal) and a given scale j (related resolution level). Actually, the variances of the detail subbands can be assumed to follow a power law with respect to the scale of the subband which can be expressed as follows [80]:

$$\log \mathrm{var}([\mathbf{x}_b]_m) = \varrho_l j + \varpi_l \tag{83}$$

where ϱ_l and ϖ_l are constants depending on the orientation l of the subband m. Once the variances of subbands in the two highest levels of decomposition have been computed using (82), we can calculate ϱ_l and ϖ_l for each orientation l using the slope and the intercept of these variances from a log plot with respect to the scale j. The remaining variances are then estimated by using (83). We then deduce from these variances an empirical estimator of $\mathbf{n}$, and set the parameters of the prior distributions of $\gamma_1, \ldots, \gamma_M$.

4.4 Experimental results

In these experiments, we consider the Hydice hyperspectral[2] data composed of 191 components in the 0.4 to 2.4 μm region of the visible and infrared spectrum. The test image is constructed by taking only a portion of size 256×256 and $B = 6$ components of Hydice using the channels 52, 67, 82, 97, 112 and 127. Hence, the problem dimension is $N = 393,216$. The original image is artificially degraded by a uniform blur of size 5×5 and an additive zero-mean white Gaussian noise with variance $\sigma^2 = 9$ so that the initial signal-to-noise ratio (SNR) is 11.16 dB. We perform an orthonormal wavelet decomposition using the Symlet wavelet of order 3, carried out over three resolution levels, hence $M = 10$ and $Q = N$. For the subband corresponding to the approximation coefficients ($m = 10$), we choose a Gaussian prior (i.e., $\beta_m = 2$, $\delta_m = 0$). For the remaining subbands ($m \in \{1, \ldots, M-1\}$), we set $\delta_m = 10^{-4}$. Moreover, we set $\beta_m = 0.2$ for the detail subbands corresponding to the lowest level of decomposition, $\beta_m = 0.4$ for the second level of decomposition, and $\beta_m = 0.5$ for the third level of decomposition.

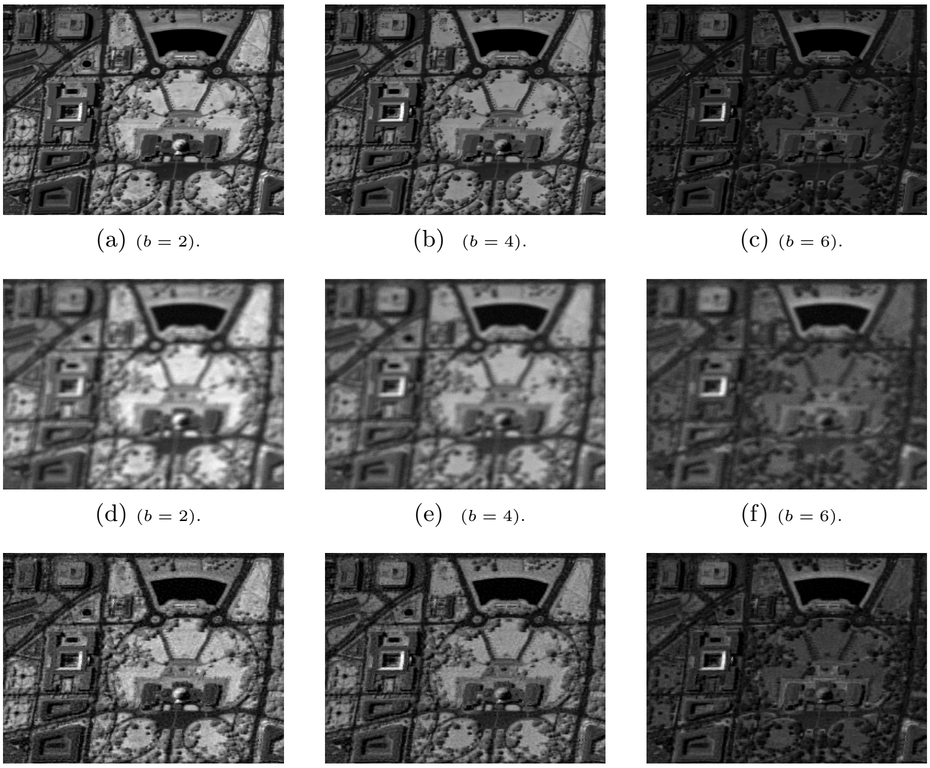
We run the Gibbs sampling Algorithm 2 with a sufficient number of iterations to reach stability. The obtained samples of the wavelet coefficients after the burn-in period are then used to compute the empirical MMSE estimator for the original image. Table 2 reports the results obtained for the different components in terms of SNR, PSNR (Peak Signal to Noise Ratio), BSNR (Blurred Signal to Noise Ratio) and SSIM (Structural SIMilarity). It can be noted that the MMSE estimator yields good numerical results. This can also be observed in Figure 1 showing the visual improvements for the different components of the multichannel image.

We propose to compare the performance of the Gibbs sampler with auxiliary variables when the posterior law of the wavelet coefficients is explored using either RW or MALA [17, 76] algorithms. We also compare the speed of our proposed approaches with standard RW and MALA without use of auxiliary variables. Figures 2 shows the evolution, with respect to the computational time, of the scale parameter γ_m in the horizontal subband for the first level of decomposition using the various algorithms. The results associated with the proposed algorithms appear in solid lines while those associated with standard algorithms without use of auxiliary variables are 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 large, the stepsize ε in standard algorithms is constrained to take very small values to allow appropriate acceptance probabilities whereas in the new augmented space, the subproblems dimension is

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

Table 2: Restoration results.

		$b=1$	$b=2$	$b=3$	$b=4$	$b=5$	$b=6$	Average
Initial	BSNR	24.27	30.28	31.73	28.92	26.93	22.97	27.52
	PSNR	25.47	21.18	19.79	22.36	23.01	26.93	23.12
	SNR	11.65	13.23	13.32	13.06	11.81	11.77	12.47
	SSIM	0.6203	0.5697	0.5692	0.5844	0.5558	0.6256	0.5875
MMSE	BSNR	32.04	38.33	39.21	38.33	35.15	34.28	36.22
	PSNR	28.63	25.39	23.98	26.90	27.25	31.47	27.27
	SNR	14.82	17.50	17.60	17.66	16.12	16.38	16.68
	SSIM	0.7756	0.8226	0.8156	0.8367	0.8210	0.8632	0.8225



(a) ($b = 2$). (b) ($b = 4$). (c) ($b = 6$).

(d) ($b = 2$). (e) ($b = 4$). (f) ($b = 6$).

