Bayesian methods for inverse problems in signal and image processing

Yosra Marnissi

École Doctorale Mathématiques et Sciences et Technologies de l'Information et de la Communication

University of Paris-Est

THESIS

Specialty: Signal and Image Processing

presented by

Yosra Marnissi

25 April 2016, Paris

Bayesian methods for large scale inverse problems in signal and image processing

---

President	Gersende Fort	DR CNRS, Institute of Mathematics, Toulouse
Reviewers	Cédric Févotte	DR CNRS, Institute of Computer Science Research, Toulouse
	Philippe Ciuciu	DR CEA/NeuroSpin, INRIA Saclay
Examiners	Amel Benazza-Benyahia	Prof., University of Carthage, Tunis
	Nelly Pustelnik	CR CNRS, ENS, Lyon
PhD-advisors	Jean-Christophe Pesquet	Prof., CentraleSupélec, University of Paris-Saclay
	Emilie Chouzenoux	Ass. Prof., University of Paris-Est Marne-La-Vallée

"*Inside every nonBayesian there is a Bayesian struggling to get out.*"

— Dennis V. Lindley

Contents

Résumé

Les interactions entre les approches bayésiennes et variationnelles ont contribué au développement de méthodes qui combinent simulation stochastique (ou approximation), et optimisation afin de fournir des nouveaux algorithmes efficaces pour la restauration des signaux. Cette thèse se place dans cette direction prometteuse et a pour objectif le développement d'algorithmes bayésiens dont l'efficacité est améliorée par des outils issus de l'optimisation déterministe.

Dans plusieurs domaines d'application en traitement du signal et des images, on se retrouve le plus souvent confronté au problème suivant: on n'a pas accès directement aux paramètres d'intérêt mais seulement à des mesures indirectes. On doit alors résoudre un problème inverse dont le but est d'estimer les paramètres inconnus à partir de ces observations. Le chapitre 2 introduit le lecteur aux approches bayésiennes pour la résolution des problèmes inverses. Nous commençons par un bref aperçu des principaux problèmes inverses rencontrés en traitement de signal et les différentes méthodes qui permettent de les résoudre. Une attention particulière est accordée à la modélisation bayésienne. Cette dernière est basée sur la loi *a posteriori* qui utilise des informations *a priori* sur les paramètres inconnus à estimer ainsi que des informations sur les observations, pour construire des estimateurs. L'estimateur optimal au sens du coût quadratique est l'un des estimateurs les plus couramment employés. Toutefois, un défi majeur dans de telles méthodes est le calcul de la loi *a posteriori* ou plus précisément son exploration. En effet, dans de nombreux domaines du traitement d'image tels que la médecine, l'astronomie et la microscopie, il est assez courant de traiter des données de grandes tailles avec des modèles de plus en plus sophistiqués. Dans ces circonstances, même si la loi *a priori* et celle des observations sont simples, la loi *a posteriori* est généralement compliquée dans le sens où elle ne peut être connue qu'à une constante multiplicative près et/ou elle possède une forme non usuelle ou son traitement nécessite des ressources informatiques massives. A cet égard, plusieurs méthodes d'approximation ont été proposées. D'une part, les algorithmes d'échantillonnage de Monte Carlo par chaînes de Markov (MCMC) sont des outils bayésiens puissants pour explorer des lois compliquées. Dans le but d'échantillonner suivant une certaine loi de probabilité, l'idée sous-jacente aux algorithmes MCMC est de produire, à partir d'une certaine loi de transition donnée, une chaîne de Markov qui converge vers la loi cible. Une classe importante de méthodes MCMC est

inspirée de l'algorithme Metropolis-Hastings (MH) où la loi de transition est définie par une loi de proposition dont l'échantillonnage est simple et une règle d'acceptation-rejet assurant la réversibilité de la chaîne. Une tâche difficile lors de la mise en oeuvre de cette méthode est le choix de la densité de proposition. Elle doit idéalement fournir une bonne approximation locale de la densité cible avec un faible coût de calcul. On peut d'autre part recourir à des approches basées sur des approximations bayésiennes variationnelles (VBA). Au lieu de simuler à partir de la vraie loi *a posteriori*, les approches VBA visent à l'approcher avec une autre loi plus simple à partir de laquelle la moyenne *a posteriori* peut être facilement calculée. Ces méthodes peuvent généralement conduire à une complexité de calcul relativement faible par rapport aux algorithmes basés sur l'échantillonnage.

Nos contributions s'orientent selon deux axes principaux: Dans une première partie de la thèse, nous proposons des algorithmes de simulation MCMC et nous fournissons des solutions pour contourner leurs limitations dans les problèmes de grande dimension (chapitres 3 et 4). Dans une seconde partie, nous proposons des approches VBA et, en particulier un algorithme pour la restauration de signaux en présence d'un bruit non gaussien (chapitre 5). Dans chaque chapitre, nos contributions résident dans la conception de nouveaux algorithmes et en la proposition de nouvelles solutions à des problèmes applicatifs issus de la restauration de signaux/images.

Dans le chapitre 3, nous nous intéressons aux algorithmes de MH basés sur la diffusion de Langevin définie pour les lois différentiables. Inspirées des outils d'optimisation de type descente de gradient, ces méthodes exploitent des informations sur la géométrie de la loi cible pour construire la densité de proposition dans le but de guider la chaîne vers l'espace cible où la plupart des échantillons doivent être concentrés. Pour ce faire, la composante directionnelle de la loi de proposition est choisie comme une itération d'un algorithme de descente de gradient préconditionné. Le bruit est ensuite injecté dans cette mise à jour de telle sorte que la trajectoire de la chaîne explore toute la distribution *a posteriori* plutôt que de converger vers un mode. Ainsi, un nouvel échantillon de cette loi de proposition est plus probablement accepté, ce qui tend à accélérer la convergence de la chaîne vers la loi stationnaire. Toutefois, l'échantillonnage devient difficile quand la dimension du problème augmente. Ces difficultés sont principalement dues au coût élevé de chaque itération et aux mauvaises propriétés de mélange de la chaîne lorsque la matrice de préconditionnement est mal choisie. Notons que des problèmes similaires sont aussi rencontrés dans les algorithmes d'optimisation de descente de gradient préconditionnés. Dans ce travail de thèse, nous exploitons les connexions entre les méthodes déterministes et les approches bayésiennes stochastiques pour accélérer les algorithmes d'échantillonnage de type MH. En s'inspirant des approches de Majoration-Minimisation, nous développons

un algorithme de Langevin MH préconditionné par une matrice adaptative construite à chaque itération à partir d'une fonction tangente majorante quadratique de l'opposé du logarithme de la densité *a posteriori*. Nous proposons différentes variantes de fonctions tangentes majorantes quadratiques construites avec des matrices de courbure pleines, constantes ou diagonales qui permettent d'adapter l'algorithme proposé à la grande dimension du problème. Nous démontrons ensuite l'ergodicité géométrique de l'algorithme d'échantillonnage proposé en nous basant sur des résultats théoriques concernant les algorithmes de MH classiques. L'algorithme conçu est enfin validé sur un problème de déconvolution de signal parcimonieux en adoptant une loi *a priori* de Cauchy. Ce test nous permet d'étudier l'impact de la matrice de préconditionnement sur la performance de l'algorithme. Les résultats expérimentaux confirment la rapidité de cette nouvelle approche par rapport à l'échantillonneur usuel de Langevin.

Dans les problèmes de grande taille, la performance des algorithmes d'échantillonnage stochastiques est très sensible aux dépendances entre les paramètres. Par exemple, ce problème se pose lorsqu'on cherche à échantillonner selon une loi gaussienne de grande dimension dont la matrice de covariance ne présente pas de structure simple, (i.e., ni parcimonieuse, ni circulante, ni Toeplitz, etc.) Dans ce contexte, il est usuel de recourir à des méthodes de simulation basées sur le principe de Perturbation-Optimisation, qui nécessitent de résoudre à chaque itération un problème de minimisation avec un algorithme itératif, ce qui peut rendre le processus d'échantillonnage prohibitif surtout lorsqu'il est intégré dans un échantillonneur de Gibbs. Un autre défi est la sélection de bonnes lois de propositions MH qui utilisent des informations sur la géométrie locale de la densité cible afin d'accélérer la convergence et améliorer les propriétés de mélange dans l'espace des paramètres, sans être trop coûteuses en termes de calcul. Ces deux problèmes sont principalement liés à la présence de deux sources de dépendances hétérogènes provenant soit du terme d'attache aux données, soit de la loi *a priori*, dans le sens où les matrices de covariance associées ne peuvent pas être diagonalisées dans le même domaine. Pour pallier ces difficultés, nous proposons d'ajouter des variables auxiliaires au modèle dans le but de dissocier les deux sources de dépendances. Dans le nouvel espace augmenté, une seule source de corrélation reste directement liée aux paramètres cibles, les autres sources de corrélations n'interviennent que par le biais des variables auxiliaires. Notons que cette stratégie est étroitement liée aux approches semi-quadratiques très souvent utilisées dans les problèmes d'optimisation déterministes [Allain et al., 2006]. Dans le chapitre 4, nous commençons par proposer une stratégie pour ajouter des variables auxiliaires dans le cas d'une loi gaussienne puis, nous l'étendons aux mélanges de gaussiennes. Une nouvelle étape est ajoutée dans l'algorithme de Gibbs pour tirer des échantillons de la loi associée aux variables auxiliaires. Nous considérons plusieurs stratégies pour effectuer

cette tâche simplement en utilisant les propriétés de la matrice de covariance liée aux variables auxiliaires. Nous étudions tout d'abord l'efficacité de l'approche proposée sur un problème de déconvolution d'images multi-composantes où l'on dispose de plusieurs observations de la même scène, acquises dans différentes bandes spectrales, qui sont dégradées par un flou et un bruit additif gaussien de variance connue. Le problème est abordé dans le domaine des ondelettes. Une loi *a priori* multivariée est adoptée pour modéliser le vecteur des coefficients d'ondelettes situés à la même position spatiale, à travers toutes les bandes spectrales, afin d'exploiter l'intercorrélation spectrale. Une stratégie de séparation est proposée pour estimer les hyperparamètres impliqués dans la loi *a priori* à partir de l'observation dégradée de l'image à reconstruire. En ajoutant des variables auxiliaires dans le terme d'attache aux données, la matrice d'observation n'est plus directement liée à l'image. Grâce à la propriété de séparation de la loi *a priori* (coefficients d'ondelettes situés dans différentes positions spatiales, orientations ou échelles, supposés indépendants), les vecteurs des coefficients d'ondelettes appartenant à différentes sous-bandes d'ondelettes peuvent être échantillonnés indépendamment en parallèle. Dans le cas d'un grand nombre de composantes spectrales, les performances de notre méthode peuvent être améliorées en utilisant une architecture parallèle multi-coeurs. Nous montrons aussi les bonnes performances de l'approche proposée pour résoudre le problème d'échantillonnage de lois gaussiennes de grande dimension à travers un exemple de restauration d'images dégradées par un mélange de bruits gaussiens. Comme la loi *a posteriori* est gaussienne, les variables auxiliaires peuvent être ajoutées soit à la *loi a priori* soit au terme d'attache aux données, soit au deux en même temps, en fonction des propriétés des matrices de covariance qui leur sont associées. Les résultats expérimentaux montrent la capacité de notre méthode à surmonter les difficultés liées à la présence de corrélations hétérogènes entre les coefficients du signal.

Dans le chapitre 5, nous nous intéressons aux méthodes VBA. Le but de ces méthodes est d'approcher la loi *a posteriori* d'un ensemble de paramètres inconnus par une nouvelle loi séparable plus simple, qui est aussi proche que possible de la vraie loi au sens de la divergence de Kullback-Leibler. Dans de nombreux cas, les expressions analytiques de ces lois approchantes sont inextricables, surtout lorsqu'il n'est pas possible de calculer directement l'espérance du logarithme de la loi jointe par rapport à la loi approchante. C'est le cas par exemple dans les problèmes impliquant des modèles de bruit non gaussiens et dépendants du signal. En recourant à des stratégies de majoration basées sur des approches semi-quadratiques, nous construisons une borne inférieure pour la divergence de Kullback-Leibler que nous souhaitons minimiser. Les lois approchantes sont alors obtenues en minimisant cette borne. En particulier, la loi approchante de l'image est réduite à une loi gaussienne, dont la matrice de covariance peut être approchée soit par une

matrice diagonale soit par un estimateur de Monte Carlo basé sur un algorithme de simulation stochastique. Nous illustrons les performances de notre algorithme sur des exemples de restauration d'images dégradées par un bruit mixte Poisson-gaussien. Les résultats expérimentaux confirment la bonne performance de la méthode proposée par rapport à des méthodes supervisées de l'état de l'art.

Dans le chapitre 6, nous résumons nos principales contributions et nous proposons plusieurs pistes pour de futurs travaux à court, moyen et long termes.

Abstract

Bayesian approaches are widely used in signal processing applications. In order to derive plausible estimates of original parameters from their distorted observations, they rely on the posterior distribution that incorporates prior knowledge about the unknown parameters as well as informations about the observations. The posterior mean estimator is one of the most commonly used inference rule. However, as the exact posterior distribution is very often intractable, one has to resort to some Bayesian approximation tools to approximate it. In this work, we are mainly interested in two particular Bayesian methods, namely Markov Chain Monte Carlo (MCMC) sampling algorithms and Variational Bayes approximations (VBA).

This thesis is made of two parts. The first one is dedicated to sampling algorithms. First, a special attention is devoted to the improvement of MCMC methods based on the discretization of the Langevin diffusion. We propose a novel method for tuning the directional component of such algorithms using a Majorization-Minimization strategy with guaranteed convergence properties. Experimental results on the restoration of a sparse signal confirm the performance of this new approach compared with the standard Langevin sampler. Second, a new sampling algorithm based on a Data Augmentation strategy, is proposed to improve the convergence speed and the mixing properties of standard MCMC sampling algorithms. Our methodological contributions are validated on various applications in image processing showing the great potentiality of the proposed method to manage problems with heterogeneous correlations between the signal coefficients.

In the second part, we propose to resort to VBA techniques to build a fast estimation algorithm for restoring signals corrupted with non-Gaussian noise. In order to circumvent the difficulties raised by the intricate form of the true posterior distribution, a majorization technique is employed to approximate either the data fidelity term or the prior density. Thanks to its flexibility, the proposed approach can be applied to a broad range of data fidelity terms allowing us to estimate the target signal jointly with the associated regularization parameter. Illustration of this approach through examples of image deconvolution in the presence of mixed Poisson-Gaussian noise, show the good performance of the proposed algorithm compared with state of the art supervised methods.

Notations

Linear algebra

$\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}^*$: set of real numbers, positive real numbers and non-zero real numbers respectively.

$\mathbb{N}$: set of positive integers.

$\mathbb{R}^{m\times n}$: set of matrices of m rows and n columns.

s : scalars will be denoted by lowercase letters.

$\mathbf{v}$: vectors will be denoted by lowercase bold letters.

$\mathbf{M}$: matrices will be denoted by uppercase bold letters.

$\mathbf{M}^\top$: transposition of the matrix $\mathbf{M}$.

$\mathbf{M}^{-1}$: inverse of the matrix $\mathbf{M}$.

$\mathbf{I}_m$: square identity matrix in $\mathbb{R}^m$.

$\mathbf{1}_m$, $\mathbf{0}_m$: vectors of $\mathbb{R}^m$ with all entries equal to 1 and 0 respectively.

v_i : i^{th} coefficient of the vector $\mathbf{v}$.

$M_{i,j}$: element in the i^{th} row and j^{th} column of the matrix $\mathbf{M}$.

$|s|$: absolute value of the scalar s.

$|\mathbf{M}|$: determinant of the matrix $\mathbf{M}$.

$\|\mathbf{v}\|$: ℓ_2 norm of the vector $\mathbf{v}$.

$\|\mathbf{v}\|_\mathbf{M}$: weighted norm equals to $\sqrt{\mathbf{v}^\top\mathbf{M}\mathbf{v}}$.

$\|\mathbf{M}\|$: ℓ_2 norm of the matrix $\mathbf{M}$.

diag(M) : vector constructed from the diagonal elements of the matrix $\mathbf{M}$.

Diag(v) : diagonal matrix whose elements are given by $\mathbf{v}$.

Spec(M) : set of eigenvalues of $\mathbf{M}$.

Ker(M) : kernel of the matrix $\mathbf{M}$.

$\succcurlyeq$: partial order on the set of Hermitian matrices, we write $\mathbf{A} \succcurlyeq \mathbf{B}$ iff $\mathbf{A}-\mathbf{B}$ is positive semidefinite.

$\nabla f(\mathbf{v})$: gradient vector of the function f with respect to $\mathbf{v}$.

$\nabla^2 f(\mathbf{v})$: Hessian matrix of the function f with respect to $\mathbf{v}$.

$\frac{\partial \mathbf{f}(\mathbf{v})}{\partial v_i}$: partial derivatives of the function f with respect to v_i.

log : natural logarithm.

$\mathcal{KL}(f||g)$: Kullback-Leibler divergence between f and g.

$1_{[a,b]}$: indicator function in interval $[a, b]$.

δ_a : discrete measure concentrated at the value a.

Probability calculus

$\mathbf{x}^{(t)}$: the state of $\mathbf{x}$ at iteration t.

$\mathsf{P}(E)$: probability of event E.

$\mathcal{F}_{\cdot}$: probability distributions will be denoted by italic uppercase letters. For example, $\mathcal{N}$, $\mathcal{ST}$, $\mathcal{GMEP}$, $\mathcal{G}, \mathcal{U}$ will denote respectively Gaussian, Student't, Generalized multivariate exponential power, Gamma and uniform distributions.

$\mathsf{f}(.)$: probability density functions will be denoted by straight lowercase letters. For example, n will denote the density function of a Gaussian distribution.

$\mathsf{p}(\mathbf{x})$: prior probability density function of $\mathbf{x}$.

$\mathsf{p}(\mathbf{x}|\mathbf{z})$: conditional probability density function of $\mathbf{x}$ given $\mathbf{z}$.

$\mathcal{P}_{\mathbf{x}}$: prior probability of $\mathbf{x}$.

$\mathcal{P}_{\mathbf{x}|\mathbf{z}}$: conditional probability of $\mathbf{x}$ given $\mathbf{z}$.

$\mathsf{E}_{\mathbf{x}}(\xi(\mathbf{x}))$: expectation of the function $\xi(\mathbf{x})$ with respect to $\mathcal{P}_{\mathbf{x}}$.

$\mathsf{E}_{\mathbf{x}|\mathbf{z}}(\xi(\mathbf{x}))$: expectation of the function $\xi(\mathbf{x})$ with respect to $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$.

$\mathcal{J}(\mathbf{x})$: minus logarithm of $\mathsf{p}(\mathbf{x}|\mathbf{z})$.

$\mathcal{J}(\mathbf{x}|\mathbf{y})$: minus logarithm of $\mathsf{p}(\mathbf{x}|\mathbf{y}, \mathbf{z})$.

List of acronyms

MAP : Maximum a posteriori.

MMSE : Minimum mean squared error.

MCMC : Markov Chain Monte Carlo.

MH : Metropolis Hastings.

RW : Random Walk.

MALA : Metropolis adapted Langevin algorithm.

MM : Majorize-Minimize.

3MH : Majorize-Minimize adapted Metropolis Hastings.

VBA : Variational Bayesian approximation.

PG : Poisson Gaussian.

TV : Total variation.

SLTV : Semi-local total variation.

NLTV : Non-local total variation.

GAST : Generalized Anscombe transform.

SPoiss : Shifted Poisson.

WL2 : Weighted least squared.

HMC : Hamiltonian Monte Carlo.

ACT : Integrated autocorrelation time.

ESS : Effective sample size.

MSJ : Mean square jump.

MCI : Multichannel image.

SNR : Signal-to-noise ratio.

SSIM : Structural similarity.

BSNR : Blurred signal-to-noise ratio.

- Chapter 1 -

General introduction

§ 1 Motivation

There has been an increasing need for building unsupervised methods where all parameters of interest are automatically estimated on the fly avoiding their manual tuning by the designer. In that respect, the Bayesian framework is becoming an increasingly popular approach to perform such tasks. Contrary to classical approaches, Bayesian modeling considers parameters of interest as random variables rather than deterministic quantities. Hence, this approach requires to specify a prior distribution that describes what is known about those parameters before data are observed. Estimates are then computed by reasoning on the resulting law that takes into account these prior probabilities combined with information about observations via Baye's formula. The aforementioned law is the so-called "Posterior distribution" which constitutes the core of the Bayesian framework.

The use of Bayesian methods can be motivated from many different viewpoints. First, the Bayesian approach provides an elegant framework for modeling uncertainty about the data and the target parameters via probabilities. Second, unlike deterministic approaches that propose point estimators, the outcome of Bayesian methods is the entire posterior distribution that describes the dispersion of the unknown parameters in the space given the observed data. The posterior distribution can be useful to make decisions, as well as to deduce point wise estimates of some statistics including the mean, variance, higher order moments, quantiles and also to capture uncertainty about the target parameters via confidence intervals. Third, many problems such as those involving hierarchical models [Damlen et al., 1999] are often much easier to handle within a Bayesian framework than with classical methods.

§ 2 Challenges

A major challenge in Bayesian methods is the calculation of the posterior distribution or more precisely, its exploration. In fact, we have reached a

time when it is common in many fields such as medicine, astronomy and microscopy, to handle large amounts of data with increasingly sophisticated models. In these challenging settings, even if the prior and the observation model are simple, the posterior law is almost always intractable in the sense that it can only be known up to a multiplicative constant and/or has a complicated form or requires massive computing resources to be handled. Regarding the difficulty in directly dealing with the posterior distribution, many methods have been proposed. This thesis is devoted to two particular families of methods namely Markov chain Monte Carlo simulation based techniques and Variational Bayesian methods and their application to large scale signal processing problems.

2.1 MCMC simulation methods

Markov chain Monte Carlo (MCMC) methods are stochastic simulation methods that allow to approximate a given target distribution such as the posterior law, by relying on Markov chain theory and Monte Carlo integration. They proceed in two main steps. First, a Markov chain is built with a given transition rule such that its stationary states follow the posterior law [Hastings, 1970; Gamerman and Lopes, 2006]. Once the Markov chain has reached its stationary distribution, Monte Carlo approximation is used to infer the posterior characteristics. Because of the lack of knowledge about the posterior distribution, the Markov chain often starts at a random point far from the target high density regions. Then, if the MCMC algorithm is not run a sufficiently long time, the resulting estimators are likely to be, highly biased leading to unreliable inference and poor forecasts. Actually, MCMC algorithms are only useful if the two following requirements are satisfied. First, the Markov chain should converge to its stationary distribution, namely the posterior law, in a finite time. This motivates the introduction of a burn-in period also known as the transient phase, which should be long enough to guarantee that the Markov chain has reached its stationary states without being too long to avoid throwing away useful samples. Second, a sufficient number of samples should be available in order to obtain accurate estimates of the posterior law. The required number of states depends on how well the MCMC algorithm is mixing the stationary space at convergence. Hence, in the case of large scale problems, we are faced with two main challenges:

- The first challenge concerns the design of a suitable MCMC algorithm, i.e., the choice of an appropriate transition kernel, with a low computational burden and limited memory requirements, such that the Markov chain reaches stability in a finite acceptable time.

- In order to produce consistent estimates in a finite time, the Markov chain should have good mixing properties in convergence. Otherwise, the number of samples required for accurate estimates can be infeasibly

large. The mixing behavior of the chain is generally related to how well the transition rule of the Markov chain allows to approach globally or locally the curvature of the target distribution.

In this thesis, a special attention is paid to Metropolis Hastings algorithms whose transition kernel are defined based on a given proposal density and a rejection-acceptance rule guaranteeing that the stationary distribution of the Markov chain is the one of interest [Metropolis et al., 1953; Hastings, 1970]. Hence, the problem of setting the transition kernel is equivalent to the problem of choosing a proposal density that allows large transitions over the parameter space with low computational cost. In particular, we are interested in adaptive algorithms that permit to self-adjust the proposal at each iteration based on the information provided by earlier samples.