(g) ($b = 2$). (h) ($b = 4$). (i) ($b = 6$).

Figure 1: From top to bottom: Original images-Degraded images-Restored images.

smaller allowing large moves to be accepted with high probability values. Note that MALA algorithm with auxiliary variables exhibits the best performance in terms of convergence speed. We summarize the obtained samples using the proposed algorithms by showing the marginal means and standard deviations of the hyperparameters in Table 3. It can be noted that the two proposed algorithms provide similar estimation results.

It is worth noting that for larger dimensional problems (i.e., larger values of B), one could further improve the efficiency of the proposed algorithm by exploiting the parallel structure of the sampling tasks.

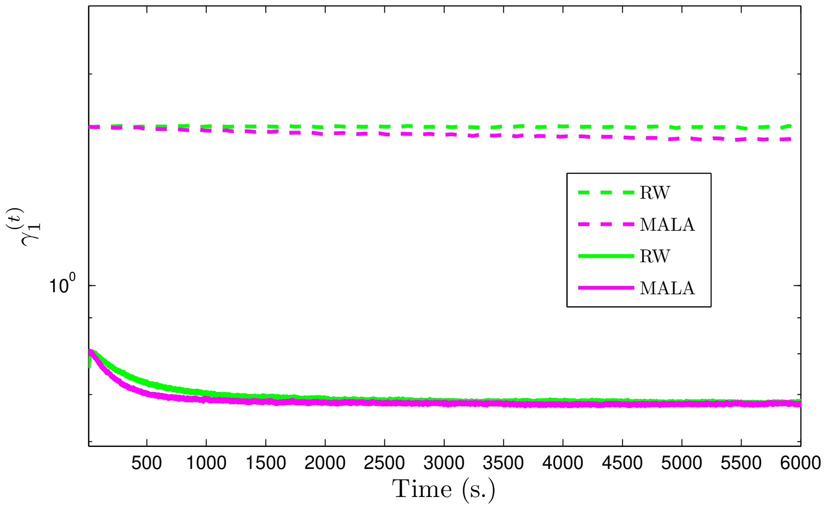


Figure 2: Trace plot of the scale parameter in subband $m = 1$ as time (horizontal subband in the first level of decomposition) with (dashed lines) and without (continuous line) auxiliary variables.

Table 3: Mean and variance estimates of hyperparameters.

		RW	MALA
$\hat{\gamma_1}$ (γ_1=0.71)	Mean Std.	0.67 (1.63 e-3)	0.67 (1.29 e-3)
$\hat{\gamma_2}$ (γ_2=0.99)	Mean Std.	0.83 (1.92 e-3)	0.83 (2.39 e-3)
$\hat{\gamma_3}$ (γ_3=0.72)	Mean Std.	0.62 (1.33 e-3)	0.61 (1.23 e-3)
$\hat{\gamma_4}$ (γ_4=0.0.24)	Mean Std.	0.24 (1.30 e-3)	0.24 (1.39 e-3)
$\hat{\gamma_5}$ (γ_5=0.40)	Mean Std.	0.37 (2.10 e-3)	0.37 (2.42 e-3)
$\hat{\gamma_6}$ (γ_6=0.22)	Mean Std.	0.21 (1.19 e-3)	0.21 (1.25 e-3)
$\hat{\gamma_7}$ (γ_7=0.0.07)	Mean Std.	0.08 (0.91 e-3)	0.08 (1.08 e-3)
$\hat{\gamma_8}$ (γ_8=0.13)	Mean Std.	0.13 (1.60 e-3)	0.13 (1.64 e-3)
$\hat{\gamma_9}$ (γ_9=0.07)	Mean Std.	0.07 (0.83 e-3)	0.07 (1 e-3)
$\hat{\gamma_1}0$ (γ_{10}=7.44 e-4)	Mean Std.	7.80 e-4 (1.34 e-5)	7.87 e-4 (2.12 e-5)
det($\hat{\mathbf{R}}$) det($\mathbf{R}$)= 5.79 e-8	Mean Std.	1.89 e-8 (9.96 e-10)	2.10 e-8 (2.24 e-9)

5 Application to image recovery in the presence of two terms mixed Gaussian noise

5.1 Problem formulation

In this second experiment, we consider the observation problem defined in (29) where $\mathbf{H}$ corresponds to a spatially invariant blur with periodic boundary conditions and the noise is a two-terms mixed Gaussian variable i.e., for every $i \in \{1, \dots, N\}$, $w_i \sim \mathcal{N}(0, \sigma_i^2)$ such that

$$\sigma_i \sim (1-\beta)\delta_{\kappa_1} + \beta\delta_{\kappa_2} \tag{84}$$

where κ_1, κ_2 are positive, $0 < \beta < 1$ is the probability that the variance of the noise σ_i equals κ_2, and δ_{κ_1} and δ_{κ_2} denote the discrete measures concentrated at the values κ_1 and κ_2 respectively. Model (84) can approximate for example mixed impulse Gaussian noise arising in radar, acoustic, and mobile radio applications [81, 82]. In this case, the impulse noise is approximated with a Gaussian one with a large variance $\kappa_2 \gg \kappa_1$ and β represents the probability of occurrence of the impulse noise. In the following, we assume without loss of generality that $\kappa_2 \geqslant \kappa_1$. We address the problem of estimating $\mathbf{x}$, $\boldsymbol{\sigma}$, β, κ_1 and κ_2 from the observations $\mathbf{z}$.

Prior distributions: We propose to use conjugate priors for the unknown variances namely inverse Gamma distributions i.e., $\kappa_i^2 \sim \mathcal{IG}(a_i, b_i)$, $i \in \{1, 2\}$ where a_i and b_i are positive constants. Here, a_1, a_2, b_1, and b_2 are set in practice to small values to ensure weakly informative priors. For the occurrence probability β, we choose a uniform prior distribution i.e., $\beta \sim \mathcal{U}(0, 1)$. Furthermore, the target image is assumed to follow a zero-mean Gaussian prior with a covariance matrix $\mathbf{G}_{\mathbf{x}}^{-1} = \gamma^{-1} \left(\mathbf{L}^\top \mathbf{L}\right)^{-1}$ known up to a precision parameter $\gamma > 0$, i.e.,

$$\mathsf{p}(\mathbf{x}|\gamma) \propto \gamma^{-Q/2} \exp\left(-\frac{\gamma}{2}\|\mathbf{L}\mathbf{x}\|^2\right). \tag{85}$$

Different covariance matrices may be chosen depending on which properties one wants to impose on the estimated image. In this example, we propose to enforce smoothness by setting $\mathbf{L} = \delta\mathbf{I}_Q - \nabla_2$ where ∇_2 is the circulant convolution matrix associated with a Laplacian filter and $\delta > 0$ is a small constant that aims at ensuring the positive definiteness of the prior covariance matrix. We further assume that the regularization parameter γ follows an inverse Gamma prior with parameters $a_\gamma > 0$ and $b_\gamma > 0$. The resulting hierarchical model is displayed in Figure 3.

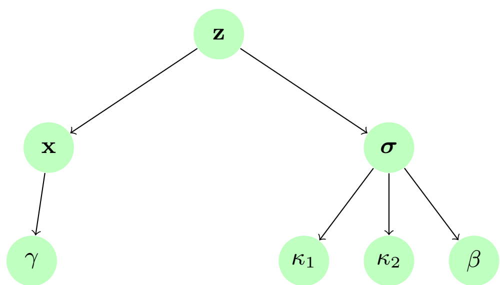


Figure 3: Hierarchical model for image deblurring under two term mixed Gaussian noise.

Posterior distributions: Given the observation model and the prior distribution, we can deduce that the posterior distribution of the target signal given $\boldsymbol{\sigma}$, β, κ_1^2, κ_2^2, γ and $\mathbf{z}$ is also Gaussian with mean $\mathbf{m}$ and precision matrix $\mathbf{G}$ given by:

$$\mathbf{G} = \mathbf{H}^\top \mathbf{D}\mathbf{H} + \gamma\mathbf{L}^\top\mathbf{L}, \tag{86}$$

$$\mathbf{m} = \mathbf{G}^{-1}\mathbf{H}^\top\mathbf{D}\mathbf{y}, \tag{87}$$

where $\mathbf{D}$ is the diagonal matrix with diagonal elements $D_{i,i} = \sigma_i^{-2}$, $i \in \{1, \dots, N\}$.

The posterior distribution of the remaining unknown parameters are given by:

- $(\forall i \in \{1,\dots,N\}) \quad \sigma_i|\mathbf{x},\beta,\kappa_1^2,\kappa_2^2,\mathbf{z} \sim (1-p_i)\delta_{\kappa_1} + p_i\delta_{\kappa_2}$ where $p_i = \dfrac{\eta_i}{1+\eta_i}$ such that

$$\eta_i = \frac{\beta}{1-\beta}\exp\left(-\frac{1}{2}\left(\kappa_2^{-2}-\kappa_1^{-2}\right)\left([\mathbf{H}\mathbf{x}]_i - z_i\right)^2\right)\frac{\kappa_1}{\kappa_2}, \tag{88}$$

- $\beta|\mathbf{x},\mathbf{z},\boldsymbol{\sigma},\kappa_1^2,\kappa_2^2 \sim \mathcal{B}(n_2+1,n_1+1)$, where $\mathcal{B}$ is the Beta distribution and n_1 and n_2 are the cardinals of the sets $\{i \in \{1,\dots,N\}, \mid \ \sigma_i = \kappa_1\}$ and $\{i \in \{1,\dots,N\}, \mid \ \sigma_i = \kappa_2\}$, respectively, so that $n_1+n_2 = N$,
- $\kappa_1^2|\mathbf{x},\boldsymbol{\sigma},\beta,\mathbf{z} \sim \mathcal{IG}\left(a_1 + \frac{n_1}{2}, b_1 + \sum_{i|\sigma_i=\kappa_1}\dfrac{\left([\mathbf{H}\mathbf{x}]_i - z_i\right)^2}{2}\right)$,
- $\kappa_2^2|\mathbf{x},\boldsymbol{\sigma},\beta,\mathbf{z} \sim \mathcal{IG}\left(a_2 + \frac{n_2}{2}, b_2 + \sum_{i|\sigma_i=\kappa_2}\dfrac{\left([\mathbf{H}\mathbf{x}]_i - z_i\right)^2}{2}\right)$,
- $\gamma|\mathbf{x} \sim \mathcal{G}\left(\dfrac{Q}{2} + a_\gamma, \dfrac{1}{2}\|\mathbf{L}\mathbf{x}\|^2 + b_\gamma\right)$.

5.2 Sampling from the posterior distribution of x

In the Gibbs algorithm, we need to draw samples from the multivariate Gaussian distribution of parameters (86) and (87) changing along the sampling iterations. In particular, even if $\mathbf{H}$ and $\mathbf{L}$ are circulant matrices, sampling cannot be done in the Fourier domain because of the presence of $\mathbf{D}$. In the sequel, we will use the method proposed in Section 3.3 to sample from this multivariate Gaussian distribution. More specifically, we exploit the flexibility of the proposed approach by resorting to two variants. In the first variant, we take advantage of the fact that $\mathbf{L}$ and $\mathbf{H}$ are diagonalizable in the Fourier domain and we propose to add the auxiliary variable to the data fidelity term in order to get rid of the coupling caused by $\mathbf{D}$ when passing to the Fourier domain. In the second variant, we introduce auxiliary variables for both the data fidelity and the prior terms in order to eliminate the coupling effects induced by all linear operators in the posterior distribution of the target image.

First variant: We introduce the variable $\mathbf{v}$ whose conditional distribution, given the set of main parameters of the problem, is the Gaussian distribution of mean $\left(\dfrac{1}{\mu}\mathbf{I}_N - \mathbf{D}\right)\mathbf{H}\mathbf{x}$ and covariance matrix $\left(\dfrac{1}{\mu}\mathbf{I}_N - \mathbf{D}\right)$ where $\mu = \epsilon\|\mathbf{D}\|_{\mathrm{S}}^{-1}$ with $0<\epsilon<1$. In practice, we set $\epsilon = 0.99$. It follows that the new conditional distribution of the target signal is

$$\mathbf{x}|\boldsymbol{\sigma},\beta,\kappa_1^2,\kappa_2^2,\gamma,\mathbf{v},\mathbf{z} \sim \mathcal{N}(\widetilde{\mathbf{m}},\widetilde{\mathbf{G}}^{-1}) \tag{89}$$

where $\widetilde{\mathbf{m}}$ and $\widetilde{\mathbf{G}}$ are defined as follows:

$$\widetilde{\mathbf{G}} = \frac{1}{\mu}\mathbf{H}^\top\mathbf{H} + \gamma\mathbf{L}^\top\mathbf{L}, \tag{90}$$

$$\widetilde{\mathbf{m}} = \widetilde{\mathbf{G}}^{-1}\mathbf{H}^\top\left(\mathbf{H}^\top\mathbf{D}\mathbf{z} + \mathbf{v}\right). \tag{91}$$