2.2 Variational Bayesian approximation methods

Variational Bayesian approximation (VBA) provides an alternative tool to estimate complicated target distributions via deterministic approximations [Šmídl and Quinn, 2005]. Unlike stochastic simulation techniques, the goal of VBA approaches is to seek for an additional distribution called the approximating distribution that should well approximate the target law while being simpler to compute. To make this approximation as close as possible to the true one, we select, over the space of possible solutions, the one which achieves the minimal distance to the target distribution. More specifically, this distance is measured in terms of Kullback-Leibler (KL) divergence. To make this minimization tractable, additional assumptions about the approximating distribution such as independence between parameters are further introduced. The minimization task is then performed by an iterative coordinate-ascent algorithm: updating each component of the factorized approximating distribution while holding the others fixed. The advantage of the VBA algorithm is that, instead of giving point estimates, it provides an approximation of the entire posterior distribution while achieving almost the same computational and implementation complexity as classical methods. It is worth noting that, each step of the iterative algorithm ideally requires the ability to integrate a sum of terms in the log joint likelihood using a factorized approximating distribution which is the case for example for conjugate exponential models. However, very often, these integrals do not have closed forms especially when the observations statistic are far from the additive Gaussian models and no longer independent from the target parameters.

§ 3 Main contributions

The contribution of this thesis is developed along two main lines. First, we contribute to MCMC simulation algorithms and we provide solutions to

overcome their limitations in high dimensional problems (Chapters 3 and 4). Second, we contribute to Variational Bayesian approximation approaches and we propose a general methodology for dealing with signal recovery in the presence of non-Gaussian noise (Chapter 5).

More specifically, our contributions in each part can be classified into the design of new algorithms and the proposition of novel solutions to applicative problems. This is explained in more details below:

MCMC simulation methods

Proposed algorithms:

- *Majorize-Minimize Metropolis Hastings algorithm: (Chapter 3)* This algorithm is a special case of Metropolis Hastings algorithms where the proposal density is adjusted by resorting to some deterministic optimization tools.
  - We derive the proposal density from the Langevin stochastic differentiable equation. The later offers a tool for constructing Metropolis Hastings proposal densities that incorporate first order derivatives information of the target law to guide the chain toward the target space where most of the samples should be concentrated. More precisely, the directional component of the proposed algorithm is chosen to be one iteration of a preconditioned gradient descent algorithm where the preconditioning matrix is constructed using a Majorize-Minimize strategy. This allows us to push the Markov chain in each iteration from the current state toward regions with high probabilities values.
  - We propose different variants of the curvature metrics that may be used as alternatives to replace costly matrices in large scale problems.
  - We demonstrate the geometric ergodicity of the proposed algorithm using some theoretical works concerning the convergence properties of Metropolis Hastings algorithms.
- *Gibbs sampling algorithm with auxiliary variables: (Chapter 4)* This chapter addresses the problem of sampling from multivariable distributions when the variables of interest are strongly correlated with heterogeneous sources of dependencies preventing the construction of efficient proposals in Metropolis Hastings algorithms. In particular, we add additional variables to the model without changing the marginal laws in such a way that simulation can be performed more efficiently in the new larger space.

- We investigate the case of Gaussian models involving high dimensional covariance matrices.
- We extend the proposed approach to scale mixture of Gaussian models.
- We demonstrate the efficiency of the proposed algorithm to sampling from high dimensional Gaussian distributions.
- We propose some strategies allowing to sample the auxiliary variables with low computational cost.

Applications to multichannel image recovery:
In multichannel imaging, several observations of the same scene acquired in different spectral ranges are available. Very often, the spectral components are degraded by a blur and an additive noise. In this thesis, we address the problem of recovering the image components in a wavelet domain by adopting a Bayesian approach. Our contribution is threefold.

- We take advantage of the inter-channels dependencies by jointly processing the spectral components. Hence, we adopt a multivariate prior model for the distribution of vectors composed of wavelet coefficients at the same spatial position in a given subband through all the channels. The proposed model is flexible enough as it allows us to consider various levels of sparsity.
- We propose a methodology for estimating the prior scale matrices within a Gibbs sampling algorithm by resorting to a separation strategy.
- We address the challenging issue of computing estimates of the wavelet coefficients jointly with prior scale matrices using the Gibbs sampling algorithm with auxiliary variables. The sampling step of the wavelet coefficients is performed using the Majorize-Minimize Metropolis Hastings algorithm proposed in Chapter 3. The proposed algorithm is able to exploit the potentials offered by multicore parallel system architectures for a large number of channels.

Variational Bayesian approximation methods

Proposed algorithm:

- *VBA algorithm for image restoration: (Chapter 5)* We address the problem of image recovery beyond additive Gaussian noise models. Unlike classical approaches often adopted in the literature, the regularization parameter is estimated throughout a VBA algorithm from the observations. In order to address the problem of the intricate form

of either the observation model or the prior distribution, we resort to majorization techniques to construct a lower bound on the Kulback divergence that we want to minimize. This bound takes additional variables which determine how tightly it approximates the true distance. The approximating distribution are then derived by optimizing this bound using an iterative scheme. Note that the proposed algorithm is flexible as it can be applied to a wide range of prior distributions and data fidelity terms.

Applications to Poisson Gaussin image recovery:
We demonstrate the potential of the proposed VBA method in the context of image recovery under Poisson Gaussian noise. Several tests are made using various approximations of the exact Poisson Gaussian data fidelity term as well as different penalization strategies. Results show that the proposed approach is efficient and achieves performance comparable with other methods where the regularization parameter is manually tuned from the ground truth.

§ 4 Publications

Journal papers

- Y. Marnissi, Y. Zheng, E. Chouzenoux and J.-C. Pesquet. A Variational Bayesian Approach for Image Restoration. Application to Image Deblurring with Poisson-Gaussian Noise. *To appear in IEEE Transactions on Computational imaging.*

Conference papers:

- Y. Marnissi, E. Chouzenoux, J.-C. Pesquet and A. Benazza-Benyahiya. An Auxiliary Variable Method for Langevin based MCMC algorithms. *In Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP 2016)* pages 297-301, Palma of Mallorca, Spain, 26-29 June 2016.

- Y. Marnissi, Y. Zheng and J.-C. Pesquet. Fast Variational Bayesian signal recovery in the presence of Poisson-Gaussian noise. *In Proceedings of the IEEE International Conference on Acoustics, Speech ans Signal Processing (ICASSP 2016)*, pages 5, Shanghai, China, 20-25 Mar. 2016.

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

- Y. Marnissi, A. Benazza-Benyahia, E. Chouzenoux and J.-C. Pesquet. Majorize-Minimize adapted Metropolis Hastings algorithm. Application to multichannel image recovery. *In Proceedings of the European Signal Processing Conference (EUSIPCO 2014)*, pages 1332-1336, Lisbon, Portugal, 1-5 Sep. 2014.
- M. A. Moussa,Y. Marnissi and Y. Ghamri Doudane. A Primal-Dual algorithm for data gathering based on matrix completion for wireless sensor networks . *In Proceedings of the IEEE International Conference on Communications (ICC 2016)*, pages 5, Kuala Lumpur, Malaysia, 23-27 May 2016.
- M. A. Moussa,Y. Marnissi and Y. Ghamri Doudane. Matrix Completion with Convex Constraints for Data Gathering in Wireless Sensor Networks . *In Proceedings of the IEEE Global Communications Conference, (GLOBECOM 2016)*, pages 6, Washington, DC USA, 4-8 Dec. 2016.

Talks

- Schéma d'échantillonnage Metropolis-Hastings adapté par Majoration-Minimisation. Journée GDR ISIS *Sur le filtrage Bayésien en grande dimension par méthodes de Monte Carlo.* Paris, France, 21 May. 2014.
- Majorize-Minimize adapted Metropolis-Hastings algorithm. Sustain Image Processing SSOIP Workshop *On high-dimensional stochastic simulation and optimisation in image processing.* Bristol, England, 27-29 Aug. 2014.

§ 5 Outline

As stated earlier, the aim of this thesis is to design new approaches to overcome the potential limitations of standard Bayesian methods, in order to approximate complicated posterior distributions in high dimensional inverse problems. The forthcoming chapters are organized as follows:

In the background chapter, we will begin with an introduction to the topic of inverse problems. We present the Bayesian framework and the main related algorithms.

In Chapter 3, we propose a new sampling algorithm which can be seen as a preconditioned version of the standard Metropolis Hastings adapted Langevin sampling algorithm using an adaptive matrix based on a Majorize-Minimize strategy. A particular attention is paid to its convergence properties namely its geometric ergodicity in case of super-exponential distributions. The proposed algorithm is also validated over an illustrative example for sparse seismic signal deconvolution with a Student't prior distribution.

In Chapter 4, we propose a second approach for Bayesian sampling in large scale problems using data augmentation type strategies allowing to address the problem of standard sampling algorithms when applied to models involving highly correlated variables. This idea is first discussed in the case of Gaussian models and then extended to scale mixture of Gaussian ones. The proposed algorithms are applied to the deconvolution of multispectral images affected by blur and additive Gaussian noise. Furthermore, we show the performance of the proposed approach for sampling from large scale Gaussian distributions with an application to image recovery under two-terms mixed Gaussian noise.

In Chapter 5, we focus on signal recovery beyond the standard additive independent Gaussian noise model by resorting to VBA approaches. The proposed method allows us to jointly estimate the original signal and the required regularization parameter from the observed data by providing good approximations of the MMSE estimators for the problem of interest. We provide simulations results together with comparisons with state-of-the-art methods in terms of image restoration performance and computation time in the case of mixed Poisson Gaussian noise often encountered in microscopy and astronomy imaging systems.

Finally, we draw some conclusions and perspectives in Chapter 6.

- Chapter 2 -

Background

In a wide range of real applications, we do not have access to the signal of interest by direct measurements but only to a distorted version of this signal. These distortions may arise due to various phenomena which are often unavoidable in practical situations. They may depend on the physics of the studied phenomenon, on the process of signal formation, on the employed acquisition system and also on the communication channel. In inverse problems, one tries to get some useful information from the observed data about either the unknown signal itself or the value of some other unknown parameters (e.g., identification of the distortion parameters). Among well known fields of applications of inverse problems, we can mention medical imaging (MRI, CT scans, ultrasound imaging) [Fitzpatrick and Sonka, 2000], image processing [Benvenuto et al., 2008; Murtagh et al., 1995; Oliveira et al., 2009], geophysics [Parker, 1994], spectroscopy [Provencher, 1979], machine learning [Bach et al., 2012], video processing [Pizurica et al., 2004], astronomy [Lantéri and Theys, 2005], microscopy [Dupé et al., 2009] etc.

Several researches have been focused on this topic and proposed some methodologies to solve inverse problems. In this chapter, we review some of them that are relevant to our studies. First, we provide the reader with an introduction to inverse problems and to the associated mathematical modeling. Second, we present the Bayesian principle applied to inverse problems. Finally, the related algorithms are presented while a special attention is devoted to stochastic simulation and approximation algorithms.

§ 1 Inverse problems

The resolution of an inverse problem is the process of estimating an unknown signal from measurements based on the direct model linking the target signal to the observed one. However, the perfect direct model is generally not satisfied due to some random parasite signals that, once incorporated into the direct model, alter the extraction of useful information. In this work, we consider the following observation model:

$$\mathbf{z} = \mathcal{D}(\mathbf{H}\overline{\mathbf{x}}) \tag{2.1}$$

where $\overline{\mathbf{x}} \in \mathbb{R}^Q$ denotes the target signal, $\mathbf{z} \in \mathbb{R}^N$ is the measured data, $\mathbf{H} \in \mathbb{R}^{N \times Q}$ is the observation matrix describing the linear degradation and $\mathcal{D}$ is the model expressing nonlinear degradation and measurement errors called noise. Such model arises in several signal processing applications (deblurring, denoising, super resolution, reconstruction, segmentation, compressive sensing, inpainting) with appropriate definitions of the operator $\mathbf{H}$ and the model $\mathcal{D}$.

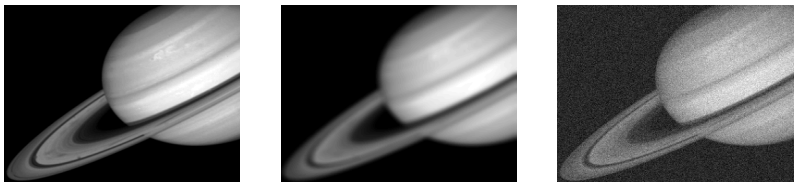
For many applications such as optical remote sensing imaging and microscopy, acquisition sensors generally suffer from internal fluctuations referred to an electrical or thermal noise. This type of noise is additive, independent of the signal of interest, and it can be modeled by a Gaussian distribution. Moreover, signals can be affected by blur resulting from a motion between the scene and the camera, a defocus of an optical imaging system, sensor imperfections and atmospheric conditions. Thereby, the observation model (2.1) reduces to the following linear additive noise model:

$$\mathbf{z} = \mathbf{H}\overline{\mathbf{x}} + \mathbf{w} \tag{2.2}$$

where $\mathbf{w}$ is an additive zero mean Gaussian noise and $\mathbf{H}$ is the blurring operator. In such case, we say that we have to solve a restoration problem. When $\mathbf{H} = \mathbf{I}_Q$, it reduces to a denoising problem. Note that, in many applications, the matrix $\mathbf{H}$ may not express a blur operator and it can model for example a projection such as in transmission tomography [Shepp and Vardi, 1982]. Then, (2.2) reduces to a reconstruction problem.

However, it has been experimentally proven that in many situations, the signal of interest may suffer from noise with more complex characteristics than the Gaussian one. In fact, many devices lead to measurements distorted by heteroscedastic noise whose characteristics depend on that of the unknown signal [Healey and Kondepudy, 1994; Tian et al., 2001; Janesick, 2007; Azzari and Foi, 2014; Liu et al., 2014; Chakrabarti and Zickler, 2012; Boubchir et al., 2014]. For example, to better reflect the physical properties of optical communication, the involved noise remains additive Gaussian but its variance is assumed to be dependent on the unknown signal [Moser, 2012]. Signals can also be corrupted by multiplicative noise [Aubert and Aujol, 2008; Buades et al., 2005; Dong and Zeng, 2013; Huang et al., 2009] such as the speckle noise which commonly affects synthetic aperture radar (SAR), medical ultrasound and optical coherence tomography images [Parrilli et al., 2012], or by impulsive noise [Cai et al., 2010]. A mixture of Gaussian and impulsive noise has also been studied in [Xiao et al., 2011; Yan, 2013]. Furthermore, in applications such as astronomy, medicine, and fluorescence microscopy where signals are acquired via photon counting devices, like CMOS and CCD cameras, the number of collected photons is related to some non-additive counting errors resulting in a shot noise [Boulanger et al., 2015; Healey and Kondepudy, 1994; Tian et al., 2001; Janesick, 2007;

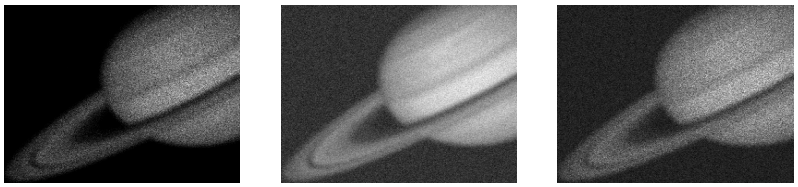
Petropulu et al., 2000]. The latter is non-additive, signal-dependent and it can be modeled by a Poisson distribution [Salmon et al., 2014; Setzer et al., 2010; Jeong et al., 2013; Harizanov et al., 2013; Dupé et al., 2012] [Chaux et al., 2009] [Bonettini and Prato, 2015; Anthoine et al., 2012; Bonettini and Ruggiero, 2011; Altmann et al., 2016]. In this case, when the noise is assumed to be Poisson distributed, the implicit assumption is that Poisson noise dominates over all other noise kinds. Otherwise, the involved noise is a combination of Poisson and Gaussian (PG) components [Roberts and Kingsbury, 2014; Mäkitalo and Foi, 2014, 2012, 2013; Luisier et al., 2011; Li et al., 2015; Lantéri and Theys, 2005; Jezierska et al., 2012, 2014; Foi et al., 2008; Chouzenoux et al., 2015; Kittisuwan, 2016; Calatroni et al., 2016]. The influence of blur and of Poisson and Gaussian noise is illustrated in the example in Figure 2.1.



(a) *Original image of size* 256×256 *and whose intensity values range between 0 and 30.*

(b) *Blurred image with a uniform blur kernel of size* 8×8.

(c) *Noisy image with a centered Gaussian noise of variance 4.*



(d) *Degraded image with a uniform blur kernel of size* 8×8 *and Poisson noise.*

(e) *Degraded image with a uniform blur kernel of size* 8×8 *and additive zero mean Gaussian noise of variance 4.*

(f) *Degraded image with a uniform blur kernel of size* 8×8, *Poisson noise and additive zero mean Gaussian noise of variance 4.*

Figure 2.1: *Influence of different sources of degradation.*

Solving the inverse problem defined by (2.1) amounts to finding a solution $\hat{\mathbf{x}}$ which is reasonably close to the signal of interest $\overline{\mathbf{x}}$ from the observation $\mathbf{z}$. The observation matrix $\mathbf{H}$ and some parameters about the statistics of the noise can also be unknown and have to be estimated either before or jointly with the estimation of $\overline{\mathbf{x}}$. In such case, we say that we have to solve an unsupervised or blind problem. In the context of non blind problems, the best

known approach to solve (2.2) is certainly the least squares method that consists in seeking to minimize the quadratic distance between the measurement and the target signal as follows [Gauss, 1995]:

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in\mathbb{R}^Q}{\operatorname{argmin}} \|\mathbf{z} - \mathbf{H}\overline{\mathbf{x}}\|^2. \tag{2.3}$$

However, among the main difficulties encountered in the resolution of (2.3), the inverse problem can be ill posed in the sense of Hadamard [Hadamard, 1902], which means that it does not satisfy one of the following conditions:

- a solution exists,
- the solution is unique,
- the solution is stable.

In particular, this problem arises when the whole measured data do not allow the existence of a solution to the problem, which may occur for example in many overdetermined problems where the number of parameters to be estimated in the model is smaller than the number of collected data. Furthermore, the solution may not be unique which is the case for example of underdetermined problems where the number of collected data compared to the number of the unknown parameters is insufficient to find the exact solution. The solution can be further highly unstable and sensitive to small disturbance in the data when it is not continuous with respect to the data. Moreover, when the statistics of the noise are far from the Gaussian hypothesis, the least square mean criterion (2.3) fails generally to provide accurate solutions. In this respect, many researches have been devoted to solve (2.1) in a more reliable manner. In the following, we will concentrate on the Bayesian methodology to solve such inverse problems.

§ 2 Bayesian methodology for solving inverse problems

2.1 Bayesian framework

A standard statistical approach to solve the inverse problem (2.1) consists in assuming that $\mathbf{z}$ and $\mathbf{x}$ are random variables with probability distributions $\mathcal{P}_{\mathbf{z}|\mathbf{x}}$ and $\mathcal{P}_{\mathbf{x}}$ respectively instead of unknown constants. Let $\mathsf{p}(\mathbf{z}|\mathbf{x})$ and $\mathsf{p}(\mathbf{x})$ denote their associated densities hence assumed to exist. While $\mathsf{p}(\mathbf{z}|\mathbf{x})$ is the likelihood of the observations and whose expression is derived from the direct model (2.1), $\mathsf{p}(\mathbf{x})$ describes some prior knowledge on $\overline{\mathbf{x}}$ and should be well designed to incorporate the desired characteristics of the solution. In such case, we say that we are working in a Bayesian framework [Bernardo and Smith, 2001; Robert, 2007].

The common Bayesian procedure for signal estimation consists in computing estimators from the posterior distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ that captures all informations inferred about the target signal from the collected data. The posterior density $\mathsf{p}(\mathbf{x}|\mathbf{z})$ is derived from the likelihood and the prior density using Bayes rules:

$$\mathsf{p}(\mathbf{x}|\mathbf{z}) = \frac{\mathsf{p}(\mathbf{x})\mathsf{p}(\mathbf{z}|\mathbf{x})}{\mathsf{p}(\mathbf{z})} \tag{2.4}$$

where

$$\mathsf{p}(\mathbf{z}) = \int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{x} \tag{2.5}$$

is the density of the distribution of $\mathbf{z}$ and which can be viewed as a normalizing constant of $\mathsf{p}(\mathbf{x}|\mathbf{z})$. Note that even if the integral in (2.5) suggests that all x_i, $1 \leqslant i \leqslant Q$, are continuous random variables, discrete values could also be considered by replacing the integral with a sum when required and using discrete probabilities instead of densities.

Under this framework, given a cost function $C(\hat{\mathbf{x}} - \overline{\mathbf{x}})$ that measures the quality of an estimator $\hat{\mathbf{x}}$ in comparison to the true signal $\bar{\mathbf{x}}$, we define a Bayesian estimator as the one with minimum expected cost:

$$\hat{\mathbf{x}} \in \operatorname*{argmin}_{\mathbb{R}^Q} \mathsf{E}_{\mathbf{x}|\mathbf{z}} \left(C(\hat{\mathbf{x}} - \mathbf{x}) \right) \tag{2.6}$$

where

$$\mathsf{E}_{\mathbf{x}|\mathbf{z}} \left(C(\hat{\mathbf{x}} - \mathbf{x}) \right) = \int_{\mathbf{x} \in \mathbb{R}^Q} C(\hat{\mathbf{x}} - \mathbf{x}) \mathsf{p}(\mathbf{x}|\mathbf{z}) \, \mathrm{d}\mathbf{x}. \tag{2.7}$$

In particular, $C(t)$ can be typically one of the following cost functions displayed in Figure 2.2:

- Hit-or-miss: $C_1(t) = \begin{cases} 0, & \text{if} \quad |t| \leqslant \delta \\ 1, & \text{if} \quad |t| > \delta \end{cases}$
- Quadratic: $C_2(t) = t^2$
- Absolute: $C_3(t) = |t|$

Note that these three cost functions are often preferred, because they allow to find the minimum cost solution in a closed form. In fact, the solution to (2.6), with $C = C_1$ and $\delta \ll 1$, arises at the maximum of $\mathsf{p}(\mathbf{x}|\mathbf{z})$. Therefore, the estimator is the mode of the posterior density and is called Maximum a Posteriori (MAP) estimator. Since the normalization constant $\mathsf{p}(\mathbf{z})$ does not depend on $\mathbf{x}$, the MAP estimator can be computed by maximizing only the numerator of (2.4) which is equivalent to solve

$$\hat{\mathbf{x}} \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^Q} - \log \mathsf{p}(\mathbf{z}|\mathbf{x}) - \log \mathsf{p}(\mathbf{x}). \tag{2.8}$$

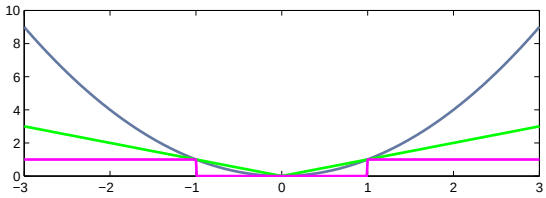


Figure 2.2: *The three classical cost functions: quadratic (blue), absolute (green) and Hit-or-miss for $\delta = 1$ (purple).*

In most applications, no closed-form solution is available to (2.8). In this case, we resort to iterative optimization methods to compute numerically the estimator.

Using the quadratic cost function C_2, the Bayesian estimator reduces to the posterior mean i.e.,

$$\hat{\mathbf{x}} = \mathsf{E}_{\mathbf{x}|\mathbf{z}}(\mathbf{x}) = \int_{\mathbb{R}^Q} \mathbf{x}\, \mathsf{p}(\mathbf{x}|\mathbf{z})\, \mathrm{d}\mathbf{x}. \tag{2.9}$$

This estimator has the desired property of being optimal in terms of minimal average squared error. However, in general due to the involved high-dimensional integral in (2.9), its computation is difficult. Moreover, the posterior distribution is generally known up only to some multiplicative constant as the evaluation of the normalizing constant in (2.4) can be computationally expensive, or even intractable in most real models. In the following, we will refer to this estimator as the minimum mean square error (MMSE) estimator or the posterior mean.

If we choose the absolute value as the cost function C_3, the Bayesian estimator reduces to the median of the posterior distribution i.e., the value which splits the total probability into equal proportions. Note that if the posterior probability density function is Gaussian, the three Bayesian estimators coincide because the mean, the median and the mode of a normal density are identical. However, in most case, as soon as the posterior density is not symmetric, the MAP, the median and the MMSE produce different estimators. In that case, the best choice of the point estimator remains an open problem and it highly depends on the properties of the posterior distribution of interest. For example, when the posterior mean estimator is misleading, this results in a high variance value [Kaipio and Somersalo, 2006, Chap. III, example 1]. In order to qualify the uncertainty on the obtained solutions, one can further use posterior credible regions i.e., Bayesian confidence intervals [Robert, 2007].