It is worth noting that the auxiliary variable $\mathbf{v}$ depends on $\mathbf{x}$, and also on $\boldsymbol{\sigma}$ through μ and $\mathbf{D}$, but does not depend on β, κ_1, κ_2, γ when conditioned to $\mathbf{x}$, $\boldsymbol{\sigma}$ and $\mathbf{z}$. Thus, we propose to use the partially collapsed Gibbs sampling algorithm in order to collapse the auxiliary variables in the sampling step of $\boldsymbol{\sigma}$. At each iteration $t \in \mathbb{N}$, the proposed algorithm goes through the following steps in an ordered manner:

AuxV1

1) Sample $(\kappa_1^2)^{(t+1)}$ from $\mathcal{P}_{\kappa_1^2|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},\beta^{(t)},\mathbf{z}}$.
2) Sample $(\kappa_2^2)^{(t+1)}$ from $\mathcal{P}_{\kappa_2^2|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},\beta^{(t)},\mathbf{z}}$.
3) Sample $\beta^{(t+1)}$ from $\mathcal{P}_{\beta|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},(\kappa_1^2)^{(t+1)},(\kappa_2^2)^{(t+1)}}$.
4) Sample $\gamma^{(t+1)}$ from $\mathcal{P}_{\gamma|\mathbf{x}^{(t)}}$.
5) Sample $\boldsymbol{\sigma}^{(t+1)}$ from $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x}^{(t)},\beta^{(t+1)},(\kappa_1^2)^{(t+1)},(\kappa_2^2)^{(t+1)},\mathbf{z}}$.
6) Set $\mu^{(t+1)} = \epsilon \min\left(\sigma_i^{(t+1)}\right)^{-2}_{1\leqslant i\leqslant N}$ and sample $\mathbf{v}^{(t+1)}$ from $\mathcal{P}_{\mathbf{v}|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t+1)}}$.
7) Sample $\mathbf{x}^{(t+1)}$ from $\mathcal{P}_{\mathbf{x}|\boldsymbol{\sigma}^{(t+1)},\gamma^{(t+1)},\mathbf{v}^{(t+1)},\mathbf{z}}$.

Second variant: Another strategy is to introduce two independent auxiliary variables $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbb{R}^Q$ following Gaussian distributions of means $\mathbf{\Gamma}_1\mathbf{x}$ and $\mathbf{\Gamma}_2\mathbf{x}$ and covariance matrices $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$, respectively, where

$$\mathbf{\Gamma}_1 = \frac{1}{\mu_1} - \mathbf{H}^\top\mathbf{D}\mathbf{H} \tag{92}$$

and

$$\mathbf{\Gamma}_2 = \frac{1}{\mu_2} - \mathbf{L}^\top\mathbf{L}. \tag{93}$$

In practice, we set $\mu_1 = \epsilon\|\mathbf{H}\|_{\mathrm{S}}^{-2}\|\mathbf{D}\|_{\mathrm{S}}^{-1}$ and $\mu_2 = \epsilon\|\mathbf{L}\|_{\mathrm{S}}^{-2}$ where $\epsilon = 0.99$. Then, the posterior distribution of $\mathbf{x}$ conditioned to $\boldsymbol{\sigma}$, β, κ_1^2, κ_2^2, γ, $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{z}$ is Gaussian with mean $\widetilde{\mathbf{m}}$ and precision matrix $\widetilde{\mathbf{G}}$ defined as

$$\widetilde{\mathbf{G}} = \left(\frac{1}{\mu_1} + \frac{\gamma}{\mu_2}\right)\mathbf{I}_Q \tag{94}$$

and

$$\widetilde{\mathbf{m}} = \mu_1\mu_2\left(\gamma\mu_1 + \mu_2\right)^{-1}\left(\mathbf{H}^\top\mathbf{D}\mathbf{y} + \mathbf{v}_1 + \sqrt{\gamma}\mathbf{v}_2\right). \tag{95}$$

The auxiliary variable $\mathbf{v}_1$ depends implicitly on $\boldsymbol{\sigma}$ through $\mathbf{D}$ and μ but does not depend on the remaining parameters when conditioned to $\mathbf{x}$, $\boldsymbol{\sigma}$ and $\mathbf{z}$. Similarly, $\mathbf{v}_2$ does not depend on $\boldsymbol{\sigma}$, β, κ_1^2, κ_2^2, $\mathbf{v}_1$, γ when conditioned to $\mathbf{x}$ and $\mathbf{z}$. We propose a PCGS algorithm that allows to collapse $\mathbf{v}_1$ in the sampling step of $\boldsymbol{\sigma}$. Each iteration $t \in \mathbb{N}$ of the proposed PCGS algorithm is composed of the following arranged sampling steps.

AuxV2

1) Sample $(\kappa_1^2)^{(t+1)}$ from $\mathcal{P}_{\kappa_1^2|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},\beta^{(t)},\mathbf{z}}$.

2) Sample $(\kappa_2^2)^{(t+1)}$ from $\mathcal{P}_{\kappa_2^2|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},\beta^{(t)},\mathbf{z}}$.

3) Sample $\beta^{(t+1)}$ from $\mathcal{P}_{\beta|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t)},(\kappa_1^2)^{(t+1)},(\kappa_2^2)^{(t+1)}}$.

4) Sample $\gamma^{(t+1)}$ from $\mathcal{P}_{\gamma|\mathbf{x}^{(t)}}$.

5) Sample $\boldsymbol{\sigma}^{(t+1)}$ from $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x}^{(t)},\beta^{(t+1)},(\kappa_1^2)^{(t+1)},(\kappa_2^2)^{(t+1)},\mathbf{z}}$.

6) Sample $\mathbf{v}_2^{(t+1)}$ from $\mathcal{P}_{\mathbf{v}_2|\mathbf{x}^{(t)}}$.

7) Set $\mu_1^{(t+1)} = \epsilon\|\mathbf{H}\|_{\mathrm{S}}^{-2} \min\left(\sigma_i^{(t+1)}\right)^{-2}_{1\leqslant i\leqslant N}$ and sample $\mathbf{v}_1^{(t+1)}$ from $\mathcal{P}_{\mathbf{v}_1|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t+1)}}$.

8) Sample $\mathbf{x}^{(t+1)}$ from $\mathcal{P}_{\mathbf{x}|\boldsymbol{\sigma}^{(t+1)},\gamma^{(t+1)},\mathbf{v}_1^{(t+1)},\mathbf{v}_2^{(t+1)},\mathbf{z}}$.

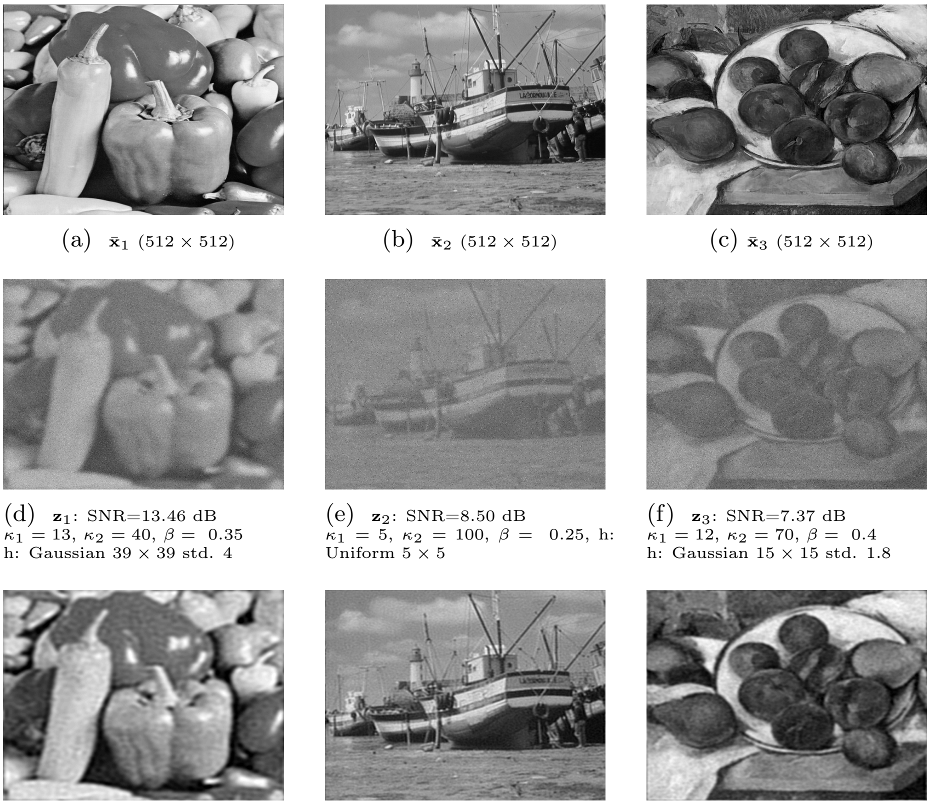
Note that, since $\mathbf{H}$ and $\mathbf{L}$ are circulant matrices and $\mathbf{D}$ is diagonal, sampling the auxiliary variables in the proposed methods can be easily performed following Section 3.4.

5.3 Experimental results

We consider a set of three test images denoted by $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\bar{\mathbf{x}}_3$, of size 512×512. These images are artificially degraded by a spatially invariant blur with point spread function h and further corrupted with mixed Gaussian noise. The Gibbs algorithms are run for 6,000 iterations and a burn-in period of 4,000 iterations is considered. Estimators of the unknown parameters are then computed using the empirical mean over the 2,000 obtained samples. Visual results are displayed in Figure 4 as well as estimates of hyper-parameters using AuxV1.

We focus now on image $\bar{\mathbf{x}}_1$ in order to compare the two variants of our proposed method with the Reversible Jump Perturbation Optimization (RJPO) algorithm [32]. For this method, we use the conjugate gradient algorithm as a linear solver at each iteration whose maximal number of iterations and tolerance are adjusted to correspond to an acceptance probability close to 0.9. We use the same initialization for all compared algorithms. Figures 5-8 display samples of hyperparameters as a function of iteration or time. Table 4 shows the marginal posterior mean and standard deviation of β, κ_1, κ_2, γ and the value of one randomly chosen pixel x_i in the reconstructed images. Figures 5-8 show that all the tested algorithms reach the same stationary distribution. In particular, it can be noted from Table 4 that β, κ_1 and κ_2 are correctly estimated by all the algorithms and the remaining parameters have similar estimators. While RJPO and AuxV1 have similar iterative behavior, AuxV2 needs more iterations to reach stability. However, the proposed two algorithms need less time to converge compared to RJPO algorithm since the computational cost of each iteration is highly reduced.

We also report comparisons in terms of mixing properties in convergence. Table 5 shows comparisons results in terms of time per iteration after the burn-in period (time needed to produce one sample), mean square jump in stationarity, and efficiency with respect to RJPO.



(a) $\bar{\mathbf{x}}_1$ (512 × 512)   (b) $\bar{\mathbf{x}}_2$ (512 × 512)   (c) $\bar{\mathbf{x}}_3$ (512 × 512)

(d) $\mathbf{z}_1$: SNR=13.46 dB $\kappa_1 = 13$, $\kappa_2 = 40$, $\beta = 0.35$ h: Gaussian 39 × 39 std. 4

(e) $\mathbf{z}_2$: SNR=8.50 dB $\kappa_1 = 5$, $\kappa_2 = 100$, $\beta = 0.25$, h: Uniform 5 × 5

(f) $\mathbf{z}_3$: SNR=7.37 dB $\kappa_1 = 12$, $\kappa_2 = 70$, $\beta = 0.4$ h: Gaussian 15 × 15 std. 1.8

(g) $\hat{\mathbf{x}}_1$: SNR=19.35 dB $\hat{\kappa}_1 = 12.98$, $\hat{\kappa}_2 = 39.80$ $\hat{\beta} = 0.35$, $\hat{\gamma} =$ 4.8e-3

(h) $\hat{\mathbf{x}}_2$: SNR=22 dB $\hat{\kappa}_1 = 5.10$, $\hat{\kappa}_2 = 100.13$ $\hat{\beta} = 0.25$, $\hat{\gamma} = 1.8$e-3

(i) $\hat{\mathbf{x}}_3$: SNR=18.74 dB $\hat{\kappa}_1 = 12.08$, $\hat{\kappa}_2 = 69.89$ $\hat{\beta} = 0.39$, $\hat{\gamma} = 4.7$e-3

Figure 4: Visual results. From top to bottom: Original images. Degraded images. Restored images.

Table 4: Mean and variance estimates.