2.2 Link with penalized approaches

The optimization criterion of type (2.8), has been widely considered in energy minimization approaches without an explicit reference to probabilities or priors. Solutions to (2.1) are proposed to be the minimizers of some objective function under the generic form:

$$\hat{\mathbf{x}} \in \underset{\mathbf{x}\in\mathbb{R}^Q}{\operatorname{argmin}}\ \Phi(\mathbf{x}) + \Psi(\mathbf{x}) \tag{2.10}$$

where Φ is the data fidelity term whose role is to keep the reconstructed signal close enough to the observation so that useful information will not be discarded in the solution, while Ψ is the regularization term that promotes solutions with some desired properties. At the first glance, penalized approaches and Bayesian estimation seemingly have distinct ways of interpretation of the different terms that constitute the objective function and the posterior law leading to a long lasting disconnection between the two worlds in inverse problems. However, by a simple identification, one can find a strong connection between the two approaches. On the one hand, the MAP estimate in (2.8) can be viewed as the solution to the penalized problem by setting $\Phi(\mathbf{x}) = -\log \mathsf{p}(\mathbf{z}|\mathbf{x})$ and $\Psi(\mathbf{x}) = -\log \mathsf{p}(\mathbf{x})$. On the other hand, the solution to the problem in (2.10) has a statistical interpretation as the MAP estimate under the prior $\mathsf{p}(\mathbf{x}) \propto \exp(-\Psi(\mathbf{x}))$ and the likelihood $\mathsf{p}(\mathbf{z}|\mathbf{x}) \propto \exp(-\Phi(\mathbf{x}))$ provided that $\mathsf{p}(\mathbf{z}|\mathbf{x})$ and $\mathsf{p}(\mathbf{x})$ obey the basic laws of probability (positive, continuous and integrable with respect to $\mathbf{x}$). This result has a very interesting generalization by omitting the integrability condition and allowing the use of some specific prior laws such as degenerate distributions and improper laws [Balakrishnan and Nevzorov, 2004] (see Sec. 2.3), the latter being characterized by:

$$\int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{x}) = +\infty. \tag{2.11}$$

However, the MAP estimate is only one of many possible Bayesian interpretations of the solution in (2.10). In fact, the MMSE in (2.9), can also be viewed as the solution to the optimization problem (2.10), for some regularization term Ψ, in most case distinct from $-\log \mathsf{p}(\mathbf{x})$. This interesting fact stipulates that the optimal estimator in terms of quadratic distance coincides with the MAP estimate with a modified prior. Reciprocally, in some cases, for a given regularization function $\Psi(\mathbf{x})$, the solution in (2.10), can be interpreted as the MMSE or the MAP estimate for two different prior distributions [Gribonval and Machart, 2013; Gribonval, 2011].

2.3 Choice of the prior

The Bayesian paradigm is founded on the subjective view regarding probabilities. In this respect, the choice of the prior is highly correlated with personal

belief and may differ according to the requirement of the designer. This step constitutes one of the important parts in the inference setup and it has been widely discussed from a philosophical point of view in [Box and Tiao, 2011; Irony and Singpurwalla, 1997]. However, regardless the arbitrariness in the selection of the prior, several works have been devoted to provide formal rules to set up the prior distribution. An interesting overview about the formal rules in the selection of the prior can be found in [Kass and Wasserman, 1996]. In general, there are two key questions that should be addressed in this respect:

1. what kind of information do we want to highlight?

2. what are the properties of the resulting posterior distribution?

Most of the theoretical work on prior distributions can be divided into two classes. In the first class, efforts have been made to design informative prior distributions that reflect some common desired properties that one aims to incorporate in the Bayesian estimation such as sparsity in some transform domain, spatial smoothness, dependencies between the signal coefficients etc. Moreover, other researchers have been interested in designing priors which are said to be non-informative in order to remove subjectiveness in the choice of the prior law and to address the problem of the lack of information about the unknown signal. In the second class, conjugate priors have been introduced for a large number of data fidelity models in order to obtain more tractable posterior distributions. Note that these two classes are highly dependent since the prior distribution should achieve a good tradeoff between the two following requirements: it should be well designed to incorporate the wanted properties of the solution without inducing a high computational burden for calculating the posterior distribution. In the following, we will provide a non exhaustive enumeration of the common prior models used in signal processing applications.

2.3.1 Regularization

In this part, we give some examples on how we can choose highly informative models in Bayesian estimation that attempt to reflect the prior knowledge about the target parameters as fully as possible which is equivalent somehow to the role of regularization terms in penalized approaches.

Tikhonov regularization: Perhaps, the most well known method is the Tikhonov regularization which can be seen as a Gaussian prior whose logarithm reads up to an additive constant as follows [Tikhonov, 1963]

$$\log \mathsf{p}(\mathbf{x}) = -\frac{1}{2}\|\boldsymbol{\Gamma}\mathbf{x}\|^2 \tag{2.12}$$

where $\mathbf{\Gamma}$ is a linear operator that promotes some desirable properties. Note that if $\mathbf{\Gamma}$ is not injective, the prior is said to be a degenerate Gaussian [Patel and Read, 1996]. When $\mathbf{\Gamma} = \gamma \mathbf{I}_Q$ where $\gamma > 0$, we promote solutions with smaller norms which leads to the so-called the ℓ_2^2 regularization. Otherwise, $\mathbf{\Gamma}$ can model a high-pass operator such as a Laplacian filter or a difference matrix giving preference to smooth solutions.

Promoting sparsity: A signal is said to be sparse if most of its coefficients are zero. A less strict definition is to qualify a signal as sparse when only a few of its coefficients have large magnitudes while the remaining ones take values around zero. Several prior laws have been proposed to model sparse signals. In the following, we will give some examples of prior distributions that have been widely used in signal processing.

- *Generalized Gaussian distributions*: Zero mean Generalized Gaussian ($\mathcal{GG}$) distributions have the following expression of their density of probability [Woods, 2013]:

$$(\forall t \in \mathbb{R})\ \mathrm{gg}(t;\beta,\gamma) = \frac{\gamma^{\frac{1}{2\beta}}}{2\Gamma(1+1/\beta)} \exp\left(-\gamma t^{2\beta}\right) \tag{2.13}$$

  where Γ is the Gamma function, $\beta > 0$ is the shape parameter and $\gamma > 0$ is the inverse scale parameter. It is a generalization of a family of probability density functions called the generalized Gamma distributions, first introduced in [Stacy, 1962]. Depending on the shape parameter β, two kinds of distributions can be obtained. In particular, $\beta = 1$ yields the Gaussian distribution whereas $\beta = 0.5$ yields the Laplace distribution. Generally, for $0 < \beta < 1$, we have leptokurtic distributions which are characterized by a thinner peak and heavier tails compared to the Gaussian one. In contrast, for $\beta > 1$, a platykurtic distribution is obtained, which is characterized by a flatter peak and thinner tails compared to the Gaussian one. For large values of β, the $\mathcal{GG}$ tends to the uniform distribution. Figure 2.3 depicts the probability density function of the $\mathcal{GG}$ for different shape parameters and for $\gamma = 1$.

- *Student's t distributions*: In order to promote the sparsity of the signal, we can also use the Student's t ($\mathcal{ST}$) distribution whose density is given by

$$(\forall t \in \mathbb{R})\ \mathrm{st}(t;\nu,\gamma) = \frac{\Gamma(\frac{\nu+1}{2})\gamma^{\nu}}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})} \left(\gamma^2 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{2.14}$$

  where $\nu > 0$ is the number of degrees of freedom determining the shape of the distribution and $\gamma > 0$ is the scale parameter. Note that the Cauchy distribution is a particular case when $\nu = 1$. Small values of

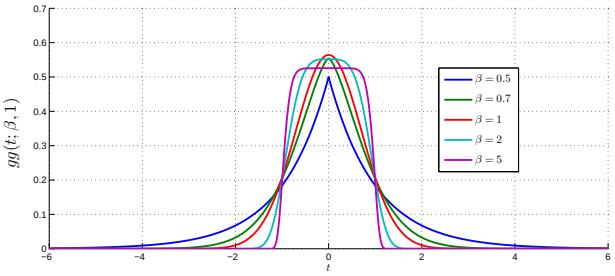


Figure 2.3: *$\mathcal{GG}$ density plot for different values of β and for $\gamma = 1$.*

ν define distributions with heavy tails while we approach the normal distribution as ν increases. Figure 2.4 displays the probability density function of the $\mathcal{ST}$ distribution for different shape parameters and for $\gamma = 1$.

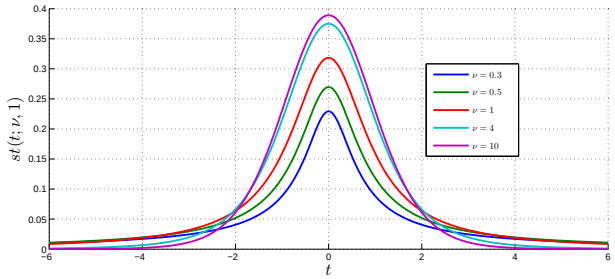


Figure 2.4: *$\mathcal{ST}$ density plot for different values of ν and for $\gamma = 1$.*

- *Scale mixtures of Gaussian distributions*: A more general family of densities can be constructed from scale mixtures of Gaussian distributions ($\mathcal{SMG}$). A discrete scale mixture of Gaussian can be made up using $S > 0$ distinct Gaussian distributions as follows:

$$(\forall t \in \mathbb{R})\ \mathrm{smg}(t) = \sum_{s=1}^{S} w_s \mathrm{n}(t; \mu, \sigma_s^2) \tag{2.15}$$

where $\mathrm{n}(t; \mu, \sigma_s^2)$ denotes the probability density of the Gaussian distribution of mean $\mu \in \mathbb{R}$ and variance σ_s^2, and $w_1, \ldots, w_S$ are positive weights called the mixing constants such that $\sum_{s=1}^{S} w_s = 1$. This type of mixture, being a finite sum, is referred to as a finite mixture. Infinite discrete mixture are covered by setting $S = \infty$. Note that the mixing variables can be either constants or random. In the case of random mixing variables, the probability of the mixing variables is referred to as the mixing distribution. For example, the Bernoulli Gaussian distribution is a scale mixture of Gaussian distribution where the mixing

law is the Bernoulli distribution. Furthermore, this definition can be extended to continuous scale mixtures as follows:

$$(\forall t \in \mathbb{R})\ \mathrm{smg}(t) = \int_{\mathbb{R}^+} \mathsf{p}(w)\mathrm{n}(t;\mu, w\sigma^2)\mathrm{d}w \tag{2.16}$$

where $\mathsf{p}(w)$ denotes the density of the mixing distribution. Different laws for w lead to different prior distributions. For instance, the $\mathcal{ST}$ distribution is a particular case of $\mathcal{SMG}$ when $\mathsf{p}(w)$ is an inverse Gamma distribution with both parameters equal to $\nu/2$. The Laplace distribution is also defined as $\mathcal{SMG}$ for Gamma mixing probability [West, 1987]. More generally, the $\mathcal{GG}$ can be written as a scale mixture of Gaussian provided that $0 < \beta \leqslant 1$ and the mixing distribution is related to alpha-stable distributions [Wainwright and Simoncelli, 1999].

- *Multivariate distributions:* In many applications involving multivariate data, the coefficients of the target signal exhibit high correlations. Then, it is usually preferable to estimate them jointly using their joint distribution rather than their marginal ones. In this respect, multivariate distributions have been proposed to model multivariate random variable. Elliptically contoured ($\mathcal{EC}$) distribution class has particularly deserved considerable attention in the literature [Kai-Tai and Yao-Ting, 1990; Kai-Tai et al., 1990; Kelker, 1970; Bartlett, 1934]. A B-dimensional random vector is said to have an elliptical distribution if its probability density is of the following form:

$$\left(\forall \mathbf{u} \in \mathbb{R}^B\right) \quad \mathrm{ec}(\mathbf{u};\boldsymbol{\Sigma},\boldsymbol{\mu},g) = C_B|\boldsymbol{\Sigma}|^{-1/2} g\left((\mathbf{u}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right) \tag{2.17}$$

where g is the functional parameter, $\boldsymbol{\mu}$ is the location vector, $\boldsymbol{\Sigma}$ is the scale matrix and C_B is a normalization constant [Zozor and Vignat, 2010]. These distributions owe their success to the fact that many results holding for multivariate Gaussian vectors can remain valid for a large number of elliptically symmetric distributions [Chmielewski, 1981]. In particular, this class is stable with linear combination (i.e., every linearly combined elliptical random vector is also elliptical) and when passing to conditional laws (i.e., the conditional distributions of $\mathcal{EC}$ random vectors remain $\mathcal{EC}$). Another interesting property proven in [Gómez-S-M. et al., 2006] is that a necessary and sufficient condition for an $\mathcal{EC}$ distribution to be expressed as a scale mixture of multivariate Gaussian distributions is the alternation of sign of the successive derivatives of its functional parameter g i.e.,

$$(-1)^k g^{(k)}(t) \geqslant 0 \text{ ,for } k \in \mathbb{N} \text{ and } t > 0. \tag{2.18}$$

This property is fulfilled by the multivariate exponential power distribution $\mathcal{MEP}$ used in several applications in signal processing [Gómez-S-M. et al., 2008; Khelil-Cherif and Benazza-Benyahia, 2004; Pizurica

and Philips, 2006; Kwitt et al., 2009] and defined with the following functional parameter:

$$(\forall t \geqslant 0) \quad g(t) = \exp(-\frac{1}{2}t^{\beta}) \tag{2.19}$$

with $\beta > 0$. The normalization constant is defined by:

$$C_B = \frac{\Gamma(\frac{B}{2})}{\pi^{\frac{B}{2}}\Gamma(\frac{B}{2\beta})2^{\frac{B}{2\beta}}}. \tag{2.20}$$

Note that the $\mathcal{MEP}$ can be seen as the extension of the $\mathcal{GG}$ to the multivariate case. Thereby, $\mathcal{MEP}$ are also known as multivariate $\mathcal{GG}$. Similarly to the univariate case, we find the multivariate Laplace distribution for $\beta = 0.5$, and the multivariate Gaussian distribution for $\beta = 1$ while leptokurtic distributions usually used to model sparse random vectors are defined for $0 < \beta < 1$.

In this thesis, we will also consider the Generalized $\mathcal{MEP}$ ($\mathcal{GMEP}$) model [Marnissi et al., 2013], which can be seen as a smooth extension of the $\mathcal{MEP}$ distribution. It is defined as the $\mathcal{EC}$ distribution with the following functional parameter

$$(\forall t \geqslant 0) \quad \varrho_g(t) = g(t + \delta) \tag{2.21}$$

with $\delta > 0$ and whose normalization constant is given by:

$$C_B^{-1} = \frac{2\pi^{\frac{B}{2}}}{\Gamma(\frac{B}{2})} \int_0^{+\infty} t^{B/2-1} \exp\left(-\frac{1}{2}(t+\delta)^{\beta}\right) \mathrm{d}t. \tag{2.22}$$

Note that many standard multivariate distributions can be seen as $\mathcal{EC}$ such as the multivariate $\mathcal{ST}$ distribution, the generalized Laplace distribution, alpha-stable distributions etc. [Zozor and Vignat, 2010].

If the signal of interest is not sparse in the spatial domain, one can use another representation of it where the signal can be represented only by few relevant coefficients. For example, the sinus signal is clearly not sparse in the spatial domain but when we pass to the Fourier domain, the signal becomes extremely sparse. Sparsity can for example be achieved using the discrete gradient transform or higher order differences. More generally, the best sparsifying transform can be qualified as the one leading to the sparsest representation with low complexity. Therein, linear transforms are the simplest way for getting such representation. The coefficients of the signal $\mathbf{x}$ in the new representation denoted by $\mathbf{c} \in \mathbb{R}^M$, are then given with respect to a preassigned operator $\mathbf{F} \in \mathbb{R}^{M \times Q}$ such that $M \geqslant Q$ as follows:

$$\mathbf{c} = \mathbf{F}\mathbf{x}. \tag{2.23}$$

Perhaps, the Fourier and the cosine transforms are the most famous linear representations [Ahmed et al., 1974]. However, their good frequency localization properties are achieved at the expense of a poor spatial (or temporal) localization. Many other transforms have been proposed in the literature to achieve a better tradeoff between both the spatial (or temporal) and frequency localizations such as wavelets [Mallat, 1999], curvelets [Candes et al., 2006], bandlets [Le Pennec and Mallat, 2005] etc. Other approaches have been also proposed to build representation dictionaries directly from the data and adapted to their characteristics [Olshausen and Field, 1996; Aharon et al., 2006]. In the following, we illustrate the sparsity concept with some examples of linear operators often employed in inverse problems of signal/image processing.

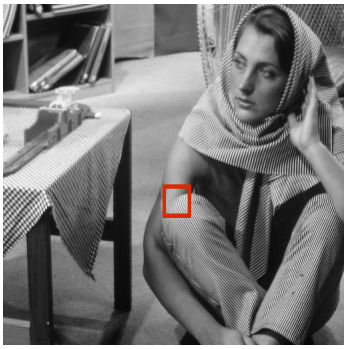
Example 2.1 *Total variation*:
Gradient operators are based on the differences between neighboring values for a given direction. In the context of image processing, we usually consider both horizontal and vertical directions. In the following, $\boldsymbol{\Delta} = \left[\boldsymbol{\Delta}_1^\top, \ldots, \boldsymbol{\Delta}_Q^\top\right]^\top$ will denote the discrete gradient operator having the following formula:

$$(\forall i \in \{1, \ldots, Q\}) \quad \boldsymbol{\Delta}_i \mathbf{x} = \begin{pmatrix} [\boldsymbol{\Delta}^h \mathbf{x}]_i \\ [\boldsymbol{\Delta}^v \mathbf{x}]_i \end{pmatrix} \tag{2.24}$$

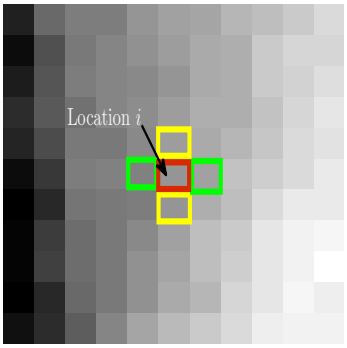
where $\boldsymbol{\Delta}_i$ is the gradient in location i and $\boldsymbol{\Delta}^h$ and $\boldsymbol{\Delta}^v$ are the discrete gradient computed along the horizontal and vertical directions. Figure 2.5 shows an illustration of the gradient across horizontal and vertical directions in the Barbara image. The local gradient at location i is given by the differences between the intensity in i and one of its adjacent pixels. This results in sparse gradient images where the relevant coefficients are concentrated in areas corresponding to edges in the original image. In energy minimization approaches, the sparsity of the gradient is enforced using suitable regularization functions Ψ describing the dynamics of the gradient images [Rudin et al., 1992]. As the quadratic function tends to give over smoothed solutions, many works have proposed to replace the quadratic regularization (ℓ_2^2) with the ℓ_1 norm which result on the so called anisotropic total variation defined by

$$\Psi(\boldsymbol{\Delta}\mathbf{x}) = \sum_{i=1}^{Q} \left(|[\boldsymbol{\Delta}^h \mathbf{x}]_i| + |[\boldsymbol{\Delta}^v \mathbf{x}]_i| \right). \tag{2.25}$$

In fact, the ℓ_1 norm and the quadratic regularization are both concentrated on zero to promote zero values but the ℓ_1 norm has a curve that increases linearly and thus is slower than the quadratic shape of the ℓ_2^2 regularization allowing the solution to take few high values. As such, edges in the image corresponding to the relevant gradient values are less affected with the ℓ_1

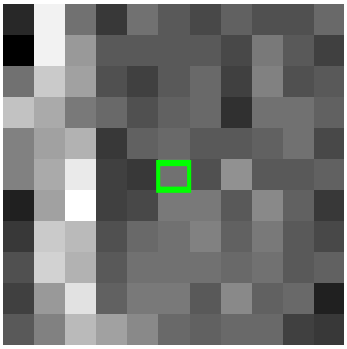


(a) *Original image.*

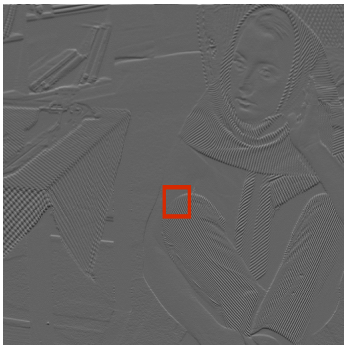


(b) *Fragment of the original image. Green rectangles highlight horizontal neighbors. Yellow rectangles highlight vertical neighbors.*

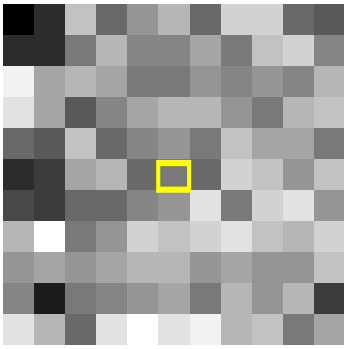
(c) *Horizontal gradient.*



(d) *Local horizontal direction in the location i in green.*



(e) *Vertical gradient.*



(f) *Local vertical direction in the location i in yellow.*

Figure 2.5: *Illustration of concept of the gradient operator.*

norm. This property holds for functions of the form $f(t) = |t|^p$, with $0 \leqslant p < 1$ as it can be seen in Figure 2.6. In this case, the regularization function

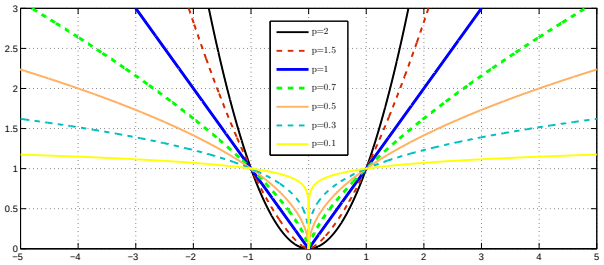


Figure 2.6: *Plot of* $f(t) = |t|^p$ *for* $0 \leqslant p < 2$.

reads:

$$\Psi(\boldsymbol{\Delta}\mathbf{x}) = \sum_{i=1}^{Q} \left(|[\boldsymbol{\Delta}^h \mathbf{x}]_i|^p + |[\boldsymbol{\Delta}^v \mathbf{x}]_i|^p \right). \tag{2.26}$$

However, the ℓ_1 norm is often preferred to other functions since it is considered as the sparsest convex regularization.
Isotropic total variation is used by applying non-separable functions, for instance an ℓ_2 norm on the vectors $\boldsymbol{\Delta}_i \mathbf{x}$, $1 \leqslant i \leqslant Q$. Then, the regularization function has the following form:

$$\Psi(\boldsymbol{\Delta}\mathbf{x}) = \sum_{i=1}^{Q} \sqrt{[\boldsymbol{\Delta}^h \mathbf{x}]_i^2 + [\boldsymbol{\Delta}^v \mathbf{x}]_i^2}. \tag{2.27}$$

Note that since we are interested in Bayesian modeling, we have to design prior distributions and not only regularization functions. It remains possible to define the total variation prior distribution proportional to $\exp(-\gamma\Psi(\boldsymbol{\Delta}\mathbf{x}))$ where $\gamma > 0$ is a positive regularization constant. In many Bayesian applications, we need to know the normalization constant of $\mathsf{p}(\mathbf{x})$ or at least the constant that depends on γ. From [Pereyra et al., 2015], we can deduce that for each total variation regularization defined above, we can associate a prior distribution that has the following form:

$$\mathsf{p}(\mathbf{x}) = C\gamma^{\frac{Q}{\kappa}} \exp(-\gamma\Psi(\boldsymbol{\Delta}\mathbf{x})) \tag{2.28}$$

where $C > 0$ is a constant independent of γ, and $\kappa > 0$ is such that Ψ is $\kappa-$homogenous. For instance, $\kappa = 1$ in (2.25) and (2.27), and $\kappa = p$ in (2.26). However, it is worth noting that the prior (2.28) may be improper in the sense that C can be equal to infinity if $\boldsymbol{\Delta}$ is not corrected in the border of the image.

Example 2.2 *Frame operators*:
Over the past decade, Wavelet transform (WT) has achieved great success in image processing tasks especially when performed recursively [Mallat, 1999].