		RJPO	AuxV1	AuxV2
$\hat{\gamma}$	Mean	4.78 e-3	4.84 e-3	4.90 e-3
(γ =5.30 e-3)	Std.	(1.39 e-4)	(1.25 e-4)	(9.01 e-5)
$\hat{\kappa_1}$	Mean	12.97	12.98	12.98
(κ_1=13)	Std.	(4.49 e-2)	(4.82 e-2)	(4.91 e-2)
$\hat{\kappa_2}$	Mean	39.78	39.77	39.80
(κ_1=40)	Std.	(0.13)	(0.14)	(0.13)
$\hat{\beta}$	Mean	0.35	0.35	0.35
(β=0.35)	Std.	(2.40 e-3)	(2.71 e-3)	(2.72 e-3)
$\hat{x_i}$	Mean	143.44	143.19	145.91
(x_i=140)	Std.	(10.72)	(11.29)	(9.92)

The speed improvement of the proposed algorithms comes with a deterioration of the quality of the generated samples due to the correlation existing between successive samples. For instance, RJPO algorithm gives the best results in terms of mean square jump in stationary. However, the generation of every sample is costly, which deteriorates the efficiency of the algorithm for large scale problems compared with AuxV2. The best trade-off between convergence speed and

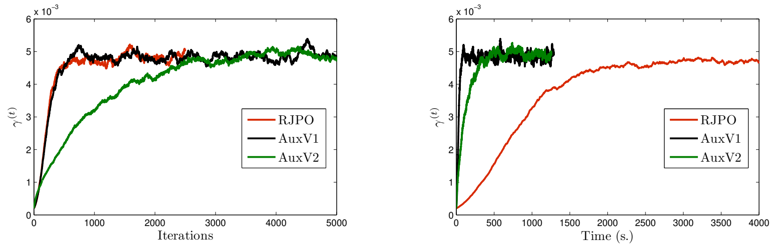


Figure 5: Chains of γ versus iteration/time.

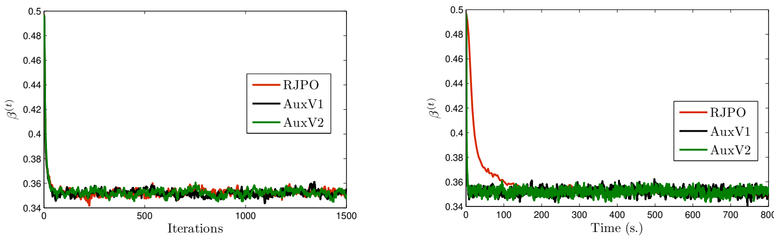


Figure 6: Chains of β versus iteration/time.

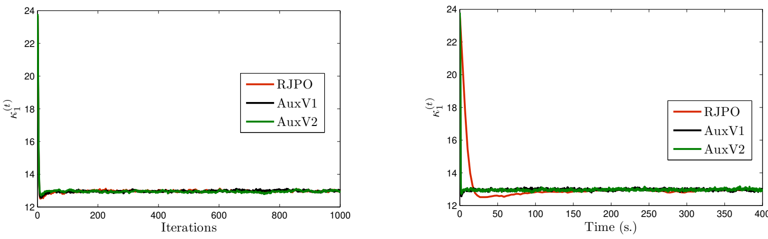


Figure 7: Chains of κ_1 versus iteration/time.

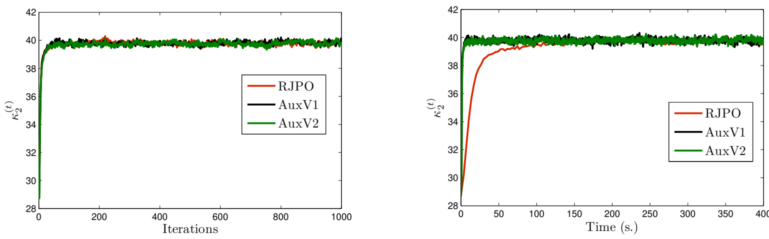


Figure 8: Chains of κ_2 versus iteration/time.

mixing properties of the chain is achieved by the proposed AuxV1 algorithm.

6 Conclusion

In this paper, we have proposed an approach for sampling from probability distributions in large scale problems. By adding some auxiliary variables to the model, we succeeded in addressing separately the different sources of correlations in the target posterior density. We have illustrated the usefulness of the proposed Gibbs sampling algorithms in two application examples. In the first application, we have proposed a wavelet-based Bayesian method to restore multichannel

Table 5: Mixing results for the different proposed algorithms. First row: Time per iteration. Second row: Estimates of the mean square jump in stationarity. Third row: Estimates of the mean square jump per second in stationarity. Fourth row: Relative efficiency to RJPO.

	RJPO	AuxV1	AuxV2
T(s.)	5.27	0.13	0.12
MSJ	15.41	14.83	4.84
MSJ/T	2.92	114.07	40.33
Efficiency	1	39	13.79

images degraded by blur and Gaussian noise. We have adopted a multivariate prior model that takes advantage of the cross-component correlation. Moreover, a separation strategy has been applied to construct prior models of the related prior hyperparameters. We have then employed the proposed Gibbs algorithm with auxiliary variables to derive optimal estimators for both the image and the unknown hyperparameters. In the new augmented space, the resulting model makes sampling much easier since the coefficients of the target image are no longer updated jointly but in a parallel manner. Experiments carried out on a set of multispectral satellite images have shown the good performance of the proposed approach with respect to standard algorithms. Several issues could be investigated as future work such as the ability of the proposed algorithm to deal with inter-scale dependencies in addition to the cross-channel ones. In the second application, we have applied the proposed method to the recovery of signals corrupted with mixed Gaussian noise. When compared to a state-of-the-art method for sampling from high dimensional scale Gaussian distributions, the proposed algorithms achieve a good tradeoff between the convergence speed and the mixing properties of the Markov chain even if the generated samples are not independent. Note that the proposed method can be applied to a wide class of applications in inverse problems, in particular, those including conditional Gaussian models either for the noise or the target signal.

References

[1] Bertero, M., Boccacci, P.: Introduction to inverse problems in imaging. CRC press (1998)

[2] Demoment, G.: Image reconstruction and restoration: Overview of common estimation structure and problems. IEEE Transactions on Acoustics, Speech and Signal Processing **37**(12), 2024–2036 (1989)

[3] Marnissi, Y., Zheng, Y., Chouzenoux, E., Pesquet, J.C.: A Variational Bayesian Approach for Image Restoration. Application to Image Deblurring with Poisson-Gaussian Noise. IEEE Transactions on Computational Imaging p. 16 (2017). DOI 10.1109/TCI.2017.2700203. URL `https://hal.archives-ouvertes.fr/hal-01613200`

[4] Chouzenoux, E., Jezierska, A., Pesquet, J.C., Talbot, H.: A Convex Approach for Image Restoration with Exact Poisson-Gaussian Likelihood. SIAM Journal on Imaging Sciences **8**(4), 2662–2682 (2015). URL `https://hal.archives-ouvertes.fr/hal-00922151`

[5] Chaari, L., Pesquet, J.C., Tourneret, J.Y., Ciuciu, P., Benazza-Benyahia, A.: A hierarchical Bayesian model for frame representation. IEEE Transactions on Signal Processing **58**(11), 5560–5571 (2010)

[6] Pustelnik, N., Benazza-Benhayia, A., Zheng, Y., Pesquet, J.C.: Wavelet-Based Image Deconvolution and Reconstruction, pp. 1–34. John Wiley & Sons, Inc. (1999)

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

[8] Liu, J.S.: Monte Carlo strategies in scientific computing. Springer Series in Statistics. Springer-Verlag, New-York (2001)

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

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

[11] Glynn, P.W., Iglehart, D.L.: Importance sampling for stochastic simulations. Management Science **35**(11), 1367–1392 (1989)

[12] Gilks, W.R., Wild, P.: Adaptive rejection sampling for Gibbs sampling. Applied Statistics **41**(2), 337–348 (1992)

[13] Neal, R.M.: MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo **2**, 113–162 (2011)

[14] Jarner, S.F., Hansen, E.: Geometric ergodicity of metropolis algorithms. Stochastic Processes and their Applications **85**(2), 341 – 361 (2000)

[15] Gilks, W.R., Best, N., Tan, K.: Adaptive rejection Metropolis sampling within Gibbs sampling. Applied Statistics **44**(4), 455–472 (1995)

[16] Dobigeon, N., Moussaoui, S., Coulon, M., Tourneret, J.Y., Hero, A.O.: Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery. IEEE Transactions on Signal Processing **57**(11), 4355–4368 (2009)

[17] Roberts, G.O., Gelman, A., Gilks, W.R.: Weak convergence and optimal scaling or random walk Metropolis algorithms. Annals of Applied Probability pp. 110–120 (1997)

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

[19] Roberts, G.O., Stramer, O.: Langevin diffusions and Metropolis-Hastings algorithms. Methodology and Computing in Applied Probability **4**(4), 337–357 (2002)

[20] Martin, J., Wilcox, C.L., Burstedde, C., Ghattas, O.: A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM Journal on Scientific Computing **34**(3), 1460–1487 (2012)

[21] Zhang, Y., Sutton, C.A.: Quasi-Newton methods for Markov chain Monte Carlo. In: Proc. Neural Information Processing Systems (NIPS 2011), 24, pp. 2393–2401. Granada, Spain (2011)

[22] Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. Journal of the Royal Statistical Society. Series B, Statistical Methodology **73**(91), 123–214 (2011)

[23] Van Dyk, D.A., Meng, X.L.: The art of data augmentation. Journal of Computational and Graphical Statistics (2012)

[24] Féron, O., Orieux, F., Giovannelli, J.F.: Gradient scan Gibbs sampler: an efficient algorithm for high-dimensional Gaussian distributions. IEEE Journal of Selected Topics in Signal Processing **10**(2), 343–352 (2016)

[25] Rue, H.: Fast sampling of Gaussian Markov random fields. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(2), 325–338 (2001)

[26] Geman, D., Yang, C.: Nonlinear image recovery with half-quadratic regularization. IEEE Transactions on Image Processing **4**(7), 932–946 (1995)

[27] Chellappa, R., Chatterjee, S.: Classification of textures using Gaussian Markov random fields. IEEE Transactions on Acoustic, Speech and Signal Processing **33**(4), 959–963 (1985)

[28] Rue, H., Held, L.: Gaussian Markov random fields: theory and applications. CRC Press (2005)

[29] Bardsley, J.M.: MCMC-based image reconstruction with uncertainty quantification. SIAM Journal on Scientific Computing **34**(3), A1316–A1332 (2012)

[30] Papandreou, G., Yuille, A.L.: Gaussian sampling by local perturbations. In: J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (eds.) Proc. Neural Inform. Process. Syst. Conf. (NIPS 2010), pp. 1858–1866. Vancouver, Canada (2010)

[31] Orieux, F., Féron, O., Giovannelli, J.F.: Sampling high-dimensional Gaussian distributions for general linear inverse problems. IEEE Signal Processing Letters **19**(5), 251–254 (2012)

[32] Gilavert, C., Moussaoui, S., Idier, J.: Efficient Gaussian sampling for solving large-scale inverse problems using MCMC. IEEE Transactions on Signal Processing **63**(1), 70–80 (2015)

[33] Parker A.and Fox, C.: Sampling Gaussian distributions in Krylov spaces with conjugate gradients. SIAM Journal on Scientific Computing **34**(3), B312–B334 (2012)

[34] Féron, O., Orieux, F., Giovannelli, J.F.: Gradient Scan Gibbs Sampler: an efficient algorithm for high-dimensional Gaussian distributions. IEEE Journal on Selected Topics in Signal Processing **10**(2), 343–352 (2016)

[35] Lasanen, S.: Non-Gaussian statistical inverse problems. Inverse Problems & Imaging **6**(2) (2012)

[36] Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning **4**(1), 1–106 (2012)

[37] Kamilov, U., Bostan, E., Unser, M.: Generalized total variation denoising via augmented Lagrangian cycle spinning with Haar wavelets. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2012), pp. 909–912. Kyoto, Japan (2012)

[38] Kolehmainen, V., Lassas, M., Niinimäki, K., Siltanen, S.: Sparsity-promoting Bayesian inversion. Inverse Problems **28**(2), 025,005 (2012)

[39] Stuart, M.A., Voss, J., Wiberg, P.: Conditional path sampling of SDEs and the Langevin MCMC method. Communications in Mathematical Sciences **2**(4), 685–697 (2004)

[40] Marnissi, Y., Chouzenoux, E., Benazza-Benyahia, A., Pesquet, J.C., Duval, L.: Reconstruction de signaux parcimonieux à l'aide d'un algorithme rapide d'échantillonnage stochastique. In: Actes du 25e colloque GRETSI. Lyon, France (2015). 4 p.