In fact, by decomposing the image with a set of multiscale bandpass oriented filters in a recursive manner, the two dimensional WT achieves good frequency selectivity. This operation is known as multi-resolution analysis and it results on two classes of coefficients: details and approximations. On the one hand, the subband of wavelet coefficients resulting from high (respectively low) horizontal frequency and low (respectively high) vertical frequency matches the horizontal (respectively vertical) contours (or details) whereas the wavelet coefficients resulting from high frequency in both directions correspond to the diagonal details. On the other hand, the subband of approximation coefficients results from low frequencies in both directions and can be seen as a rough approximation of the image satisfying its spatial statistics. Therefore, we can further apply a WT to this subband to obtain more horizontal, diagonal and vertical details. Figure 2.7 shows a two levels decomposition of the Barbara image using Symmlet wavelet basis of order 8. It can be noted that most of the energy in each of the details subbands is concentrated in areas corresponding to edges in the original image. Several researches have been devoted to improve the efficiency of

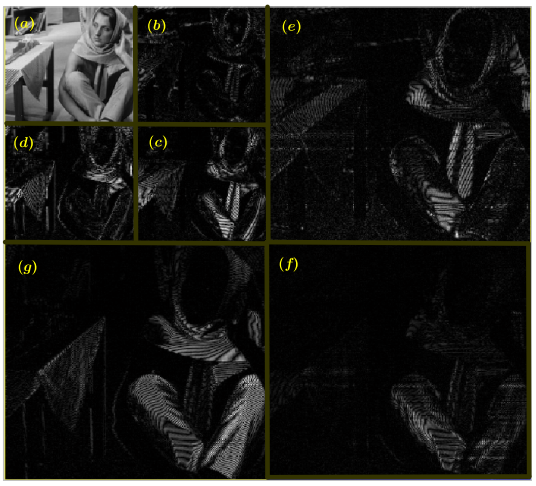


Figure 2.7: *Two decomposition levels of Barbara image using Symmlet wavelet basis of order 8. (a) denotes appoximation subband, (b) and (e) are the horizontal details, (c) and (f) are the diagonal details, (d) and (g) are the vertical details.*

WT towards geometric features of the image (namely textures and edges), by proposing more general representations with a higher degree of redundancy, as well as increased directional selectivity known as frames ($M > Q$). The interest behind the use of frames is related to the efficiency of these overcomplete representations to capture local structures of the signal and their quasi shift-invariance properties [Candes and Donoho, 2002; Fadili and Starck, 2009; Han and Larson, 2000; Pustelnik et al., 2016].

Many researchers have proposed to address image processing tasks such as denoising in multiscale oriented representations by using probability models for the wavelet coefficients [Pustelnik et al., 2016]. In this respect, regularization using such sparse representations in inverse problems, can be divided into two classes:

- Regularization when using a synthesis frame.
- Regularization when using an analysis frame.

In the first class, regularization is applied directly to the sparse coefficients. Consequently, the observation model in (2.1) becomes:

$$\mathbf{z} = \mathcal{D}(\mathbf{H}\mathbf{F}^{*}\overline{\mathbf{c}}) \tag{2.29}$$

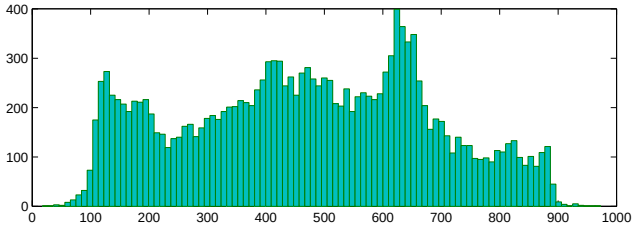
where $\mathbf{F}^{*} \in \mathbb{R}^{Q \times M}$ is the synthesis operator such that $\mathbf{x} = \mathbf{F}^{*}\mathbf{c}$. The objective is then to find the frame coefficients $\hat{\mathbf{c}}$ that are as close as possible to $\overline{\mathbf{c}}$. To this end, a prior distribution is associated with the frame coefficients having the following form:

$$\mathsf{p}(\mathbf{c}) \propto \exp(-\Psi(\mathbf{c})), \tag{2.30}$$

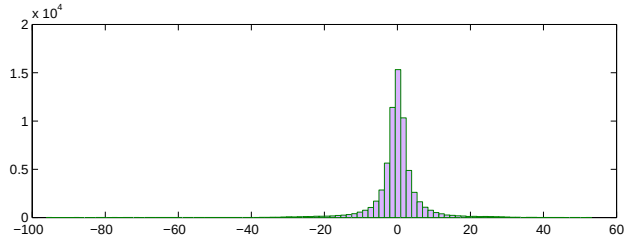
and the reconstructed signal is then computed by $\hat{\mathbf{x}} = \mathbf{F}^{*}\hat{\mathbf{c}}$. Specifically, the marginal distributions of the wavelet detail coefficients are highly leptokurtic. In fact, their histograms show a much higher peak around the mean value, and fat tails at the extreme ends of the probability curve as it can be seen in Figure 2.8. Then, their statistics can be well described using suitable long-tailed distributions such as $\mathcal{GG}$ and $\mathcal{ST}$ distributions. Furthermore, recent work has investigated multivariate models by taking into account the dependencies between coefficients. In fact, it has been observed that frame coefficients of similar position, orientation and scale may exhibit high correlations [Portilla et al., 2003]. Moreover, when dealing with multicomponent images, high correlations also exist between the frame coefficients in the same position through all the components [Marnissi et al., 2013]. In the second class, regularization is applied indirectly to the sparse coefficients through the signal of interest $\mathbf{x}$, and the density of distribution has generally the following form:

$$\mathsf{p}(\mathbf{x}) \propto \exp(-\Psi(\mathbf{F}\mathbf{x})) \tag{2.31}$$

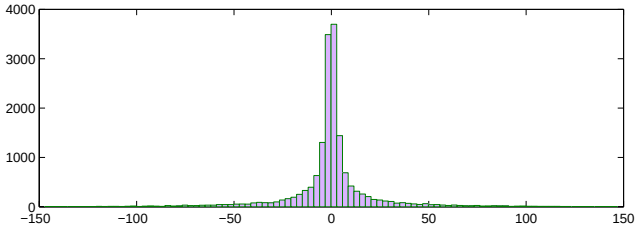
where the frame coefficients are given by $\mathbf{c} = \mathbf{F}\mathbf{x}$. For instance, a common approach is to use ℓ_1 norm regularization and then the prior distribution has the same form as in (2.25). Ψ can also be the ℓ_2 norm applied to a vector gathering the wavelet coefficients of similar positions through all the subbands in each level of decomposition. In particular, a connection can be uncovered between such regularization using undecimated wavelet transform involving a Haar filter and the total variation regularization. However, due



(a) *Subband of apporximations coefficients (subband (a) in Figure 2.7).*



(b) *Subband of horizontal details (subband (g) in Figure 2.7).*



(c) *Subband of diagonal details (subband (c) in Figure 2.7).*

Figure 2.8: *Histograms of wavelet coefficients.*

to the flexibility offered by frames and their ability to generate multiscale analyses, the use of the Haar transform may yield a better performance than gradient operator for natural images [Kamilov et al., 2012]. Note that, analysis and synthesis based regularizations are equivalent in the case of orthonormal transforms.

2.3.2 Conjugate priors

In many situations, prior knowledge about the target parameters is limited or not concrete enough to favor a specific prior probability distribution. Such situations may arise for example in the estimation of the characteristics of the noise or of the regularization parameters on which we have only some general prior information as positivity and mean value but generally few ideas about their statistics. In this case, we have some freedom in selecting the prior distribution and we would rather select it in such a way to result

in a tractable posterior distribution. In fact, for many observation models, there exists a family of prior models for which when combined via Bayes rule, the resulting posterior distribution has a closed form. This observation has motivated the use of conjugate families [Joksch et al., 1964]. However, we should pay attention to the fact that we cannot select any conjugate prior that mathematically gives satisfactory results. The selected prior should indeed be compatible with the available prior knowledge about the target parameter. Hence, conjugate priors define only the sub-class of conjugate family models, that produce analytically tractable posterior distribution and are flexible enough to model prior information about the parameters of interest. A discussion on conjugate models can be found in [Fink, 1997]. In the following, we will give some examples of conjugate priors often used in Bayesian inferences.

Example 2.3 *Conjugate prior for the parameters of the univariate Gaussian model*:
In many Bayesian applications, we have to use univariate Gaussian distributions with unknown parameters. This is the case for example, in denoising and deconvolution problems where the noise is assumed to be Gaussian and uncorrelated.
Let us consider the generic problem of estimating the mean and the variance of a Gaussian distribution given some independent observations $\mathbf{z} \in \mathbb{R}^N$ such that

$$(\forall i \in \{1, \ldots, N\}) \quad \mathsf{p}(z_i|\mu, \tau) = \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{1}{2\tau}(z_i - \mu)^2\right) \tag{2.32}$$

where $\mu \in \mathbb{R}$ and $\tau > 0$. Let $\mathsf{p}(\mu)$ and $\mathsf{p}(\tau)$ be the conjugate prior densities of μ and τ which will be defined later. The joint distribution of $\mathbf{z}$, μ and τ is then given by

$$\mathsf{p}(\mathbf{z}, \mu, \tau) = \mathsf{p}(\mu)\mathsf{p}(\tau) \prod_{i=1}^{N} \mathsf{p}(z_i|\mu, \tau). \tag{2.33}$$

The density of the conditional posterior distribution of μ is given by:

$$\mathsf{p}(\mu|\mathbf{z}, \tau) \propto \exp\left(-\frac{N}{2\tau}\left(\mu - N^{-1}\sum_{i=1}^{N} z_i\right)^2\right) \mathsf{p}(\mu). \tag{2.34}$$

Then, by using a normal prior for μ i.e., $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, its posterior distribution remains Gaussian with variance $\sigma^2 = \left(\frac{N}{\tau} + \frac{1}{\sigma_0^2}\right)^{-1}$ and mean $\sigma^2(\sigma_0^{-2}\mu_0 + \tau^{-1}\sum_{i=1}^{N} z_i)$.
The density of the conditional posterior distribution of τ is given by:

$$\mathsf{p}(\tau|\mathbf{z}, \mu) \propto \tau^{-N/2} \exp\left(-\frac{1}{\tau}\sum_{i=1}^{N} \frac{(\mu - z_i)^2}{2}\right) \mathsf{p}(\tau). \tag{2.35}$$

It can be noted that, if we choose the prior distribution for τ as the inverse Gamma distribution with parameter a and b, the resulting posterior distribution is also an inverse Gamma with parameter $a+N/2$ and $b+\sum_{i=1}^{N}\frac{(\mu-z_i)^2}{2}$.

Example 2.4 *Conjugate prior for the covariance matrix of the multivariate Gaussian model:*
Let us consider the problem of estimating the covariance matrix of a zero mean multivariate Gaussian distribution given some P independent observation vectors $\mathbf{z}_1,\dots,\mathbf{z}_P \in \mathbb{R}^N$ where

$$(\forall i \in \{1,\dots,P\}) \quad \mathsf{p}(\mathbf{z}_i|\mathbf{\Sigma}) = \frac{|\mathbf{\Sigma}|^{-1/2}}{(\sqrt{2\pi})^N}\exp\left(-\frac{1}{2}\|\mathbf{\Sigma}^{-1/2}\mathbf{z}_i\|^2\right) \tag{2.36}$$

where $\mathbf{\Sigma}$ is the unknown positive definite matrix. Given a prior distribution $\mathsf{p}(\mathbf{\Sigma})$, the posterior distribution reads:

$$\mathsf{p}(\mathbf{\Sigma}|\mathbf{z}_1,\dots,\mathbf{z}_P) \propto |\mathbf{\Sigma}|^{-P/2}\exp\left(-\frac{1}{2}\ \text{trace}\left(\mathbf{\Sigma}^{-1}\sum_{i=1}^{P}\mathbf{z}_i\mathbf{z}_i^\top\right)\right)\mathsf{p}(\mathbf{\Sigma}). \tag{2.37}$$

Then, when $\mathsf{p}(\mathbf{\Sigma})$ is chosen to be the probability density of the inverse Wishart distribution of parameters $R > N$ and $\mathbf{A}$, the posterior distribution of $\mathbf{\Sigma}$ reduces to the inverse Wishart distribution of parameters $R+P$ and $\mathbf{A}+\sum_{i=1}^{P}\mathbf{z}_i\mathbf{z}_i^\top$. Similarly, when considering instead the precision matrix $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$, the conjugate prior for $\mathbf{\Lambda}$ is a Wishart distribution.

Example 2.5 *Conjugate prior for the regularization parameter in the case of a κ-homogenous regularization:*
A common class of prior is derived from homogeneous regularizations such as $\mathcal{GG}$, total variation prior and more generally all those derived from norms and compositions of norms with linear operators. In such case, the prior density is written as follows:

$$\mathsf{p}(\mathbf{x}|\gamma) = C\gamma^{\frac{Q}{\kappa}}\exp(-\gamma g(\mathbf{x})), \tag{2.38}$$

where g is a positive κ-homogeneous function, $\gamma > 0$ is the regularization parameter and $C > 0$ is a constant independent of γ [Pereyra et al., 2015]. Here, we are interested in the problem of estimating γ. Let $\mathsf{p}(\gamma)$ be the density of probability of the regularization parameter. It follows that the posterior distribution of γ is

$$\mathsf{p}(\gamma|\mathbf{x}) \propto \gamma^{\frac{Q}{\kappa}}\exp(-\gamma g(\mathbf{x}))\mathsf{p}(\gamma). \tag{2.39}$$

It can be noted that when using the Gamma prior with parameter a and b for γ, $\mathsf{p}(\gamma|\mathbf{x})$ reduces to a Gamma prior of parameter $a+Q/\kappa$ and $b+g(\mathbf{x})$.

2.3.3 Non informative priors

In many applications, we have no sufficient prior information about the target parameters to be interpreted into a mathematical form as a probability distribution. Then, it is not recommended to use a strict subjective prior probability that may sway the inference process in a wrong direction. However, the Bayesian framework requires the presence of a prior distribution. In such cases, the prior law is only used as a technical way of expressing available information necessary to perform the Bayesian inference. Therein, many works have been done in order to provide priors that impart as little information as possible about the target parameters in order to allow the observations to have their maximal weight in the posterior distribution. These priors are often called non-informative or objective [Kass and Wasserman, 1996] since they generally express some objective information such as, the positiveness or the definition domain.

Certainly, the simplest and oldest rule to define such prior is based on the indifference principle that gives equal probabilities to all the domain where the parameter is defined. Otherwise, a constant prior is assigned to the target parameter $\mathbf{x}$ (i.e., $\mathsf{p}(\mathbf{x}) = 1$) which is known as the Laplace non-informative prior [Laplace, 1820]. Otherwise, reference priors are among the most commonly used non-informative priors in the literature. Their introduction is mainly related to an attempt to find the prior that maximizes the expected amount of information provided by the observations in the posterior law. Such prior is found by maximizing $\mathcal{I}$ defined as follows [Bernardo, 1979]:

$$\mathcal{I}(\mathsf{p}(\mathbf{x})) = \int_{\mathbb{R}^N} \mathsf{p}(\mathbf{z}) \mathcal{KL}\left(\mathsf{p}(\mathbf{x}|\mathbf{z}) \| \mathsf{p}(\mathbf{x})\right) \mathrm{d}\mathbf{z} \tag{2.40}$$

where

$$\mathcal{KL}\left(\mathsf{p}(\mathbf{x}|\mathbf{z}) \| \mathsf{p}(\mathbf{x})\right) = \int_{\mathbb{R}^Q} \mathsf{p}(\mathbf{x}|\mathbf{z}) \log\left(\frac{\mathsf{p}(\mathbf{x}|\mathbf{z})}{\mathsf{p}(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \tag{2.41}$$

is the Kullback-Leibler divergence measuring the similarities between the prior and the posterior densities. Reference priors are defined in the asymptotic case i.e., as the limit of the obtained solution in (2.40) when the number of independent observations N tends to infinity. In particular, for a finite discrete parameter space, the integral is replaced by a finite sum and the reference prior is shown to be the uniform distribution over all the possible values [Bernardo, 1979]. For continuous variables, the reference prior is given for several models satisfying some regularity conditions such as asymptotic normality [Dawid, 1970; Lindley, 1961; Bernardo, 1979] by:

$$\mathsf{p}(\mathbf{x}) \propto |\boldsymbol{\Omega}(\mathbf{x})|^{1/2} \tag{2.42}$$

where $\boldsymbol{\Omega}(\mathbf{x})$ is the Fisher information matrix defined by

$$(\forall i, j \in \{1, \ldots, Q\}) \quad \Omega_{i,j} = -\mathsf{E}_{\mathbf{z}}\left(\frac{\partial \log \mathsf{p}(\mathbf{z}|\mathbf{x})}{\partial x_i x_j}\right) \tag{2.43}$$

and $|\boldsymbol{\Omega}(\mathbf{x})|$ is its determinant. This prior is known as the Jeffreys' prior [Jeffreys, 1998]. In the following, we give examples of applications to inverse problems where non-informative priors are commonly used.

Example 2.6 *Jeffrey priors for the parameters of a univariate Gaussian model*:

Let us reconsider Example 2.3 and assume that we have no prior information about the unknown parameters (μ, τ). Then, we will instead use non-informative priors. The minus of the log-likelihood is given by:

$$-\log \mathsf{p}(\mathbf{z}|\mu,\tau) = \frac{N}{2}\log(2\pi\tau) + \frac{1}{2\tau}\sum_{i=1}^{N}(z_i-\mu)^2. \tag{2.44}$$

Consequently, the Fisher matrix is given by:

$$\begin{aligned}\boldsymbol{\Omega}(\mu,\tau) &= \mathsf{E}_{\mathbf{z}}\begin{pmatrix} \dfrac{N}{\tau} & \dfrac{1}{\tau^2}\displaystyle\sum_{i=1}^{N}(\mu-z_i) \\ \dfrac{1}{\tau^2}\displaystyle\sum_{i=1}^{N}(\mu-z_i) & -\dfrac{N}{2\tau^2}+\dfrac{1}{\tau^3}\displaystyle\sum_{i=1}^{N}(z_i-\mu)^2 \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{N}{\tau} & 0 \\ 0 & \dfrac{N}{2\tau^2} \end{pmatrix}. \end{aligned} \tag{2.45}$$

It follows that the Jeffrey's prior of (μ, τ) is

$$\mathsf{p}(\mu,\tau) \propto \frac{1}{\tau^{3/2}}. \tag{2.46}$$

Note that, in general, due to the lack of information, μ and τ are supposed to be independent i.e, the prior density is written as follows [Robert, 2007]

$$\mathsf{p}(\mu,\tau) = \mathsf{p}(\mu)\mathsf{p}(\tau). \tag{2.47}$$

Then, the Jeffrey's priors are given by:

$$\mathsf{p}(\mu) = 1 \tag{2.48}$$

and

$$\mathsf{p}(\tau) \propto \frac{1}{\tau}. \tag{2.49}$$

In this case, the Jeffrey's prior $\mathsf{p}(\mu,\tau) \propto \dfrac{1}{\tau}$ and the reference prior coincide [Bernardo, 1979]. In particular, (2.49) can be seen as the inverse Gamma distribution with both parameters equal to zero. The posterior distribution for τ reduces to the inverse Gamma with shape parameter $N/2$ and rate parameter $\displaystyle\sum_{i=1}^{N}\frac{(\mu-z_i)^2}{2}$ while the posterior distribution of μ is Gaussian with variance $\dfrac{\tau}{N}$ and mean $N^{-1}\sum_{i=1}^{N} z_i$.

Example 2.7 *Jeffrey prior for the regularization parameter in the case of a κ-homogenous regularization:*
Let us consider Example 2.4 and assume that we have no prior information about γ and that we are not able to choose suitable values a and b for the conjugate Gamma distribution. The Jeffrey's prior is given by:

$$\mathsf{p}(\gamma) \propto \frac{1}{\gamma} \tag{2.50}$$

which can be understood as the Gamma distribution with both parameters equal to zero. $\mathsf{p}(\gamma|\mathbf{x})$ reduces to the density of a Gamma prior of parameters Q/κ and $g(\mathbf{x})$.

In practice, most improper priors of interest can be also interpreted as limits of standard distributions (e.g., Gaussians of infinite variance, Gamma and inverse Gamma densities with both parameters equal to zero, uniform densities on infinite intervals). Note that, in most applications, even if the prior density is improper, the resulting posterior distribution remains proper.

2.4 Hierarchical Bayesian modeling

In Bayesian inference, we often have more than one unknown variable to estimate. Prior distributions are assigned to these variables either through their joint distributions if the variables are supposed to be dependent or through their marginal distributions if they are independent. The unknown variables can be generally structured into different groups (see Figure 2.9). First, we have the main variables which include the target signal, the blurring operator in the case of blind deconvolution, the noise statistics etc. Informative priors (regularization or conjugate) are generally assigned to these variables when prior knowledge is available. These informative priors may introduce some new variables which are generally unknown (e.g., we assign to the target image a total variation prior with unknown regularization parameter, to the variance of the noise an inverse Gamma prior with unknown rate parameter, to the signal of interest a scale mixture of Gaussian distributions which involves unknown mixing variables etc.) These new variables define the second group and are usually called parameters. The latter are generally modeled with conjugate or non-informative priors. Conjugate priors may also involve new unknown variables called hyperparameters which are in most case modeled by non-informative distributions. This structured model is referred to as hierarchical Bayesian modeling which is at the core of Bayesian inference.

2.5 Algorithms for computing Bayesian estimates

Once the posterior distributions of these unknown variables are determined, a Bayesian estimator is derived for each unknown variable using its posterior distribution given the remaining ones. The common Bayesian point estimates

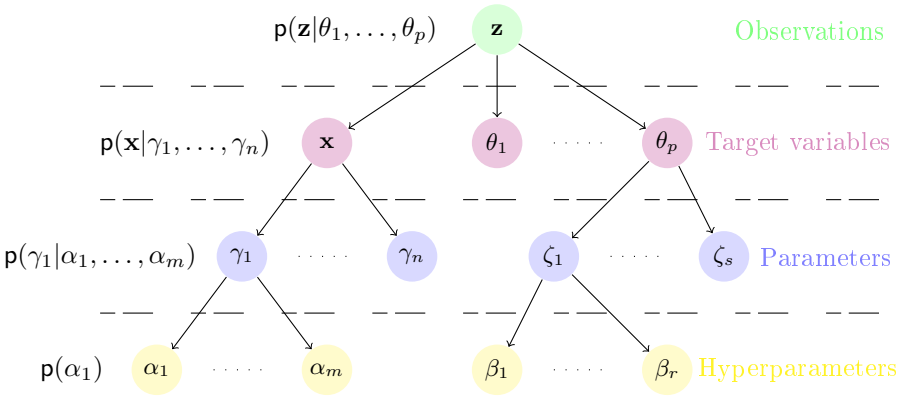


Figure 2.9: *Hierarchical Bayesian model.*

used in the literature are the mean (MMSE) or the mode (MAP) of the posterior distribution.

Algorithms for computing the MAP estimate: The MAP estimate is computed by minimizing a cost function equal up to an additive constant to the the minus logarithm of the posterior density. Based on the properties of this cost function (differentiablility, convexity, continuity etc.), several algorithms can be employed to solve this minimization problem. Among the mostly used algorithms, we can mention descent algorithms such as nonlinear conjugate gradient and quasi-Newton methods. Other algorithms are based on majorize-minimize strategies [Chouzenoux et al., 2011] such as half-quadratic approaches [Geman and Yang, 1995; Nikolova and Ng, 2005; Charbonnier et al., 1994; Ciuciu and Idier, 2002], expectation-maximization [Champagnat and Idier, 2004; Celeux and Diebolt, 1985]. For non-differentiable objective functions, one may use primal proximal algorithms [Combettes and Pesquet, 2011, 2007] and primal-dual methods [Esser et al., 2010; Chambolle and Pock, 2011].