[41] Marnissi, Y., Benazza-Benyahia, A., Chouzenoux, E., Pesquet, J.C.: Majorize-Minimize adapted Metropolis-Hastings algorithm. Application to multichannel image recovery. In: Proceedings of the European Signal Processing Conference (EUSIPCO 2014), pp. 1332–1336. Lisbon, Portugal (2014)

[42] Vacar, C., Giovannelli, J.F., Berthoumieu, Y.: Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2011), pp. 3964–3967. Prague, Czech Republic (2011)

[43] Schreck, A., Fort, G., Le Corff, S., Moulines, E.: A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. IEEE Journal of Selected Topics in Signal Processing **10**(2), 366–375 (2016)

[44] Pereyra, M.: Proximal Markov chain Monte Carlo algorithms. Statistics and Computing **26**(4), 745–760 (2016)

[45] Atchadé, Y.F.: An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. Methodology and Computing in applied Probability **8**(2), 235–254 (2006)

[46] Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. Journal of the American statistical Association **82**(398), 528–540 (1987)

[47] Mira, A., Tierney, L.: 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`

[48] Robert C.and Casella, G.: Monte Carlo statistical methods. Springer Science & Business Media (2013)

[49] Doucet, A., Sénécal, S., Matsui, T.: Space alternating data augmentation: Application to finite mixture of gaussians and speaker recognition. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2005), vol. 4, pp. 708–713. IEEE, Philadelphia, PA (2005)

[50] Févotte, C., Cappé, O., Cemgil, A.T.: Efficient Markov chain Monte Carlo inference in composite models with space alternating data augmentation. In: Proceedings of the IEEE Statistical Signal Processing Workshop (SSP 2011), pp. 221–224. IEEE, Nice, France (2011)

[51] Giovannelli, J.F.: Unsupervised Bayesian convex deconvolution based on a field with an explicit partition function. IEEE Transactions on Image Processing **17**(1), 16–26 (2008)

[52] David, H.M.: Auxiliary variable methods for Markov Chain Monte Carlo with applications. J. Am. Stat. Assoc. **93**, 585–595 (1997)

[53] Hurn, M.: Difficulties in the use of auxiliary variables in Markov chain Monte Carlo methods. Stat. Comput. **7**, 35–44 (1997)

[54] Damlen, P., Wakefield, J., Walker, S.: Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. Journal of the Royal Statistical Society. Series B, Statistical Methodology **61**(2), 331–344 (1999)

[55] Duane, S., Kennedy, A., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. Physical Letters B **195**(2), 216 – 222 (1987)

[56] Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI**(6), 721–741 (1984)

[57] Idier, J.: Convex half-quadratic criteria and interacting auxiliary variables for image restoration. IEEE Transactions on Image Processing **10**(7), 1001–1009 (2001)

[58] Geman, D., Reynolds, G.: Constrained restoration and the recovery of discontinuities. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(3), 367–383 (1992)

[59] Champagnat, F., Idier, J.: A connection between half-quadratic criteria and EM algorithms. IEEE Signal Process. Lett. **11**(9), 709–712 (2004)

[60] Nikolova, M., Ng, M.K.: Analysis of half-quadratic minimization methods for signal and image recovery. SIAM Journal on Scientific Computing **27**, 937–966 (2005)

[61] Bect, J., Blanc-Féraud, L., Aubert, G., Chambolle, A.: A l1-unified variational framework for image restoration. In: Proc. European Conference on Computer Vision (ECCV 2004), pp. 1–13. Prague, Czech Republic (2004)

[62] Cavicchioli, R., Chaux, C., Blanc-Féraud, L., Zanni, L.: ML estimation of wavelet regularization hyperparameters in inverse problems. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP 2013), pp. 1553–1557. Vancouver, Canada (2013)

[63] Ciuciu, P.: Méthodes markoviennes en estimation spectrale non paramÃľtriques. application en imagerie radar doppler. Ph.D. thesis, Université Paris Sud-Paris XI (2000)

[64] Andrews, D.F., Mallows, C.L.: Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological) pp. 99–102 (1974)

[65] West, M.: On scale mixtures of normal distributions. Biometrika **74**(3), 646–648 (1987)

[66] Van Dyk, D.A., Park, T.: Partially collapsed gibbs samplers: Theory and methods. Journal of the American Statistical Association **103**(482), 790–796 (2008)

[67] Park, T., van Dyk, D.A.: Partially collapsed Gibbs samplers: Illustrations and applications. Journal of Computational and Graphical Statistics **18**(2), 283–305 (2009)

[68] Costa, F., Batatia, H., Oberlin, T., Tourneret, J.Y.: A partially collapsed Gibbs sampler with accelerated convergence for EEG source localization. In: Proceedings of the IEEE Statistical Signal Processing Workshop (SSP 2016), pp. 1–5. IEEE, Palma de Mallorca, Spain (2016)

[69] Kail, G., Tourneret, J.Y., Hlawatsch, F., Dobigeon, N.: Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method. IEEE Transactions on Signal Processing **60**(6), 2727–2743 (2012)

[70] Chouzenoux, E., Legendre, M., Moussaoui, S., Idier, J.: Fast constrained least squares spectral unmixing using primal-dual interior-point optimization. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(1), 59–69 (2012)

[71] Marnissi, Y., Benazza-Benyahia, A., Chouzenoux, E., Pesquet, J.C.: Generalized multivariate exponential power prior for wavelet-based multichannel image restoration. In: Proceedings of the IEEE International Conference on Image Processing (ICIP 2013), pp. 2402–2406 (2013)

[72] Laruelo, A., Chaari, L., Tourneret, J.Y., Batatia, H., Ken, S., Rowland, B., Ferrand, R., Laprie, A.: Spatio-spectral regularization to improve magnetic resonance spectroscopic imaging quantification. NMR in Biomedicine **29**(7), 918–931 (2016)

[73] Delp, E., Mitchell, O.: Image compression using block truncation coding. IEEE Transactions on Communications **27**(9), 1335–1342 (1979)

[74] Khelil-Cherif, N., Benazza-Benyahia, A.: Wavelet-based multivariate approach for multispectral image indexing. In: SPIE Conference on Wavelet Applications in Industrial Processing (2004)

[75] Chaux, C., Pesquet, J.C., Duval, L.: Noise covariance properties in dual-tree wavelet decompositions. IEEE Transactions on Information Theory **53**(12), 4680–4700 (2007)

[76] Roberts, G.O., Tweedie, L.R.: Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli **2**(4), 341–363 (1996)

[77] Murphy, K.P.: Conjugate Bayesian analysis of the Gaussian distribution. Tech. rep. (2007). `https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf`

[78] Barnard, J., McCulloch, R., Meng, X.L.: Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica pp. 1281–1311 (2000)

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

[80] Flandrin, P.: Wavelet analysis and synthesis of fractional brownian motion. IEEE Trans. Inf. Theory **38**(2), 910–917 (1992)

[81] Velayudhan, D., Paul, S.: Two-phase approach for recovering images corrupted by Gaussian-plus-impulse noise. In: Proceedings of the IEEE International Conference on Inventive Computation Technologies (ICICT 2016), vol. 2, pp. 1–7. IEEE, Tamilnadu, India (2016)

[82] Chang, E.S., Hung, C.C., Liu, W., Yina, J.: A denoising algorithm for remote sensing images with impulse noise. In: Proceedings on the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS 2016), pp. 2905–2908. IEEE, Beijing, China (2016)