Note that, the parameters and the hyperparameters in the hierarchical Bayesian model can be jointly estimated with the target signal using either the MAP estimate or other approaches [Pereyra et al., 2015; Thompson et al., 1991; Archer and Titterington, 1995; Molina et al., 1999; Almeida and Figueiredo, 2013; Bardsley and Goldes, 2009; Bertero et al., 2010]. One popular approach is based on the discrepancy principle proposed in order to address the problem of selecting the regularization parameter in deconvolution problems involving Gaussian noise [Thompson et al., 1991]. The regularization parameter is chosen such that the variance of the residual (i.e, the difference between the observed image and the blurred estimate) is equal to the variance of the noise. This method has been also extended to other data fidelity terms such as the Poisson noise and the signal dependent Gaus-

sian noise [Bertero et al., 2010; Bardsley and Goldes, 2009]. Among other well known approaches, we can mention the generalized cross validation and the L-curve [Golub et al., 1979; Hansen and O'Leary, 1993].

Algorithms for computing the MMSE estimate: While the MAP estimate is computed by minimizing a cost function, the MMSE estimate requires to calculate an integral which is almost intractable analytically as the exact posterior density is generally known up to a multiplicative constant. Moreover, when the problem dimension increases, basic numerical integration methods fails to compute the involved integral. Given the difficulty in directly computing the MMSE estimate, several methods have been proposed to address this issue. They can be divided into 2 classes:

- *Stochastic simulation methods*: In order to avoid analytically computing the integral, the Monte Carlo approach is a classical alternative solution which consists of simulating a sufficient number of i.i.d. random variables from the posterior distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ and approximating the MMSE estimator by the empirical average over all these samples. However, the target posterior is often complex as it does not present a closed form, so that direct sampling is not always possible. To alleviate this difficulty, sampling algorithms have been developed to explore complicated distributions such as importance sampling, rejection sampling and Monte Carlo Markov chain algorithms [Kong, 1992; Liu, 2001; Gilks et al., 1999; Gamerman and Lopes, 2006].

- *Approximations methods*: Instead of addressing the true posterior distribution, approximation methods search for a simpler distribution that well approximates it and that is computationally tractable. The MMSE estimator is then approximated with the help of this new distribution. Various approximation strategies have been developed. In general, we can distinguish two classes of methods. First, one can make use of the MAP estimate and try to construct a local approximation of the probability mass around it. This strategy seems attractive since the MAP estimate can usually be computed in a straightforward manner. Otherwise, one can compute the approximate density by maximizing some criterion that expresses the similarity between the true and the approximate law. A well common criterion is the Kullback-Leibler divergence. In the first class of method, we can mention the Laplace approximation [Kass and Raftery, 1995] while the second class is related to Variational Bayesian methods [Parisi, 1988].

Note that, unlike optimization methods that give point wise estimates of the target variables, these Bayesian algorithms provide a good summary (respectively approximation) of the whole posterior distribution via the generated samples (respectively the approximate distribution). Thus, these algorithms

are not just proposed to compute the MMSE estimator but they can also be used to estimate others statistics such as the variance and the median. Moreover, in some cases, point wise estimates may not be a good representative of the posterior law and thus we can use these methods to quantify the uncertainty in the computed solution which is known as the credible regions [Robert, 2007].
In the following, we give a review on some stochastic simulation algorithms and approximation methods.

§ 3 Stochastic simulations methods

Monte Carlo integration is a powerful method for computing numerical integration using random variables. It estimates the expectation of a function ζ under the probability distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ as follows:

$$\mathsf{E}_{\mathbf{x}|\mathbf{z}}(\zeta(\mathbf{x})) = \int_{\mathbb{R}^Q} \zeta(\mathbf{x})\mathsf{p}(\mathbf{x}|\mathbf{z})\mathsf{d}\mathbf{x} \simeq \frac{1}{P}\sum_{i=1}^{P} \zeta(\mathbf{x}^{(i)}) \tag{2.51}$$

where $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}$ are P samples following the distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$. Note that, thanks to the law of large numbers, the Monte Carlo estimate is unbiased when P tends to infinity. Most importantly, the more numerous the samples are, the more accurate and reliable the Monte Carlo estimates are. In general, the number of needed samples for achieving good estimates depends on both the dimension of $\mathbf{x}$ and the properties of the generated samples. This method is useful to compute the posterior statistics such as the posterior mean for $\zeta(\mathbf{x}) = \mathbf{x}$ and the standard deviation for $\zeta(\mathbf{x}) = (\mathbf{x} - \mathsf{E}_{\mathbf{x}|\mathbf{z}}(\mathbf{x}))^2$. Since direct simulation from the posterior distribution is not always possible, an alternative is to resort to sampling methods.

3.1 Importance sampling

The idea behind importance sampling, is that, when $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ is difficult to sample from, one can still write the integral in (2.51) as follows [Glynn and Iglehart, 1989]:

$$\mathsf{E}_{\mathbf{x}|\mathbf{z}}(\zeta(\mathbf{x})) = \int_{\mathbb{R}^Q} \zeta(\mathbf{x})\pi(\mathbf{x})\frac{\mathsf{p}(\mathbf{x}|\mathbf{z})}{\pi(\mathbf{x})}\mathsf{d}\mathbf{x} \tag{2.52}$$

where π is a density of probability that one can easily sample from in a direct manner. Hence, it is sufficient to use samples from the auxiliary distribution associated with the density π and then compute the estimate $\mathsf{E}_{\mathbf{x}|\mathbf{z}}(\zeta(\mathbf{x}))$ by weighting these generated samples according to

$$\mathsf{E}_{\mathbf{x}|\mathbf{z}}(\zeta(\mathbf{x})) = \frac{1}{P}\sum_{i=1}^{P} \varpi^{(i)}\zeta(\mathbf{x}^{(i)}) \tag{2.53}$$

where for every $i \in \{1, \ldots, P\}$, $\mathbf{x}^{(i)}$ are random variables following the distribution of density π and $\varpi^{(i)} = \dfrac{\mathsf{p}(\mathbf{x}^{(i)}|\mathbf{z})}{\pi(\mathbf{x}^{(i)})}$ are known as the importance weights. Note that, this estimate is also unbiased like in (2.51) for sufficient large values of P. In situations when the posterior distribution can be computed only up to a multiplicative constant, one can still use the importance sampling approach as follows:

$$\mathsf{E}_{\mathbf{x}|\mathbf{z}}(\zeta(\mathbf{x})) = \frac{1}{P} \frac{\sum_{i=1}^{P} \varpi^{(i)} \zeta(\mathbf{x}^{(i)})}{\sum_{i=1}^{P} \varpi^{(i)}} \tag{2.54}$$

with $\varpi^{(i)} = \dfrac{\tilde{\mathsf{p}}(\mathbf{x}^{(i)}|\mathbf{z})}{\tilde{\pi}(\mathbf{x}^{(i)})}$ where $\tilde{\mathsf{p}}(\mathbf{x}|\mathbf{z})$ and $\tilde{\pi}(\mathbf{x})$ are equal to $\mathsf{p}(\mathbf{x}|\mathbf{z})$ and $\pi(\mathbf{x})$ up to multiplicative constants. It is worth noting that, in this case, this estimate is no longer unbiased. The choice of the auxiliary distribution with density π is important to control this error [Kong, 1992; Rubinstein and Kroese, 2013; Quang et al., 2012]. Moreover, although the importance sampling method is simple to implement, the estimate can often have very high variance. In fact, even for simple models, the variance can be unbounded [Geweke, 1989]. In order to alleviate the shortcomings of importance sampling related to the choice of importance weights, methods based on acceptance-rejection principle have been proposed.

3.2 Rejection sampling

Similarly to importance sampling idea, rejection sampling strategies make use of an auxiliary density π. However, the latter should be selected such that [Gilks and Wild, 1992]

$$\left(\forall \mathbf{x} \in \mathbb{R}^{Q}\right) \quad \mathsf{p}(\mathbf{x}|\mathbf{z}) \leqslant c\pi(\mathbf{x}) \tag{2.55}$$

where c is a positive constant. In other words, $c\pi(\mathbf{x})$ should majorize the probability density $\mathsf{p}(\mathbf{x}|\mathbf{z})$. Samples are obtained from $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ by generating samples from the distribution with density $\pi(\mathbf{x})$ and accept or reject them stochastically according to Algorithm 1. Note that the generated samples are i.i.d. according to the target distribution. The number of iterations needed to successfully generate each sample is also a random number following a geometric distribution with success probability

$$p = \mathsf{P}\left(u \leqslant \mathsf{E}_{\tilde{\mathbf{x}}}\left(\frac{\mathsf{p}(\tilde{\mathbf{x}}|\mathbf{z})}{c\pi(\tilde{\mathbf{x}})}\right)\right). \tag{2.56}$$

Thus, we need on average, $1/p$ iterations to generate one sample. As u is a uniform random number, we have then $p = \mathsf{E}_{\tilde{\mathbf{x}}}\left(\dfrac{\mathsf{p}(\tilde{\mathbf{x}}|\mathbf{z})}{c\pi(\tilde{\mathbf{x}})}\right)$. By recalling that $\tilde{\mathbf{x}}$ follows the distribution with density π, we deduce that $p =$

Algorithm 1 Rejection sampling

1: **for** $t = 1, \ldots, P$ **do**
2: Generate $\tilde{\mathbf{x}}^{(t)}$ from the auxiliary probability distribution of density $\pi(\mathbf{x})$ satisfying (2.55)
3: **Acceptance-Rejection:**
4: Generate $u \sim \mathcal{U}(0,1)$
5: Compute

$$\alpha(\tilde{\mathbf{x}}^{(t)}) = \frac{\mathsf{p}(\tilde{\mathbf{x}}^{(t)}|\mathbf{z})}{c\pi(\tilde{\mathbf{x}}^{(t)})}$$

6: **if** $u < \alpha(\tilde{\mathbf{x}}^{(t)})$ **then**
7: **Accept:** $\mathbf{x}^{(t)} = \tilde{\mathbf{x}}^{(t)}$
8: **else**
9: **Reject:** Go to 2
10: **end if**
11: **end for**

$\int_{\mathbb{R}^Q} \frac{\mathsf{p}(\tilde{\mathbf{x}}|\mathbf{z})}{c\pi(\tilde{\mathbf{x}})}\pi(\tilde{\mathbf{x}})\mathrm{d}\tilde{\mathbf{x}} = 1/c$. Then, the smaller c is, the faster is the sampling algorithm. In the following, we show an example of application of the rejection sampling algorithm.

Example 3.1 *Sampling from a Beta distribution ($\mathcal{B}$) using rejection sampling algorithm:*
The density of the $\mathcal{B}$ distribution is defined as follows:

$$(\forall t \in [0,1]) \quad \mathrm{bet}(t;a,b) = \frac{(1-t)^{a-1}t^{b-1}}{\beta(a,b)} \tag{2.57}$$

where a and b are positive and $\beta(a,b)$ is the beta function. Assume that $a > 1$ and $b > 1$. The mean of this distribution is $\dfrac{b}{a+b}$ and the mode is $\dfrac{b-1}{a+b-2}$. Let π be the density of the uniform distribution on $[0,1]$, we have:

$$c = \sup_{0 \leqslant t \leqslant 1} \frac{\mathrm{bet}(t;a,b)}{\pi(t)} = \mathrm{bet}\left(\frac{b-1}{a+b-2};a,b\right) = \frac{(a-1)^{a-1}(b-1)^{b-1}}{\beta(a,b)(a+b-2)^{a+b-2}}. \tag{2.58}$$

We propose to use the uniform distribution as auxiliary distribution. In particular, for $a = b$, the expected value is equal to $1/2$ as for the uniform distribution and the average of the acceptance rate is $p = \beta(a,a)4^{a-1}$. However, for large values of a, this algorithm is not efficient since the acceptance rate becomes very small (i.e., for $a = 10$, $p = 0.28$ and for $a = 100$, $p = 0.08$). Consequently, we need more time to generate the desired samples. This seems logical since in this case the two densities are not very similar.

In particular, unlike the uniform distribution, the variance of the $\mathcal{B}$ distribution decreases as $1/a$ when a increases and then the samples should be more concentrated around the mode of the distribution.

In general, the rejection sampling algorithm is also of limited interest in practical applications for large scale problems and with complicated distribution. In fact, it is generally difficult to find an appropriate auxiliary distribution satisfying (2.55) with c small enough to achieve reasonable values of acceptance rate across the whole high dimensional space.

3.3 Markov chain Monte Carlo methods

3.3.1 Theory of Markov chains

Definition 3.2 *Markov chain*
A Markov chain is a sequence of random variables $\mathbf{x}^{(t)}$, $t > 0$ belonging to some set $\mathbb{D} \subset \mathbb{R}^Q$ called the state space such that the conditional distribution of $\mathbf{x}^{(t+1)}$ given $\mathbf{x}^{(t)}, \ldots, \mathbf{x}^{(0)}$ depends only on $\mathbf{x}^{(t)}$, that is

$$(\forall t \geqslant 0, \mathcal{A} \subset \mathbb{D}) \;\; \mathsf{P}(\mathbf{x}^{(t+1)} \in \mathcal{A} | \mathbf{x}^{(t)}, \ldots, \mathbf{x}^{(0)}) = \mathsf{P}(\mathbf{x}^{(t+1)} \in \mathcal{A} | \mathbf{x}^{(t)}). \quad (2.59)$$

The conditional distribution in (2.59) is called the transition probability. The latter controls the moves between the possible states of the Markov chain and will be denoted by $\mathcal{T}_t$:

$$(\forall t > 0, \mathcal{A} \subset \mathbb{D}) \;\; \mathcal{T}_t(\mathcal{A} | \mathbf{x}^{(t-1)}) = \mathsf{P}(\mathbf{x}^{(t)} \in \mathcal{A} | \mathbf{x}^{(t-1)}). \quad (2.60)$$

Thus, a Markov chain is described by its initial state $\mathbf{x}^{(0)}$ and the transition probabilities $\mathcal{T}_t$, $t \geqslant 0$. The chain is said to be homogeneous if the transition probability does not depend on t that is, for all $t \geqslant 0$, $\mathcal{T}_t = \mathcal{T}$. The transition density defining the density of probability of moving from $\mathbf{x}$ to $\mathbf{y}$ for all $\mathbf{x}$, $\mathbf{y} \in \mathbb{D}$, will be denoted by $\mathsf{t}(\mathbf{y}|\mathbf{x})$. Furthermore, $\mathcal{T}^{(t)}(.|\mathbf{x}^{(0)})$ will denote the conditional distribution of $\mathbf{x}^{(t)}$ when the chain starts at $\mathbf{x}^{(0)}$ that is

$$(\forall t \geqslant 0, \mathcal{A} \subset \mathbb{D}) \;\; \mathcal{T}^{(t)}(\mathcal{A} | \mathbf{x}^{(0)}) = \mathsf{P}(\mathbf{x}^{(t)} \in \mathcal{A} | \mathbf{x}^{(0)}). \quad (2.61)$$

Definition 3.3 *Invariant distribution*
An homogeneous Markov chain with transition probability $\mathcal{T}$ has invariant (or stationary) distribution $\mathcal{F}^*$ of density f^* if

$$(\forall \mathbf{y} \in \mathbb{D}) \;\; \mathsf{f}^*(\mathbf{y}) = \int_{\mathbb{D}} \mathsf{t}(\mathbf{y}|\mathbf{x}) \mathsf{f}^*(\mathbf{x}) \mathsf{d}\mathbf{x}. \quad (2.62)$$

Then, the invariance property means that, if the marginal distribution of the initial state is $\mathcal{F}^*$, all subsequent states of the Markov chain are distributed

according to $\mathcal{F}^*$. If we suppose that the direct sampling from $\mathcal{F}^*$ is complicated, the objective is to find conditions on the transition probability, under which, from an initial arbitrary state $\mathbf{x}^{(0)}$ there exists $t_0 > 0$ such that the state $\mathbf{x}^{(t_0)}$ follows $\mathcal{F}^*$ that is the invariant distribution $\mathcal{F}^*$ is the equilibrium distribution of the Markov chain i.e.,

$$(\forall \mathcal{A} \subset \mathbb{D}) \quad \lim_{t \to +\infty} \mathcal{T}^{(t)}(\mathcal{A}|\mathbf{x}^{(0)}) = \mathcal{P}_{\mathbf{x}|\mathbf{z}}(\mathcal{A}) \tag{2.63}$$

for $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ almost all $\mathbf{x}^{(0)}$ [Tierney, 1994]. In such cases, the chain is said to be ergodic.

Proposition 3.4 *Reversibility and invariance [Tierney, 1994]*
An homogeneous Markov chain with transition probability $\mathcal{T}$ of density t is reversible if there exists a distribution $\mathcal{F}$ with density f such that

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbb{D}) \quad \mathsf{f}(\mathbf{x})\mathsf{t}(\mathbf{y}|\mathbf{x}) = f(\mathbf{y})\mathsf{t}(\mathbf{x}|\mathbf{y}). \tag{2.64}$$

In this case, $\mathcal{F}$ is an invariant distribution of the Markov chain.

Application to Monte Carlo integration: Markov Chain Monte Carlo (MCMC) methods are other extensions of importance sampling algorithms when i.i.d. simulation is too costly or when the selection of an appropriate auxiliary density is not a trivial task [Hastings, 1970; Liu, 2001; Gilks et al., 1999; Gamerman and Lopes, 2006]. The general idea of such methods can be formalized as follows: even if it is difficult to draw directly independent samples from a complicated distribution, one can often find a way of constructing a Markov chain whose stationary distribution is the target law so that, after a sufficient number of iterations, the samples drawn by the MCMC algorithm follow the distribution of interest. In the context of Bayesian estimation, the problem reduces to exploring the space from an initial state $\mathbf{x}^{(0)}$ according to a well chosen transition probability $\mathcal{T}$ such that the Markov chain admits the posterior law as the unique equilibrium distribution. In that respect, the transition probability should obey the following additional conditions:

- *Positive recurrence*: There is a non-zero probability that we will return to any state $\mathbf{x}^{(t)}$ for all $t \geqslant 0$ an infinite number of times.

- *Irreducibility*: For any state of the Markov chain, there is a positive probability of visiting all other states. Formally, a Markov chain is said to be π -irreducible if, for all $\mathbf{x} \in \mathbb{D}$ such that $\pi(\mathbf{x}) > 0$, we have $\mathsf{t}(\mathbf{x}|\mathbf{y}) > 0$ for every $\mathbf{y} \in \mathbb{D}$. In particular, if π is the target invariant density, the chain is also positive recurrent.

- *Aperiodicity*: The chain should not get trapped into cycles.

Under these assumptions, it has been shown that for $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ almost all $\mathbf{x}^{(0)}$ [Jarner and Hansen, 2000; Tierney, 1994; Meyn and Tweedie, 2012]

$$\lim_{t\to+\infty} \|\mathcal{T}^{(t)}(.|\mathbf{x}^{(0)}) - \mathcal{P}_{\mathbf{x}|\mathbf{z}}(.)\|_{TV} = 0 \tag{2.65}$$

where $\|.\|_{TV}$ is the total variation distance defined for any signed measures μ_1 and μ_2 by:

$$\|\mu_1 - \mu_2\|_{TV} = 2\sup_{\mathcal{A}\subset\mathbb{D}} |\mu_1(\mathcal{A}) - \mu_2(\mathcal{A})|. \tag{2.66}$$

It is worth noting that a considerable attention has also been paid to the convergence rate which determines the speed at which the Markov chain approaches its equilibrium distribution [Gibbs, 2000; Barndorff-Nielsen and Kluppelberg, 2000; Mengersen and Tweedie, 1996; Brooks and Roberts, 1998; Roberts and Tweedie, 1996; Roberts et al., 1997; Jarner and Hansen, 2000]. This convergence rate depends on both the target distribution and the choice of the transition probability. In particular, a Markov chain that converges to its invariant distribution at a geometric rate is said to be geometrically ergodic [Kendall, 2004].

In the following, for every positive function $V : \mathbb{D} \to [1, +\infty[$, we define the V-norm for any function f as follows

$$|f|_V = \sup_{\mathbf{x}\in\mathbb{D}} \frac{|f(\mathbf{x})|}{V(\mathbf{x})}. \tag{2.67}$$

The V- distance between two measures μ_1 and μ_2 is given by:

$$\|\mu_1 - \mu_2\|_V = \sup_{f:|f|_V\leqslant 1} |\mu_1(f) - \mu_2(f)| \tag{2.68}$$

where, for every signed measure μ, $\mu(f) = \int_{\mathbb{R}^Q} f(\mathbf{y})\mu(\mathrm{d}\mathbf{y})$. It can be noted that the V- distance coincides with the TV distance for $V = 1$.

Definition 3.5 *Geometric ergodicity* [Mengersen and Tweedie, 1996]
The Markov chain with transition probability $\mathcal{T}$ and stationary distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ is geometrically ergodic if there exist two finite constants $R > 0$ and $\rho \in (0, 1)$ and a positive finite function $V : \mathbb{D} \to [1, +\infty[$ such that

$$(\ \forall t \geqslant 0) \quad \sup_{\mathbf{x}^{(0)}\in\mathbb{D}} \frac{\|\mathcal{T}^{(t)}(.|\mathbf{x}^{(0)}) - \mathcal{P}_{\mathbf{x}|\mathbf{z}}(.)\|_V}{V(\mathbf{x})} \leqslant R\rho^t. \tag{2.69}$$

The main interest in geometric ergodicity is related to the 3 following facts at least:

- A geometrically ergodic Markov chain allows to achieve accurate simulation results in finite time [Johnson, 2009],
- It enables to approach efficiently integrals by a suitable empirical average, by relying on the central limit theorem [Jones, 2004],

- It is required for consistent estimation of Monte Carlo standard errors [Flegal et al., 2008; Hobert et al., 2002; Jones et al., 2006].

To establish geometric ergodicity, it is common to use Foster-Lyapunov drift conditions.

Definition 3.6 *Geometric drift*
$\mathcal{T}$ is said to have geometric drift towards a set $\mathcal{S}$ if there exists a function $V : \mathbb{D} \to [1, +\infty[$ and constants $\lambda < 1$ and β such that

$$(\forall \mathbf{x} \in \mathbb{D}) \;\; \mathcal{T}V(\mathbf{x}) \leqslant \lambda V(\mathbf{x}) + \beta 1_{\mathcal{S}}(\mathbf{x}) \tag{2.70}$$

where

$$(\forall \mathbf{x} \in \mathbb{D}) \quad \mathcal{T}V(\mathbf{x}) = \int_{\mathbb{R}^Q} \mathrm{t}(\mathbf{y}|\mathbf{x}) V(\mathbf{y}) \mathrm{d}\mathbf{y}. \tag{2.71}$$

In order to prove the geometric ergodicity, we need the notion of small sets.

Definition 3.7 *Small sets*
A set $\mathcal{S} \subset \mathbb{D}$ is μ-small if there exists a probability measure μ such that:

$$(\forall \mathbf{x} \in \mathcal{S})(\forall \mathcal{B} \subset \mathbb{D}) \quad \mathcal{T}(\mathcal{B}|\mathbf{x}) \geqslant \mu(\mathcal{B}). \tag{2.72}$$

Theorem 3.8 *[Mengersen and Tweedie, 1996]*
We assume that $\mathcal{T}$ is a π-irreducible, aperiodic transition probability with invariant distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$. Then, if $\mathcal{T}$ has geometric drift towards a small set $\mathcal{S}$ with drift finite function V satisfying (2.70) then V satisfies (2.69) i.e., the Markov chain with transition probability $\mathcal{T}$ is geometrically ergodic.

Convergence inspection of MCMC methods: It can be noted that a practical issue concerning the use of MCMC methods is the convergence inspection to determine when we can stop the MCMC run and be certain that the obtained samples follow the target distribution. Generally, in practice, we throw away some iterations at the beginning of an MCMC run which corresponds to the burn-in period. Then, if the latter is long enough and if the Markov chain is ensured to converge theoretically, all generated samples after the burn-in period should follow the target distribution. Moreover, other various convergence diagnoses may also be used to test whether the Markov chain is exploring the target law after the burn-in period [Gelman and Shirley, 2011; Cowles and Carlin, 1996; Brooks and Gelman, 1998; Brooks and Roberts, 1998]. For example, we can rely on the visual convergence inspection such as the trace plots analysis to detect stability which is known as “the times series plots". Rather than just the values of the target signal, we can consider the running mean, that is, we compute the mean

of samples $\mathbf{x}^{(iT+1)}, \ldots, \mathbf{x}^{((i+1)T)}$ drawn in each period of time T. A time series plot of the running mean can in many cases be more informative and consistent than simply the plots of the values of the target parameters. The convergence is detected when the running means are stabilized at the posterior mean [Smith, 2005]. Others diagnoses are based on the comparison of samples obtained from the MCMC algorithms for different independent runs. In this case, convergence is attained when the difference between the statistics such as the mean and the variance of the obtained samples from these different runs is negligible [Gelman and Rubin, 1992]. It is worthwhile to note that these diagnostics are only reliably used to detect a lack of convergence but not to prove convergence [Cowles and Carlin, 1996; Brooks, 1996].

After having obtained a set of P samples following the target stationary distribution at convergence, one may wonder about how much information we actually have about the target distribution from these samples. In fact, if there are some correlation between the successive states of the Markov chain, one may expect that the generated samples do not reveal as much information of the posterior distribution as independent samples could provide. In that respect, it is also recommended to evaluate the quality of the Markov chain in terms of correlation at convergence. For instance, one can use the integrated autocorrelation time (ACT) that represents the effective number of dependent samples that is equivalent to a single independent sample. It is defined as follows:

$$\text{ACT} = 1 + 2\sum_{i=1}^{+\infty} \gamma(i) \tag{2.73}$$

where $\gamma(i)$ is the lag i autocorrelation which defines the correlation between states of the samples distant by i [Geyer, 1992]. Thus, the smaller the ACT, the better the mixing properties of the chain are. In particular, the effective sample size defined by:

$$\text{ESS} = \frac{P}{\text{ACT}} \tag{2.74}$$

gives an idea about the number of independent samples that yields the same estimation than the P generated correlated samples. Note that, for a given MCMC run, the ACT may differ from one parameter to another. Alternatively, the Mean Squared Jump (MSJ) distance is another indicator of how well the Markov chain is mixing within the target posterior probability distribution. It is computed over all the parameters as follows:

$$\text{MSJ} = \left(\mathsf{E}_{\mathbf{x}|\mathbf{z}}\left(\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2\right)\right)^{1/2}. \tag{2.75}$$

Note that maximizing the MSJ is equivalent to minimizing a weighted sum of the lag 1 autocorrelations [Sherlock et al., 2010].

In the following, we will give two examples of the most commonly used MCMC methods namely Metropolis Hastings and Gibbs algorithms.

3.3.2 Metropolis Hastings algorithm

This algorithm has been firstly introduced by [Metropolis et al., 1953] and then generalized to a more statistical setting by [Hastings, 1970]. In order to draw a sample from a target distribution $\mathsf{p}(\mathbf{x}|\mathbf{z})$, a sample is generated according to some proposal distribution of density $\mathsf{g}(.|\mathbf{x}^{(t)})$ that may depend on the current state $\mathbf{x}^{(t)}$ at each iteration t and should be easy to simulate. The proposed variable is then accepted or rejected according to an acceptance probability. The main steps of this method are summarized in Algorithm 2. Note that, the probability density of moving from a state $\mathbf{x}$ to another state

Algorithm 2 Metropolis Hastings algorithm

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$

for $t = 0, 1, \ldots,$ **do**
  Generate $\tilde{\mathbf{x}}^{(t)}$ from the proposal distribution of density $\mathsf{g}(\cdot|\mathbf{x}^{(t)})$
  Acceptance-Rejection:
  Generate $u \sim \mathcal{U}(0,1)$
  Compute

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

  if $u < \alpha(\mathbf{x}^{(t)}, \tilde{\mathbf{x}}^{(t)})$ **then**
    Accept: $\mathbf{x}^{(t+1)} = \tilde{\mathbf{x}}^{(t)}$
  else
    Reject: $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$
  end if
end for

$\mathbf{y}$ is given by

$$\mathsf{m}(\mathbf{x}, \mathbf{y}) = \mathsf{g}(\mathbf{y}|\mathbf{x})\alpha(\mathbf{x}, \mathbf{y}) \tag{2.76}$$

and the probability density of remaining in the same state $\mathbf{x}$ is given by:

$$\mathsf{r}(\mathbf{x}) = \int_{\mathbb{R}^Q} \mathsf{g}(\mathbf{y}|\mathbf{x})\,(1 - \alpha(\mathbf{x}, \mathbf{y}))\,\mathrm{d}\mathbf{y}. \tag{2.77}$$

Then, the transition density has the following expression:

$$\left(\forall\,(\mathbf{x}, \mathbf{y}) \in (\mathbb{R}^Q)^2\right)\ \mathsf{t}(\mathbf{y}|\mathbf{x}) = \mathsf{m}(\mathbf{x}, \mathbf{y}) + \mathsf{r}(\mathbf{x})\delta(\mathbf{x} - \mathbf{y}). \tag{2.78}$$

Thanks to the accept-reject step, the transition probability of density (2.78) is reversible. Then, $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ is an invariant distribution of the MH Markov chain. Furthermore, if the support of the proposal density is large enough to explore all the domain of the target density (i.e., for every $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^Q$ such that $\mathsf{p}(\mathbf{x}|\mathbf{z}) > 0$ and $\mathsf{p}(\mathbf{y}|\mathbf{z}) > 0$, we have $\mathsf{g}(\mathbf{x}|\mathbf{y}) > 0$), the Markov chain

is also $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$-irreducible. Moreover, if there exist $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^Q$ satisfying $\mathsf{p}(\mathbf{x}|\mathbf{z}) > 0$ and $\mathsf{p}(\mathbf{y}|\mathbf{z}) > 0$ such that $\mathsf{g}(\mathbf{x}|\mathbf{y}) > 0$ and $\alpha(\mathbf{x}, \mathbf{y}) < 1$, then there exists a positive probability such that the chain remains in the state $\mathbf{x}$ (i.e., $\mathsf{r}(\mathbf{x}) > 0$). It follows that the Markov chain is also aperiodic [Mengersen and Tweedie, 1996]. Thereby, samples generated by the MH algorithm follow, at equilibrium, the distribution of interest.

Similarly to rejection sampling, the target distribution is also needed to be only known up to a multiplicative factor. However, unlike rejection sampling, MH algorithms offer more flexibility on the choice of the proposal distribution. The only theoretical requirement is that we should be able to compute the ratio $\mathsf{p}(\mathbf{x}|\mathbf{z})/\mathsf{g}(\mathbf{y}|\mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^Q$ up to a multiplicative constant. For instance, it is feasible theoretically, to select the proposal distribution such that $\mathsf{p}(\mathbf{x}|\mathbf{z})/\mathsf{g}(\mathbf{y}|\mathbf{x})$ is not bounded. However, when this is true, the performance of the MH algorithm may be degraded in practice, as detailed in [Casella et al., 2004]. In particular, independent proposals can be used where the proposal density does not depend on the past states of the chain i.e., $\mathsf{g}(\mathbf{y}|\mathbf{x}) = \mathsf{g}(\mathbf{y})$, for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^Q$. In this case, MH algorithm appears as a straightforward generalization of the rejection sampling method when the independent proposal density is the same as the auxiliary density in the rejection method. However, the generated samples by the MH algorithm will involve repeated occurrences of the same value corresponding to the rejected states on the chain which results on correlated samples. Reciprocally, another well common approach for the practical construction of the proposal is to take the current value of the chain into account to generate the next state. This amounts to considering a local exploration of the space around the current state. Thus, at each iteration t, we propose a value according to the following scheme

$$\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)} + \boldsymbol{\omega}^{(t)} \tag{2.79}$$

where $\boldsymbol{\omega}^{(t)}$ is a random perturbation independent of $\mathbf{x}^{(t)}$. For instance, it can be drawn from a uniform or Gaussian distribution. In this case, the proposal density is symmetric and the MH algorithm is referred to as the random walk [Roberts et al., 1997]. The acceptance probability reads

$$\alpha(\mathbf{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}) = \min\left(1, \frac{\mathsf{p}(\tilde{\mathbf{x}}^{(t)}|\mathbf{z})}{\mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z})}\right). \tag{2.80}$$

It follows that, every move to a more probable state is accepted with probability 1 while moves to less probable states are accepted with a probability $\frac{\mathsf{p}(\tilde{\mathbf{x}}^{(t)}|\mathbf{z})}{\mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z})} < 1$. In order to increase the acceptance probability and then decrease the number of rejected values, one should propose small moves at each iteration. However, as the newly accepted value depends on the previous state in the chain through the proposal density, one may instead prefer high moves in order to have a good mixing behavior and then a fast exploration of the space. The perturbation in the random walk algorithm should

then be set to find the ideal tradeoff between obtaining a large acceptance probability and good mixing properties of the chain. In the following, we give an example of applications of MH algorithms.

Example 3.9 *Sampling from a $\mathcal{B}$ distribution using MH algorithm*:
Let us consider Example 3.1. For illustration purposes, we set $a = b = 5$ and, we propose to use the independent MH sampling algorithm with a uniform distribution over $[0, 1]$, which means that the proposal density does not depend on the past values of the chain. We run the MH algorithm for $6,000$ iterations and discard the first $1,000$ as burn-in period. The average acceptance probability is 0.49. For the rejection sampling, the acceptance rate is $p = 0.40$. In Figure 2.10, we display the trace plot of the last 300 generated samples by both algorithms, the resulting histograms computed over $5,000$ samples and the autocorrelation function of the samples. It can be noted that, in this setting, the independent MH algorithm performs very similarly to the rejection algorithm since both histograms properly approximate the target density. The only difference is that, the generated samples are now correlated. This correlation originates from the rejected states in the MH algorithm as there are multiple occurrences of the same value in the sample.

As it has already been discussed in Example 3.1, for $a = b = 100$, the rejection algorithm is slow since the acceptance rate is very small (we need around 11 iterations of the rejection sampling to generate one sample). Similarly, when using MH algorithm with uniform proposal over $[0, 1]$, we obtain an acceptance probability of around 0.11 (which is close to the acceptance rate of rejection sampling). Therein, we propose to employ random walk methods with two different proposals. First, we use a uniform distribution centered on the current value. In this case, a new value is proposed at each iteration according to $\tilde{\mathbf{x}}^{(t)} \sim \mathcal{U}\left(\mathbf{x}^{(t)} - \delta, \mathbf{x}^{(t)} + \delta\right)$ where $\delta > 0$. Second, we use a Gaussian distribution centered on the current state i.e, $\tilde{\mathbf{x}}^{(t)} \sim \mathcal{N}(\mathbf{x}^{(t)}, \varepsilon^2)$ where $\varepsilon > 0$. We set $\delta = 0.1$ and $\varepsilon = 0.06$ which correspond to an acceptance probability around 0.5. Figure 2.11 shows that the two proposal distributions achieve similar results and that the generated samples fit the target distribution.

In order to study the influence of the proposal on the efficiency of the sampling process, we compare the samples of the Gaussian random walk run with two other different values of the tuning parameters $\varepsilon = (0.001, 0.5)$ that achieve acceptance probabilities around $(0.99, 0.08)$ respectively. Figure 2.12 shows the difference in the produced chains: Too small or too large moves (that is, a small or a large value of ε) result in higher correlation and slower convergence. In fact, in the former case, the Markov chain moves at each iteration but very slowly, while in the latter, it remains constant over long periods of time.

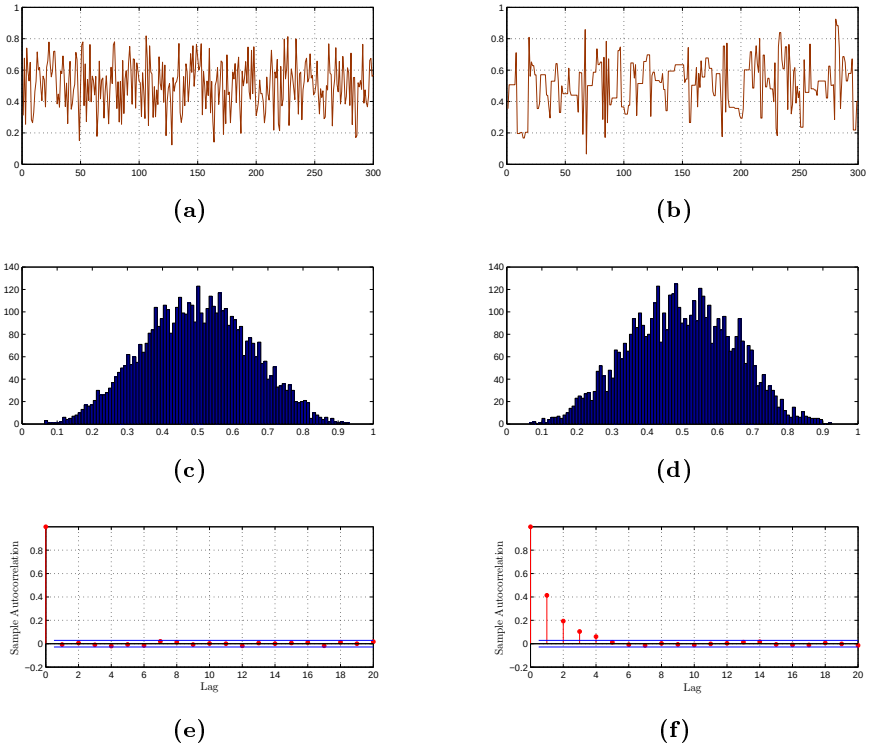


Figure 2.10: *Trace Sample, histograms and sample autocorrelation. The left column corresponds to the Rejection Sampling algorithm, the right one to independent MH algorithm with Uniform proposal.*

As it has been highlighted in Example (3.9), the performance of the MH algorithm is obviously strongly related to the choice of the proposal distribution. This issue becomes especially critical in large scale problems. In general, when selecting a proposal in MH algorithms, one should consider two issues. First, whilst MH algorithms are guaranteed to yield samples from the target distribution after a sufficient burn-in period, the number of iterations required to reach convergence can be infeasibly large. Second, unlike rejection sampling algorithms, the generated samples in convergence are correlated. This correlation originates from two main sources: the correlation introduced by retaining the same value because the newly generated value has been rejected and the correlation between successive samples for non-independent proposals. A poorly mixed chain tends to generate samples that are highly correlated which lead to an incomplete summary of the target distribution and highly biased estimators. Consequently, we need more samples to achieve the same precision as i.i.d methods. In [Roberts and Rosenthal, 2001], the efficiency of MH algorithms is discussed with respect to the acceptance probability. In general, a good proposal should be a good

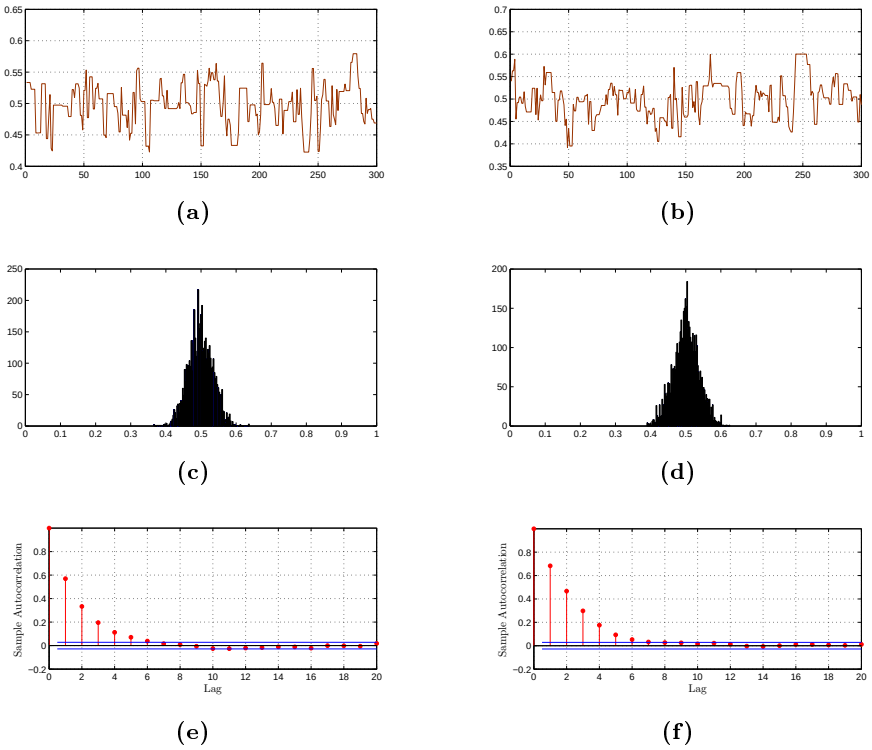


Figure 2.11: *Trace Sample, histograms and sample autocorrelation using a random walk. The left column corresponds to the uniform proposal with* $\delta = 0.1$*, the right one to the Gaussian proposal with* $\varepsilon = 0.5$*.*

approximation or good local approximation of the target density without being costly to sample from. In particular, it should reflect the dependence structure of the target distribution for large scale problems. In this respect, a large amount of works has been devoted to construct proposals in MH algorithms in attempt to meet these requirements [Roberts and Tweedie, 1996; Roberts and Rosenthal, 2009; Stuart et al., 2004; Roberts and Rosenthal, 1998; Vacar et al., 2011; Girolami and Calderhead, 2011; Zhang and Sutton, 2011; Martin et al., 2012; Pillai et al., 2012].

3.3.3 Gibbs sampler

Suppose that, we can decompose $\mathbf{x} \in \mathbb{R}^Q$, into variables or blocks of variables $\mathbf{x}_i \in \mathbb{R}^{Qi}$, $1 \leqslant i \leqslant R$, where $\sum_{i=1}^{R} Q_i = Q$ so that the conditional distribution of each $\mathbf{x}_i$ given $\mathbf{z}$ and the remaining variables $\mathbf{x}_j$, $j \neq i$ is simple to sample from. This strategy often leads to important efficiency gains, particularly if the involved conditional densities are simpler than the joint density, in the sense that it is computationally straightforward to draw samples from these

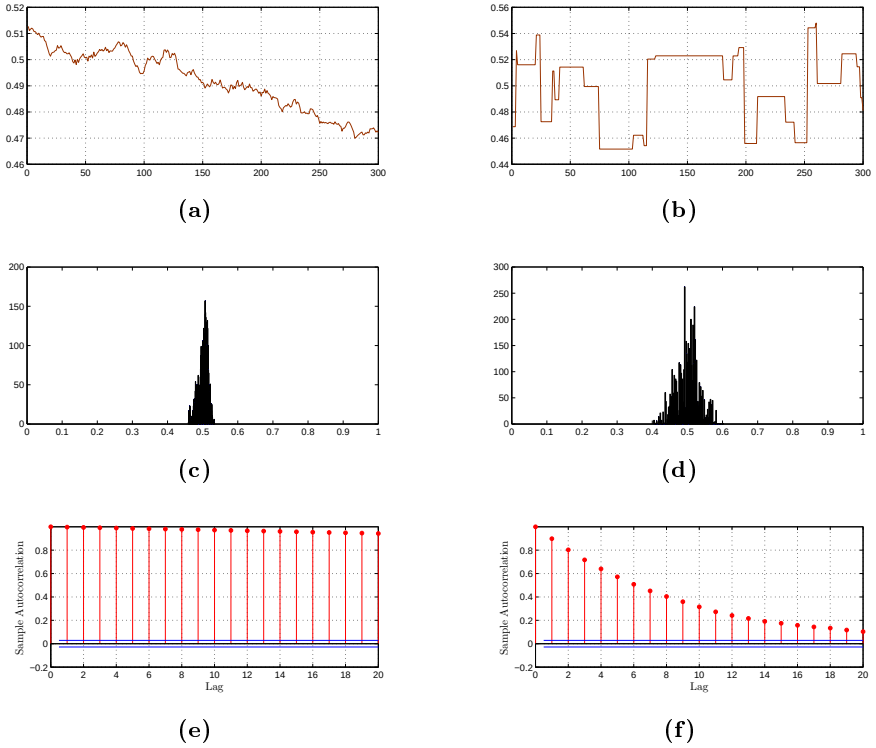


Figure 2.12: *Trace Sample, histograms and sample autocorrelation using Gaussian Random Walk. The left column corresponds to* $\varepsilon = 0.001$*, the right one* $\varepsilon = 0.06$*.*

conditional distributions rather than from the joint law. In this case, the Gibbs algorithm is a MCMC technique suitable for this task [Casella and George, 1992; Geman and Geman, 1984]. The idea behind Gibbs sampling is to generate posterior samples by sweeping through each of these variables $\mathbf{x}_i \in \mathbb{R}^{Qi}$ and to sample from its conditional distribution while keeping the remaining variables fixed to their current values. The main steps are given in Algorithm 3. The generated Markov chain resulting from concatenating all these variables admits $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ as invariant distribution, and thus the samples produced by the Gibbs sampler follow the desired distribution after a sufficient burn-in period [Robert, 2013]. MH steps can also be added to this algorithm. That is, when the full conditional distributions are available and belong to the family of standard distributions (Gamma, Gaussian, etc.), we can draw the new samples directly. Otherwise, we can draw samples using MH steps embedded within the Gibbs algorithm [Gilks et al., 1995]. This is known as the Hybrid Gibbs sampler. The convergence of the Markov chain to the desired invariant distribution is also guaranteed [Gilks et al., 1999]. However, this may result in a deterioration of the algorithm convergence

Algorithm 3 Gibbs Sampler

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$
  for $t = 0, 1, \ldots,$ **do**
    Generate $\mathbf{x}_1^{(t+1)} \sim \mathcal{P}_{\mathbf{x}_1|\mathbf{z},\mathbf{x}_2^{(t)},\ldots,\mathbf{x}_R^{(t)}}$.
    Generate $\mathbf{x}_2^{(t+1)} \sim \mathcal{P}_{\mathbf{x}_2|\mathbf{z},\mathbf{x}_1^{(t+1)},\mathbf{x}_3^{(t)},\ldots,\mathbf{x}_R^{(t)}}$.
    $\vdots$
    Generate $\mathbf{x}_R^{(t+1)} \sim \mathcal{P}_{\mathbf{x}_R|\mathbf{z},\mathbf{x}_1^{(t+1)},\ldots,\mathbf{x}_{R-1}^{(t+1)}}$.
  end for

rate. Note that the Gibbs sampler algorithm is well adapted to hierarchical Bayesian models where we have more than one unknown variable to be estimated (as acquisition parameters, prior parameters and hyperparameters) [Damlen et al., 1999].

§ 4 Approximation methods

4.1 Laplace approximation

This method is based on a local approximation of the target distribution by a Gaussian distribution around the MAP estimate. Let $\hat{\mathbf{x}}_{MAP}$ denotes a mode of the true posterior distribution. The Laplace approximation is given by the Gaussian distribution of mode $\hat{\mathbf{x}}_{MAP}$ and whose covariance matrix is given by the inverse of the Hessian matrix of posterior distribution minus logarithm computed at $\hat{\mathbf{x}}_{MAP}$.

This approximation has several shortcomings. First, the approximate density requires the computation of the derivatives of the posterior density logarithm and the inversion of a high dimensional matrix which may become burdensome for large scale problems. Besides, the second derivatives themselves may be intractable to compute. Second, the Gaussian assumption may not be suitable for example to bounded, constrained, or positive parameters and multimodal distributions. Furthermore, even if the exact posterior is unimodal, the approximate distribution may fail to well represent the whole probability density, as the approximation is only made up locally around the maximum. The asymptotic error of this approximation was studied in [Kass and Raftery, 1995].

4.2 Variational Bayes approximation

The variational Bayes approximation (VBA) approach has been first introduced in physics [Parisi, 1988]. The idea behind it is to approximate the true posterior density $\mathsf{p}(\mathbf{x}|\mathbf{z})$ with another distribution density denoted by $\mathsf{q}(\mathbf{x})$

which is as close as possible to $\mathsf{p}(\mathbf{x}|\mathbf{z})$, by minimizing the Kullback-Leibler divergence between them [Šmídl and Quinn, 2005; Pustelnik et al., 2016; Pereyra et al., 2016; Grimmer, 2010]:

$$\mathsf{q}^{\text{opt}} = \underset{\mathsf{q}(\mathbf{x})\in\mathbb{Q}}{\operatorname{argmin}}\ \mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big) \tag{2.81}$$

where $\mathbb{Q}$ is a subset of valid densities on $\mathbf{x}$ and

$$\mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big) = \int_{\mathbb{R}^Q} \mathsf{q}(\mathbf{x}) \log \frac{\mathsf{q}(\mathbf{x})}{\mathsf{p}(\mathbf{x}|\mathbf{z})} \mathrm{d}\mathbf{x}. \tag{2.82}$$

Since the true posterior density is almost known up to a multiplicative constant, the $\mathcal{KL}$ divergence is commonly decomposed as follows:

$$\mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big) = \mathcal{E}(\mathsf{q}(\mathbf{x})) + \log \mathsf{p}(\mathbf{z}) \tag{2.83}$$

where

$$\mathcal{E}(\mathsf{q}(\mathbf{x})) = \int_{\mathbb{R}^Q} \mathsf{q}(\mathbf{x}) \log \frac{\mathsf{q}(\mathbf{x})}{\mathsf{p}(\mathbf{x},\mathbf{z})} \mathrm{d}\mathbf{x} \tag{2.84}$$

is known as the Gibbs free energy or variational free energy. Here above, $\log \mathsf{p}(\mathbf{z})$ can be seen as a constant that does not depend on $\mathsf{q}(\mathbf{x})$. Thus, the minimization of $\mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big)$ can be replaced by the minimization of $\mathcal{E}(\mathsf{q}(\mathbf{x}))$ whose expression only depends on $\mathsf{p}(\mathbf{x},\mathbf{z}) = \mathsf{p}(\mathbf{z}|\mathbf{x})\mathsf{p}(\mathbf{x})$ without requiring the computation of the normalization constant, that is

$$\mathsf{q}^{\text{opt}}(\mathbf{x}) = \underset{\mathsf{q}(\mathbf{x})\in\mathbb{Q}}{\operatorname{argmin}}\ \mathcal{E}(\mathsf{q}(\mathbf{x})). \tag{2.85}$$

Furthermore, since $\mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big)$ is positive we have:

$$-\log \mathsf{p}(\mathbf{z}) = \mathcal{E}(\mathsf{q}(\mathbf{x})) - \mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big) \leqslant \mathcal{E}(\mathsf{q}(\mathbf{x})). \tag{2.86}$$

It follows that, $\mathcal{E}(\mathsf{q}(\mathbf{x}))$ is an upper bound of the negative log of the marginal density of the observation. Thus, in convergence when $\mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big)$ is nearly zero, $\mathcal{E}(\mathsf{q}(\mathbf{x}))$ can approximate $-\log \mathsf{p}(\mathbf{z})$.

Note that $\mathcal{KL}\big(\mathsf{q}(\mathbf{x})\|\mathsf{p}(\mathbf{x}|\mathbf{z})\big)$ is positive and it reaches its minimum at zero for $\mathsf{q}(\mathbf{x}) = \mathsf{p}(\mathbf{x}|\mathbf{z})$. However, this is not particularly helpful because the posterior is generally intractable. Hence, we introduce additional assumptions on the approximating density $\mathsf{q}(\mathbf{x})$ in order to make inference tractable while also providing a close approximation to the true density. Note that one of the main difficulties in the minimization of either (2.81) or (2.85) is the mutual dependence between different unknowns. In particular, solutions of such problems become tractable if a suitable factorization structure of $\mathsf{q}(\mathbf{x})$ is assumed. Therefore, the true posterior needs to be approximated by a separable distribution which facilitates the calculation of the estimators. That is, we assume that $\mathsf{q}(\mathbf{x}) = \prod_{i=1}^{R} \mathsf{q}_{\mathbf{X}_i}(\mathbf{x}_i)$ with $R \leqslant Q$. When $Q = R$,

we have a total separability, otherwise, the separability is partial. Hence, the optimal density approximation $\mathsf{q}_{\mathbf{X}_i}^{\text{opt}}(\mathbf{x}_i)$ for each variable $\mathbf{x}_i$, is obtained by minimizing the $\mathcal{KL}$ divergence while holding the remaining densities with the rest of variables fixed. In this case, there exists an optimal solution to the optimization problem (2.81) for each density $(\mathsf{q}_{\mathbf{X}_i}(\mathbf{x}_i))_{1\leq i\leq R}$, given by the exponential of the expectation of the joint density with respect to the distribution of all the unknown parameters except the one of interest i.e.,

$$(\forall i \in \{1,\ldots,R\}) \;\; \mathsf{q}_{\mathbf{X}_i}^{\text{opt}}(\mathbf{x}_i) \propto \exp\left(\langle \log \mathsf{p}(\mathbf{z},\mathbf{x})\rangle_{\prod_{j\neq i}\mathsf{q}_{\mathbf{X}_j}^{\text{opt}}(\mathbf{x}_j)}\right) \tag{2.87}$$

where

$$\langle\cdot\rangle_{\prod_{j\neq i}\mathsf{q}_{\mathbf{X}_j}(\mathbf{x}_j)} = \int \cdot \prod_{j\neq i} \mathsf{q}_{\mathbf{X}_j}(\mathbf{x}_j)\mathrm{d}\mathbf{x}_j. \tag{2.88}$$

Due to the implicit relationships existing between $\left(\mathsf{q}_{\mathbf{X}_i}^{\text{opt}}(\mathbf{x}_i)\right)_{1\leq i\leq R}$, an analytical expression of $\mathsf{q}^{\text{opt}}(\mathbf{x})$ generally does not exist. Usually, these distributions are determined in an iterative way, by updating one of the separable components $(\mathsf{q}_{\mathbf{X}_i}(\mathbf{x}_i))_{1\leq i\leq R}$ while fixing the others [Šmídl and Quinn, 2005]. Applications of classical VBA approaches can be found in [Drémeau et al., 2012; Babacan et al., 2011; Chen et al., 2014; Tramel et al., 2016; McGrory and Titterington, 2009; Forbes and Fort, 2007] while improved VBA algorithms have been proposed in [Fraysse and Rodet, 2014; Zheng et al., 2015]. Once the approximate distributions are computed, the unknown parameters are then estimated by the means of the obtained distributions. In the following, we give an example of application of VBA.

Example 4.1 *VBA method for the separable approximation of a Gaussian posterior distribution*
Consider the observation model in (2.2), where $\mathbf{w} = (w)_{1\leqslant i\leqslant N}$ is Gaussian noise of unknown variance τ. Then,

$$\mathsf{p}(\mathbf{z}|\mathbf{x},\tau) = (2\pi\tau)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\tau}\|\mathbf{H}\mathbf{x}-\mathbf{z}\|^2\right). \tag{2.89}$$

We also suppose that the prior density of $\mathbf{x}$ reads:

$$\mathsf{p}(\mathbf{x}) = C\gamma^{-\frac{Q}{2}} \exp\left(-\frac{\gamma}{2}\|\mathbf{\Lambda}\mathbf{x}\|^2\right) \tag{2.90}$$

where $\gamma > 0$ is the unknown regularization parameter and $C > 0$ is a constant independent of γ. We also use a non-informative prior for the unknown parameters that is,

$$\mathsf{p}(\gamma) = \frac{1}{\gamma} \tag{2.91}$$

and

$$\mathsf{p}(\tau) = \frac{1}{\tau}. \tag{2.92}$$

It follows that the joint posterior density of $\mathbf{x}$, γ, τ is proportional to

$$\mathsf{p}(\mathbf{x},\gamma,\tau,\mathbf{z}) = \exp\left(-\frac{\gamma}{2}\|\mathbf{\Lambda x}\|^2\right)\exp\left(-\frac{1}{2\tau}\|\mathbf{Hx}-\mathbf{z}\|^2\right)\tau^{-\frac{N}{2}-1}\gamma^{\frac{Q}{2}-1}. \quad (2.93)$$

Let us approximate this true posterior density via VBA method using partial or total separability.

VBA method using partial separability: We approximate $\mathsf{p}(\mathbf{x},\gamma,\tau|\mathbf{z})$ with another density $\mathsf{q}(\mathbf{x},\gamma,\tau)$ such that:

$$\mathsf{q}(\mathbf{x},\gamma,\tau) = \mathsf{q}_{\mathbf{X}}(\mathbf{x})\mathsf{q}_{\Gamma}(\gamma)\mathsf{q}_{T}(\tau). \quad (2.94)$$

Using (2.87), the density $\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})$ at the iteration $k+1$ of the algorithm is given by

$$\begin{aligned}\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x}) &\propto \exp\left(\langle\log\mathsf{p}(\mathbf{x},\gamma,\tau,\mathbf{z})\rangle_{\mathsf{q}_{\Gamma}^{(k)}(\gamma)\mathsf{q}_{T}^{(k)}(\tau)}\right)\\ &\propto \exp\left(-\frac{\langle\tau^{-1}\rangle^{(k)}}{2}\|\mathbf{Hx}-\mathbf{z}\|^2-\frac{\langle\gamma\rangle^{(k)}}{2}\|\mathbf{\Lambda x}\|^2\right) \quad (2.95)\end{aligned}$$

where $\langle\tau^{-1}\rangle^{(k)} = \int_{\mathbb{R}}\tau^{-1}\mathsf{q}_{T}^{(k)}(\tau)\mathrm{d}\tau$ and $\langle\gamma\rangle^{(k)} = \int_{\mathbb{R}}\gamma\mathsf{q}_{\Gamma}^{(k)}(\gamma)\mathrm{d}\gamma$. Then, $\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})$ is the density of a Gaussian distribution with mean $\mathbf{m}^{(k+1)}$ and covariance matrix $\mathbf{\Sigma}^{(k+1)}$ given by

$$\mathbf{m}^{(k+1)} = \mathbf{\Sigma}^{(k+1)}\langle\tau^{-1}\rangle^{(k)}\mathbf{H}^{\top}\mathbf{z} \quad (2.96)$$

and

$$\left(\mathbf{\Sigma}^{(k+1)}\right)^{-1} = \langle\tau^{-1}\rangle^{(k)}\mathbf{H}^{\top}\mathbf{H}+\langle\gamma\rangle^{(k)}\mathbf{\Lambda}^{\top}\mathbf{\Lambda}. \quad (2.97)$$

The density $\mathsf{q}_{\Gamma}^{(k+1)}(\gamma)$ is given by

$$\begin{aligned}\mathsf{q}_{\Gamma}^{(k+1)}(\gamma) &\propto \exp\left(\langle\log\mathsf{p}(\mathbf{x},\gamma,\tau,\mathbf{z})\rangle_{\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})\mathsf{q}_{T}^{(k)}(\tau)}\right)\\ &\propto \gamma^{\frac{Q}{2}-1}\exp\left(-\frac{\gamma}{2}\left\langle\|\mathbf{\Lambda x}\|^2\right\rangle^{(k+1)}\right) \quad (2.98)\end{aligned}$$

where

$$\begin{aligned}\left\langle\|\mathbf{\Lambda x}\|^2\right\rangle^{(k+1)} &= \int_{\mathbb{R}^Q}\|\mathbf{\Lambda x}\|^2\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \|\mathbf{\Lambda m}^{(k+1)}\|^2+\mathrm{trace}(\mathbf{\Sigma}^{(k+1)}\mathbf{\Lambda}^{\top}\mathbf{\Lambda}). \quad (2.99)\end{aligned}$$

(2.98) can be seen as the density of the Gamma distribution of parameters

$$a_{\gamma} = \frac{Q}{2} \quad (2.100)$$

and

$$b_{\gamma}^{(k+1)} = \frac{1}{2}\left(\|\mathbf{\Lambda}\mathbf{m}^{(k+1)}\|^2 + \text{trace}(\mathbf{\Sigma}^{(k+1)}\mathbf{\Lambda}^\top\mathbf{\Lambda})\right). \tag{2.101}$$

It follows that

$$\langle\gamma\rangle^{(k+1)} = \frac{a_\gamma}{b_\gamma^{(k+1)}}. \tag{2.102}$$

Similarly, we have

$$\begin{aligned}\mathsf{q}_T^{(k+1)}(\tau) &\propto \exp\left(\langle\log \mathsf{p}(\mathbf{x},\gamma,\tau,\mathbf{z})\rangle_{\mathsf{q}_X^{(k+1)}(\mathbf{x})\mathsf{q}_\Gamma^{(k+1)}(\gamma)}\right)\\ &\propto \tau^{-\frac{N}{2}-1}\exp\left(-\frac{1}{2\tau}\left\langle\|\mathbf{H}\mathbf{x}-\mathbf{z}\|^2\right\rangle^{(k+1)}\right)\end{aligned} \tag{2.103}$$

where

$$\begin{aligned}\left\langle\|\mathbf{H}\mathbf{x}-\mathbf{z}\|^2\right\rangle^{(k+1)} &= \int_{\mathbb{R}^Q}\|\mathbf{H}\mathbf{x}-\mathbf{z}\|^2\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \|\mathbf{H}\mathbf{m}^{(k+1)}-\mathbf{z}\|^2 + \text{trace}(\mathbf{\Sigma}^{(k+1)}\mathbf{H}^\top\mathbf{H}).\end{aligned} \tag{2.104}$$

(2.103) can be seen as the density of the inverse Gamma distribution of parameters

$$a_\tau = \frac{N}{2} \tag{2.105}$$

and

$$b_\tau^{(k+1)} = \frac{1}{2}\left(\|\mathbf{H}\mathbf{m}^{(k+1)}-\mathbf{z}\|^2 + \text{trace}(\mathbf{\Sigma}^{(k+1)}\mathbf{H}^\top\mathbf{H})\right). \tag{2.106}$$

It follows that

$$\langle\tau^{-1}\rangle^{(k+1)} = \frac{a_\tau}{b_\tau^{(k+1)}}. \tag{2.107}$$

Note that the parameters of the densities $\mathsf{q}_\Gamma^{(k+1)}(\gamma)$ and $\mathsf{q}_T^{(k+1)}(\tau)$ depend on the parameters of $\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})$. At each iteration, the computation of the mean of the Gaussian distribution of density $\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x})$ can be fulfilled using an iterative method for solving

$$\left(\mathbf{\Sigma}^{(k+1)}\right)^{-1}\mathbf{m}^{(k+1)} = \langle\tau^{-1}\rangle^{(k)}\mathbf{H}^\top\mathbf{z}. \tag{2.108}$$

However, the problem remains in the computation of the covariance matrix. In [Babacan et al., 2011], the authors propose to approximate it with a diagonal matrix whose elements are given by the inverse of the diagonal of $\left(\mathbf{\Sigma}^{(k+1)}\right)^{-1}$. However, this may induce errors on the estimation.

VBA method using total separability: We now assume that

$$\mathsf{q}(\mathbf{x},\gamma,\tau)=\mathsf{q}_\Gamma(\gamma)\mathsf{q}_T(\tau)\mathsf{q}_{\mathbf{X}}(\mathbf{x})=\mathsf{q}_\Gamma(\gamma)\mathsf{q}_T(\tau)\prod_{i=1}^{Q}\mathsf{q}_{X_i}(x_i). \tag{2.109}$$

Similarly to the partial separability, we can show that by using (2.87), the approximate distribution of density $\mathsf{q}_{X_i}(x_i)$, for every $i\in\{1,\ldots,Q\}$ is a Gaussian distribution with mean $m_i^{(k+1)}$ and variance $(\sigma_i^2)^{(k+1)}$ given by

$$\begin{aligned} m_i^{(k+1)} = (\sigma_i^2)^{(k+1)}\langle\tau^{-1}\rangle^{(k)}\left(\left[\mathbf{H}^\top\mathbf{z}\right]_i-\left[\mathbf{H}^\top\mathbf{H}\tilde{\mathbf{m}}^{(k+1)}\right]_i+\mathrm{diag}\left(\mathbf{H}^\top\mathbf{H}\right)_i\tilde{\mathbf{m}}_i^{(k+1)}\right)\\ -(\sigma_i^2)^{(k+1)}\langle\gamma\rangle^{(k)}\left(\left[\mathbf{\Lambda}^\top\mathbf{\Lambda}\tilde{\mathbf{m}}^{(k+1)}\right]-\mathrm{diag}\left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right)_i\tilde{\mathbf{m}}_i^{(k+1)}\right) \end{aligned} \tag{2.110}$$

and

$$(\sigma_i^{-2})^{(k+1)}=\langle\tau^{-1}\rangle^{(k)}\mathrm{diag}\left(\mathbf{H}^\top\mathbf{H}\right)+\langle\gamma\rangle^{(k)}\mathrm{diag}\left(\mathbf{\Lambda}^\top\mathbf{\Lambda}\right) \tag{2.111}$$

where $\tilde{\mathbf{m}}^{(k+1)}=\left[m_1^{(k+1)},\ldots,m_{i-1}^{(k+1)},m_i^{(k)},m_Q^{(k)}\right]^\top$. The optimization of densities $\mathsf{q}_\Gamma(\gamma)$ and $\mathsf{q}_T(\tau)$ is the same as for the partial separability. Since the covariance matrix of $\mathsf{q}_{\mathbf{X}}(\mathbf{x})$ is now diagonal, its manipulation is easier than in the first case. However, the total separability approximation may be inappropriate when the coefficients $\mathbf{x}$ exhibit high correlation.

It can be noted that VBA methods constitute a powerful tool to compute the MMSE estimator for complicated models by imposing additional properties to approximate them, in order to make the estimation tractable. However, very rough approximation of the posterior density may result on a deterioration of the quality of estimation. Moreover, the VBA procedure cannot be easily implemented for more complicated priors and likelihood models since the direct optimization of non standard distributions is not generally a trivial task.

Conclusion

In this chapter, we have provided an overview of the Bayesian framework for the resolution of inverse problems in signal processing. In such framework, the unknown signal and the observations are both modeled by two random variables. The distribution of the observations is related to the direct observation model whereas the prior distribution of the unknown signal is chosen by the designer. Depending on the amount of information and the properties of the observation law, the prior distribution can be chosen to be highly informative, conjugate or non-informative. The posterior distribution is derived from the prior and the observation models using the Bayes rule.

Bayesian estimators are then computed from this posterior by minimizing a given cost function. Among well known Bayesian estimators, we can mention the MAP and the MMSE estimators. While the MAP can be computed using deterministic minimization algorithms, the derivation of the MMSE estimator requires to calculate an integral that is in most case intractable. Therein, stochastic and approximation methods have been proposed to compute such estimator. Table 2.1 summarizes the advantages/drawbacks of the various methods that we have presented.

Table 2.1: *Comparison of the presented Bayesian methods to compute the MMSE estimator.*

Stochastic simulation methods (in particular MCMC algorithms)	Approximation methods (in particular VBA algorithm)
• Construct a Markov chain whose stationary distribution is the posterior law.	• Find a tractable approximation of the true posterior distribution that is maximally similar to it.
• Derive inferences via empirical estimates computed over the samples in convergence.	• Derive inferences using the approximate distribution.
✗ Computationally expensive.	✓ Relatively faster.
✓ Asymptotically exact.	✗ Approximation errors.
✓ Flexibility of application.	✗ Tractable only for specific laws.

- Chapter 3 -

Majorize-Minimize adapted Metropolis Hastings algorithm

Metropolis Hastings (MH) algorithms are currently the most popular simulation techniques for producing samples from posterior probabilities when direct sampling is not trivial. They consist of building a Markov chain defined by a proposal density and a rejection-acceptance rule whose stationary distribution is the desired posterior law. The asymptotic states of the constructed chain are then considered as samples from the target distribution. However, the choice of the proposal distribution is crucial as it impacts the statistical properties of the resulting Markov chain especially for complicated and high-dimensional target distributions. The proposal density should ideally provide an accurate approximation of the target density with a low computational cost. This problem is often tackled in an empirical manner. However, it is also possible to determine theoretically an optimal proposal scaling [Roberts and Rosenthal, 2001] or to use adaptive algorithms in order to find local approximation of the target distribution automatically [Atchadé, 2006]. One typical algorithm is the Random Walk (RW) algorithm whose adaptive proposal law takes the form of a Gaussian distribution centered at the current state [Roberts et al., 1997]. The popularity of this algorithm is mainly related to its simplicity of implementation. However, the RW usually takes too many steps to reach stability for high dimensional models. Furthermore, slow convergence together with bad mixing behavior could make the Markov chain more likely to get trapped into some regions and thus fail to explore efficiently the whole target space.

As the dimension and the complexity of inference problems have dramatically increased, the design of improved proposal scheme providing large proposal transitions that are accepted with high probability, is required. Intuitively, the proposal density should take advantage of the local properties of the target distribution to accelerate the exploration of regions with high probability values. In this work, we are interested in proposals based on the Euler discretization of the Langevin stochastic differential equation where the drift term accounts for the slope and curvature of the target law. In particular, we propose a preconditioned version of the standard Metropolis Hastings adapted Langevin algorithm using an adaptive matrix based on a Majorize-Minimize strategy.

This chapter is organized as follows: In Section 1, we formulate the problem and we give a brief overview of the Langevin diffusion process. In Section 2, we describe the new Majorize-Minimize adapted MH algorithm. In section 3, a particular attention is paid to the convergence proof of the proposed algorithm. Section 4 is devoted to experimental results.

§ 1 Problem statement and related work

1.1 Langevin diffusion

A Q-dimensional Langevin diffusion is a continuous time Markov process $(\mathbf{x}(t))_{t\in\mathbb{R}_+}$ with values in $\mathbb{R}^Q$ defined as the solution of the following stochastic differential equation [Roberts and Stramer, 2002]:

$$\mathrm{d}\mathbf{x}(t) = \mathbf{b}(\mathbf{x}(t))\mathrm{d}t + \mathbf{V}(\mathbf{x}(t))\mathrm{d}\mathbf{B}(t), \quad \mathbf{x}(0) = \mathbf{x}^{(0)} \tag{3.1}$$

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

$$\begin{aligned}(\forall i \in \{1,\dots,Q\}) \quad b_i(\mathbf{x}) = \frac{1}{2}\sum_{j=1}^{Q} A_{ij}(\mathbf{x})\frac{\partial \log \pi(\mathbf{x})}{\partial x_j} \\ + |\mathbf{A}(\mathbf{x})|^{\frac{1}{2}} \sum_{j=1}^{Q} \frac{\partial}{\partial x_j}\left(A_{ij}(\mathbf{x})|\mathbf{A}(\mathbf{x})|^{-\frac{1}{2}}\right)\end{aligned} \tag{3.2}$$

where $\mathbf{A}(\mathbf{x}) = \mathbf{V}(\mathbf{x})\mathbf{V}(\mathbf{x})^\top$ is a symmetric definite positive matrix and $|\mathbf{A}(\mathbf{x})|$ denotes its determinant. Note that the process is stationary and π is the density of the stationary distribution of the diffusion i.e., if a state $\mathbf{x}(t_0)$ follows the distribution of density π, all subsequent states $\mathbf{x}(t_0+\tau)$, $\tau > 0$ also follow this same distribution. Thereby, when $\pi(.) = \mathsf{p}(.|\mathbf{z})$, one can construct a Langevin Markov chain whose stationary law is the target posterior distribution. In the following, we set $\pi(.) = \mathsf{p}(.|\mathbf{z})$ which amounts to assuming that the posterior density is differentiable with respect to $\mathbf{x}$.

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

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \varepsilon^2\mathbf{b}(\mathbf{x}^{(t)}) + \varepsilon\mathbf{A}^{1/2}(\mathbf{x}^{(t)})\boldsymbol{\omega}^{(t+1)} \tag{3.3}$$

where $\varepsilon > 0$ is the stepsize resulting from the Euler's discretization and $\boldsymbol{\omega}^{(t)} \in \mathbb{R}^Q$ is a realization of zero-mean white noise with covariance matrix

$\mathbf{I}_Q$. The scheme (3.3) is referred to as the Unadjusted Langevin Algorithm (ULA) [Roberts and Tweedie, 1996]. Due to the discretization error, the Markov chain following the ULA scheme may behave differently from the diffusion process resulting from (3.2). In particular, it may sway away from the target stationary distribution as pointed out in [Roberts and Tweedie, 1996; Girolami and Calderhead, 2011]. This discrepancy can be corrected by adding a Metropolis acceptance probability at each iteration to guarantee the reversibility of the chain with respect to the posterior distribution. The resulting sampler can be seen as a MH algorithm where $\mathbf{g}(.|\mathbf{x}^{(t)})$ is the density of a Gaussian distribution with mean $\mathbf{x}^{(t)} + \varepsilon^2\mathbf{b}(\mathbf{x}^{(t)})$ and covariance matrix $\varepsilon^2\mathbf{A}(\mathbf{x}^{(t)})$. Note that the convergence properties have been also studied for some variants of ULA in [Roberts and Tweedie, 1996; Durmus and Moulines, 2015; Durmus et al., 2016].

It is worth noting that two scale parameters play an important role: ε determines the length of the proposed jumps whereas the scale matrix $\mathbf{A}$ controls the direction. Various classes of algorithms have been developed from this diffusion model depending on the choice of $\mathbf{A}$.

1.2 Choice of the scale matrix

The standard Metropolis adjusted Langevin algorithm (MALA) is the simplest form of diffusion (3.3) when $\mathbf{A}$ equals $\mathbf{I}_Q$ [Roberts and Tweedie, 1996]:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\nabla \log \mathrm{p}(\mathbf{x}^{(t)}|\mathbf{z}) + \varepsilon\boldsymbol{\omega}^{(t+1)}. \tag{3.4}$$

It can be proved that the MALA sampling algorithm has $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ as its stationary distribution and is more likely to accept proposed values than a standard RW. Indeed, the gradient information of the target distribution allows the chain to be guided toward regions of higher probability, where most of the samples should lie and hence, it enables to achieve high acceptance rates. As a consequence, the MALA algorithm explores the invariant distribution much faster than the standard RW [Roberts and Rosenthal, 2001; Breyer et al., 2004]. Moreover, it should be noted that a bad adjustment of ε can significantly affect the convergence rate especially for high dimensional problems [Pillai et al., 2012]. For this reason, many methods have focused on how to choose a suitable stepsize such that the asymptotic average acceptance rate is bounded away from zero for high dimensions [Pillai et al., 2012; Roberts and Rosenthal, 1998]. Despite these improvements, when the variables of interest are strongly correlated with widely differing variances, MALA algorithm fails to explore efficiently the target space. In fact, since the third term of the MALA algorithm is an isotropic Brownian motion, the discretization stepsize ε in such parameter space, is generally constrained to take very small values in order to promote the directions with smallest variances which results in a slow convergence of the algorithm, poor mixing

of the chain and highly correlated samples [Girolami and Calderhead, 2011]. The performance of MALA algorithm may be improved by introducing a scale matrix $\mathbf{A}$ [Roberts and Stramer, 2002]. Some approaches have been proposed to accelerate the algorithm by preconditioning the proposal density with a constant scale matrix [Stuart et al., 2004]. Such algorithms propose samples according to the following scheme:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\mathbf{A}\nabla \log \mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z}) + \varepsilon \mathbf{A}^{1/2}\boldsymbol{\omega}^{(t+1)}. \tag{3.5}$$

Whereas the step size ε can easily be tuned with respect to the asymptotic acceptance rate, there is no clear guiding strategies for the selection of the constant matrix in the absence of some knowledge about the moments of the target density which are supposed to be unknown. Furthermore, the use of the same preconditioning matrix in the whole algorithm may be inefficient since optimal scaling of the burn-in period may differ from that of the stationary phase [Christensen et al., 2005]. Therefore, rather than employing a fixed global scale matrix in the proposal density, a position dependent matrix should be employed to take into account the local structure of the target density at each state of the Markov chain. In that respect, many algorithms [Atchadé, 2006; Martin et al., 2012; Zhang and Sutton, 2011; Bui-Thanh and Ghatas, 2012; Girolami and Calderhead, 2011; Vacar et al., 2011] rely on adaptive procedures where $\mathbf{A}$ is tuned automatically according to the past behavior of the Markov chain resorting to some deterministic optimization tools. For example, when setting $\mathbf{A}$ to the inverse of the Hessian matrix of $-\log \mathsf{p}(\mathbf{x}|\mathbf{z})$ and, assuming a locally constant curvature, the term involving the derivatives of the scale matrix in (3.2) reduces to zero. Consequently, a new sample is drawn from:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\mathbf{A}(\mathbf{x}^{(t)})\nabla \log \mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z}) + \varepsilon \mathbf{A}^{1/2}(\mathbf{x}^{(t)})\boldsymbol{\omega}^{(t+1)} \tag{3.6}$$

where $\mathbf{A}^{-1}(\mathbf{x}) = -\nabla^2 \log \mathsf{p}(\mathbf{x}|\mathbf{z})$ that is for all $i \in \{1, \ldots, Q\}$, for all $j \in \{1, \ldots, Q\}$, $\left[\mathbf{A}^{-1}\right]_{i,j}(\mathbf{x}) = -\dfrac{\partial^2 \log \mathsf{p}(\mathbf{x}|\mathbf{z})}{\partial x_i \partial x_j}$. Consequently, the computation of the drift term $\mathbf{b}$ becomes a scaled Newton step for minimizing $-\log \mathsf{p}(\mathbf{x}|\mathbf{z})$. Thus, a new sample of the Newton-based MCMC is more likely drawn from a highly probable region and then more likely accepted, which can speed up the convergence of the sampling process [Martin et al., 2012; Zhang and Sutton, 2011; Bui-Thanh and Ghatas, 2012]. Note that, in practice, this method has a high computational load since it requires the computation of the full Hessian matrix and its inverse at each iteration. This is particularly critical for large scale problems and/or when the Hessian matrix is not positive definite. One appealing solution is to replace the Hessian by a scale matrix that provides similar information than the Hessian with a lower computational cost. In particular, many methods have proposed the Fisher information matrix

as a preconditioning matrix in the Langevin diffusion [Girolami and Calderhead, 2011; Vacar et al., 2011] which can be interpreted as the discretization of the MALA algorithm directly on the natural Riemannian manifold where the parameters live. In this work, we propose a new approach where the scale matrix of the Langevin diffusion is chosen according to a Majorize-Minimize strategy.

§ 2 Proposed algorithm

2.1 Majorize-Minimize Framework

The majorization-minimization (MM) principle is a powerful tool for designing algorithms to solve optimization problems. The idea behind the MM approach is to replace a complicated minimization problem with successive minimizations of some well chosen surrogate functions [Hunter and Lange, 2004]. These functions are called tangent majorants.

Definition 2.1 *Tangent majorant*
Let $\mathbf{x}' \in \mathbb{R}^Q$. A function f is said to be a tangent majorant function of $\mathcal{J}$ at $\mathbf{x}'$ provided that

$$\begin{cases} P_1: \ f(\mathbf{x}', \mathbf{x}') = \mathcal{J}(\mathbf{x}'), \\ P_2: \ f(\mathbf{x}', \mathbf{x}) \geqslant \mathcal{J}(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathbb{R}^Q). \end{cases} \tag{3.7}$$

Let $\mathbf{x}^{(0)}$ be an arbitrary initial value and $\left(\mathbf{x}^{(t)}\right)_{t\in\mathbb{N}}$ the sequence constructed according to:

$$\mathbf{x}^{(t+1)} = \underset{\mathbf{x}\in\mathbb{R}^Q}{\operatorname{argmin}} f(\mathbf{x}^{(t)}, \mathbf{x}). \tag{3.8}$$

According to the majorization properties (3.7), the scheme (3.8) will produce a monotically decreasing sequence $(\mathcal{J}(\mathbf{x}^{(t)}))_{t\in\mathbb{N}}$ that converges to a local minimum of $\mathcal{J}$. In fact, we have

$$\mathcal{J}(\mathbf{x}^{(t)}) \underset{(a)}{=} f(\mathbf{x}^{(t)}, \mathbf{x}^{(t)}) \underset{(b)}{\geqslant} f(\mathbf{x}^{(t)}, \mathbf{x}^{(t+1)}) \underset{(c)}{\geqslant} \mathcal{J}(\mathbf{x}^{(t+1)}) \tag{3.9}$$

where (a) holds from the tangency property P_1, (b) from the minimization step (3.8) and (c) from the majorization property P_2 (see Figure 3.1).

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

$$f(\mathbf{x}', \mathbf{x}) = \mathcal{J}(\mathbf{x}') + (\mathbf{x} - \mathbf{x}')^\top \nabla\mathcal{J}(\mathbf{x}') + \frac{1}{2}(\mathbf{x} - \mathbf{x}')^\top \mathbf{Q}(\mathbf{x}')(\mathbf{x} - \mathbf{x}'), \tag{3.10}$$

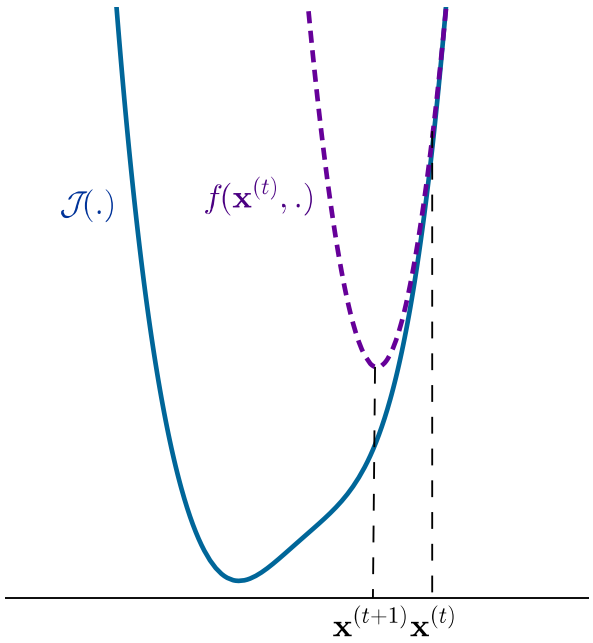


Figure 3.1: *MM algorithm: the new iterate* $\mathbf{x}^{(t+1)}$ *is the minimizer of the tangent majorant* $f(\mathbf{x}^{(t)},.)$ *of* $\mathcal{J}$ *in* $\mathbf{x}^{(t)}$.

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

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{x}^{(t)})\nabla\mathcal{J}(\mathbf{x}^{(t)}) \tag{3.11}$$

with $\varepsilon \in (0, \sqrt{2}]$ is a relaxation stepsize. Note that (3.11) implies that inequality (b) in (3.9) is satisfied, remarking that $2\varepsilon^{-2}\mathbf{Q}(\mathbf{x}') \succeq \mathbf{Q}(\mathbf{x}')$, for every $\mathbf{x}' \in \mathbb{R}^Q$ and every $\varepsilon \in (0, \sqrt{2}]$.

2.2 Proposed sampling algorithm

In this work, we propose to extend the idea behind the MM strategy to the context of sampling algorithms. More specifically, the idea is to push the proposal distribution of the MH algorithm at each iteration from the current state to a region with high density value. Contrary to the RW where the proposal is centered on the current state, we propose to pick the mean of the proposal density using a step of an MM search with form (3.11) and then to explore the space around this center according to the MM curvature matrix $\mathbf{Q}(\mathbf{x}^t)$ that should well describe the local curvature of the target distribution. This results in a preconditioned Langevin proposal where the scale matrix $\mathbf{A}$ in (3.2), equal to the inverse of the curvature matrix $\mathbf{Q}(\mathbf{x}^t)$, is constructed from the MM strategy. Similarly to Newton-based

MCMC methods, the drift term, when assuming zero curvature changes, proposes, from a current state $\mathbf{x}^{(t)}$, a state with a higher value of $\log \mathsf{p}(\mathbf{x}|\mathbf{z})$, resulting from an iteration of MM algorithm minimizing $\mathcal{J}(\mathbf{x}) = -\log \mathsf{p}(\mathbf{x}|\mathbf{z})$. Consequently, the obtained proposal reduces to a noisy version of an MM iteration for minimizing $-\log \mathsf{p}(\mathbf{x}|\mathbf{z})$. The proposed sample is then subjected to the accept/reject rule of the MH algorithm. The resulting 3MH sampler is described by Algorithm 4.

Algorithm 4 Majorize-Minimize adapted Metropolis–Hastings algorithm

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

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

Generate

$$\tilde{\mathbf{x}}^{(t)} \sim \mathcal{N}\left(\mathbf{x}^{(t)} + \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{x}^{(t)})\nabla \log \mathsf{p}(\mathbf{x}^{(t)}|\mathbf{z}), \varepsilon^2 \mathbf{Q}^{-1}(\mathbf{x}^{(t)})\right)$$

Acceptance-Rejection:

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

Compute

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

where $\mathsf{g}(.|\mathbf{v}) \propto |\mathbf{Q}(\mathbf{v})|^{\frac{1}{2}} \exp\left(-\frac{1}{2\varepsilon^2}\|. - \mathbf{v} - \frac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{v})\nabla \log \mathsf{p}(\mathbf{v}|\mathbf{z})\|^2_{\mathbf{Q}(\mathbf{v})}\right)$

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

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

else

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

end if

end for

Recall that the metric $\mathbf{Q}$ is the precision matrix of the Gaussian proposal distribution which makes the choice of $\mathbf{Q}$ crucial for the efficiency of the sampling algorithm. The matrix $\mathbf{Q}(\mathbf{x}^{(t)})$ at each iteration t should be chosen such that (3.10) is a tangent majorant to the minus logarithm of the posterior density at the current state $\mathbf{x}^{(t)}$, that is it should satisfy the properties in (3.7). Furthermore, it should provide a good approximation of the local curvature of the posterior distribution. Let $\mathbf{x}' \in \mathbb{R}^Q$. Any symmetric positive definite matrix verifying for every $\mathbf{x} \in \mathbb{R}^Q$, $\mathbf{Q}(\mathbf{x}') \succeq -\nabla^2 \log \mathsf{p}(\mathbf{x}|\mathbf{z})$, defines the curvature of a quadratic tangent majorant of $-\log \mathsf{p}(\mathbf{x}|\mathbf{z})$ at $\mathbf{x}'$. In the following, we propose a general procedure for building such a set of suitable preconditioning matrices $\{\mathbf{Q}(\mathbf{x})\}_{\mathbf{x}\in\mathbb{R}^Q}$ under some mild conditions about the

posterior distribution.

2.3 Construction of the tangent majorant

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

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathcal{J}(\mathbf{x}) = \Phi(\mathbf{H}\mathbf{x} - \mathbf{z}) + \Psi(\mathbf{V}\mathbf{x}) \tag{3.12}$$

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

$$\Psi(\mathbf{V}\mathbf{x}) = \sum_{s=1}^{S} \psi_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|) \tag{3.13}$$

with $\mathbf{V} = [\mathbf{V}_1^\top, \ldots, \mathbf{V}_S^\top]^\top$ and $(\forall s \in \{1, ..., S\})$ $\mathbf{V}_s \in \mathbb{R}^{P_s\times Q}$, $\mathbf{c}_s \in \mathbb{R}^{P_s}$ and $(\psi_s)_{1\leqslant s\leqslant S}$ is a set of nonnegative continuous functions. Note that this form of posterior density is frequently encountered in inverse problems where $\mathbf{z}$ is the observation, Φ is the data fidelity term and Ψ is the minus logarithm of the prior density defined according to some linear operators $\mathbf{V}_1, \ldots, \mathbf{V}_S$. For instance, $\mathbf{V}$ may be a matrix computing the horizontal and vertical discrete gradient (or higher order differences) between neighboring pixels useful for edge preserving in image restoration problems. In this case, by setting $P_s = 1$ and $\psi_s(.) = |.|$, we recover the anisotopic total variation while for $P_s = 2$ and ψ_s equals to ℓ_2 norm, we obtain the isotropic total variation (see chapter 2). Another important choice, is the analysis frame regularization where $\mathbf{V}$ is a frame of $\mathbb{R}^Q$. For example, $\mathbf{V}_1$ may be the operator that computes low frequency wavelet coefficients and ψ_1 a function enforcing smooth solutions while the remaining operators give the high frequency ones that can be well described using suitable heavy tailed functions ψ_s such as the ℓ_p^p penalties for $p < 1$, the Cauchy or the Bernoulli-Gaussian models. As Langevin based algorithms require the use of differentiable regularizations, one can use smoothed approximations of these functions that have a quadratic behavior near 0 [Allain et al., 2006; Charbonnier et al., 1997; Lange, 1990; Zibulevsky and Elad, 2010].

We further make the following assumptions:

1. Φ is a continuous coercive differentiable function with an L-Lipschitzian gradient, that is

$$\big(\forall \mathbf{u} \in \mathbb{R}^N\big)\,\big(\forall \mathbf{v} \in \mathbb{R}^N\big) \quad \|\nabla\Phi(\mathbf{u}) - \nabla\Phi(\mathbf{v})\| \leqslant L\|\mathbf{u} - \mathbf{v}\|, \tag{3.14}$$

2. $(\forall s \in \{1, ..., S\})$ ψ_s is a differentiable function,

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

4. $(\forall s \in \{1, ..., S\})$ $(\exists \bar{\omega}_s > 0)$ such that $(\forall u > 0)$ $0 \leqslant \dot{\psi}_s(u) \leqslant \bar{\omega}_s u$ and $\lim_{u \to 0} \dot{\psi}_s(u)/u < \infty$.

The first requirement holds for a large number of data fidelity terms. This includes for example the Gaussian noise model, the Huber function which has shown its efficiency compared to the quadratic one for limiting the influence of outliers encountered in some observed data [Huber, 2011], the Cauchy model [Antoniadis et al., 2002], and the signal-dependent Gaussian model generally used as a second order approximation of mixed Poisson-Gaussian noise [Repetti et al., 2012] as well as the exact Poisson-Gaussian likelihood [Chouzenoux et al., 2015]. More examples can be found in [Chouzenoux et al., 2013]. Furthermore, Assumptions 2-4 are satisfied for several commonly used prior models such as $\mathcal{ST}$, $\mathcal{SMG}$, $\mathcal{GMEP}$ distributions (see Chapter 2) as well as smoothed approximation of ℓ_p^p regularization functions for $p \leqslant 2$ and $\ell_2 - \ell_0$ penalties (asymptotically constant with a quadratic behavior near 0) used to approximate the ℓ_0 pseudo-norm [Veksler, 1999; Ganan and McClure, 1985; Dennis Jr and Welsch, 1978; Chouzenoux et al., 2013].[1]

Under Assumptions 1-4, convex quadratic tangent majorants of (3.12) can be obtained by setting [Chouzenoux et al., 2013]

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_1(\mathbf{x}) = \mu \mathbf{H}^\top \mathbf{H} + \mathbf{V}^\top \mathrm{Diag}\{\boldsymbol{\omega}(\mathbf{x})\} \mathbf{V} + \zeta \, \mathbf{I}_Q \tag{3.15}$$

where $\mu \in [L, +\infty[$, $\mathbf{V} = \left[\mathbf{V}_1^\top, \ldots, \mathbf{V}_S^\top\right]^\top$ and $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_P]^\top$ is such that, for all $s \in \{1, \ldots, S\}$, $p \in \{1, \ldots, P_s\}$

$$\omega_{P_1+P_2+\ldots+P_{s-1}+p}(\mathbf{x}) = \frac{\dot{\psi}_s(\|\mathbf{V}_s \mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s \mathbf{x} - \mathbf{c}_s\|}. \tag{3.16}$$

Moreover, $\zeta \geqslant 0$ is a constant that ensures the invertibility of $\mathbf{Q}_1(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^Q$.

The numerical efficiency of the proposed algorithm relies on the use of quadratic majorants that provide tight approximations of the target density but also whose curvature matrices are simple to compute. However, sampling from the proposal constructed by the MM strategy when using the curvature matrix (3.15) is often very difficult because of the high computational cost of each iteration and/or the memory limitations. In fact, similarly to Newton MCMC samplers, the main computational cost is related to the computation of the inverse of (3.15) and sampling from the associated high-dimensional Gaussian distribution at each iteration. In the following, we will propose alternative choices of the curvature matrix, when the manipulation of matrix (3.15) is intractable.

[1] Again, improper prior laws are tolerated provided that the resulting posterior distribution is proper.

Constant curvature matrix: We can resort to the following constant curvature matrix which can be seen as a constant majorant of (3.15) with respect to $\mathbf{x}$:

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_2(\mathbf{x}) = \mu \mathbf{H}^\top \mathbf{H} + \max_{1 \leqslant s \leqslant S} (\bar{\omega}_s)\, \mathbf{V}^\top \mathbf{V} + \zeta \mathbf{I}_Q. \tag{3.17}$$

It can be noted that in the special case, when $\mathbf{H}$ is circulant and $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_Q$ which is the case for example when $\mathbf{V}$ is a tight frame analysis operator, then $\mathbf{Q}_2$ is easily invertible in the Fourier domain. More generally, when $\mathbf{H}$ and $\mathbf{V}$ can be diagonalized in the same domain, the inverse and the square root decomposition of (3.17) can be easily performed in this domain.

Diagonal curvature matrix: We also propose the following alternative choice described in [Chouzenoux et al., 2014], which can be understood as a diagonal approximation of (3.15):

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_3(\mathbf{x}) = (\mu \|\mathbf{H}\|^2 + \zeta)\mathbf{I}_Q + \mathrm{Diag}\left(\mathbf{P}^\top \boldsymbol{\omega}(\mathbf{x})\right) \tag{3.18}$$

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

$$(\forall i \in \{1, \ldots, P\})(\forall j \in \{1, \ldots, Q\}) \quad \mathbf{P}_{i,j} = |\mathbf{V}_{i,j}| \sum_{k=1}^{Q} |\mathbf{V}_{i,k}|. \tag{3.19}$$

§ 3 Convergence analysis

In this section, we address the convergence properties of the proposed algorithm. Similarly to [Atchadé, 2006], we will make the following technical assumption about the drift term:

$$\mathbf{b}(\mathbf{x}) = \frac{\varepsilon^2}{2} \mathbf{Q}^{-1}(\mathbf{x}) \mathbf{D}(\mathbf{x}) \tag{3.20}$$

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

$$\mathbf{D}(\mathbf{x}) = \frac{d}{\max(d, \|\nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})\|)} \nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z}) \tag{3.21}$$

with $d > 0$. Note that, the drift term (3.20) is equivalent to the one in Algorithm 4 for large values of d.

We further make the following assumptions:

Assumption 3.1 $\mathsf{p}(.|\mathbf{z})$ is continuous and $\mathsf{p}(\mathbf{x}|\mathbf{z}) > 0$, for every $\mathbf{x} \in \mathbb{R}^Q$.

Remark 3.1 If $(\psi_s)_{s\in\{1,\dots,S\}}$ and Φ are continuous and take finite values, then Assumption 3.1 holds.

Assumption 3.2 $\mathsf{p}(\mathbf{x}|\mathbf{z})$ is the density of a super-exponential distribution that is $\mathsf{p}(\mathbf{x}|\mathbf{z})$ is positive and has continuous first derivatives such that

$$\lim_{\|\mathbf{x}\|\to+\infty} \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|},\ \nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z}) \right\rangle = -\infty. \tag{3.22}$$

Assumption 3.3 We have

$$\lim_{\|\mathbf{x}\|\to\ +\infty} \sup \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|},\ \frac{\nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})}{\|\nabla \log \mathsf{p}(\mathbf{x}|\mathbf{z})\|} \right\rangle < 0. \tag{3.23}$$

Remark 3.2 In the particular case when $\Phi(.) = \|.\|^2$ in (3.12), Assumptions 3.2 and 3.3 are satisfied if one of the following properties holds:

- $\mathbf{H}$ is injective, for example $\mathbf{H} = \mathbf{I}_Q$ which is the case for denoising problems,
- There exists $s_0 \in \{1,\dots,S\}$ such that
  - $\mathrm{Ker}(\mathbf{H}) \cap \mathrm{Ker}(\mathbf{V}_{s_0}) = \{\mathbf{0}_Q\}$,
  - $\lim_{t\to+\infty} \dot{\psi}_{s_0}(t) = +\infty$,
  - for every $s \in \{1,\dots,S\} \setminus \{s_0\}$, we have $\lim_{t\to+\infty} \frac{\dot{\psi}_s(t)}{\dot{\psi}_{s_0}(t)} < +\infty$.

Proof. Let $\mathbf{x} \in \mathbb{R}^Q$. We have

$$\nabla \mathcal{J}(\mathbf{x}) = \mathbf{H}^\top (\mathbf{H}\mathbf{x} - \mathbf{z}) + \sum_{s=1}^{S} \mathbf{V}_s^\top (\mathbf{V}_s\mathbf{x} - \mathbf{c}_s) \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|}. \tag{3.24}$$

Then

$$\|\nabla \mathcal{J}(\mathbf{x})\| \leqslant \|\mathbf{H}\|\|\mathbf{H}\mathbf{x} - \mathbf{z}\| + \sum_{s=1}^{S} \|\mathbf{V}_s\| \dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|) \tag{3.25}$$

$$\leqslant \|\mathbf{H}\|\|\mathbf{H}\mathbf{x} - \mathbf{z}\| + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\|\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|. \tag{3.26}$$

We have

$$\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x}) = \|\mathbf{H}\mathbf{x}\|^2 + \sum_{s=1}^{S} \|\mathbf{V}_s\mathbf{x}\|^2 \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|} + h(\mathbf{x}) \tag{3.27}$$

where

$$h(\mathbf{x}) = -\mathbf{x}^\top \left(\mathbf{H}^\top \mathbf{z} + \sum_{s=1}^{S} \mathbf{V}_s^\top \mathbf{c}_s \frac{\dot{\psi}_s(\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|)}{\|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\|} \right). \tag{3.28}$$

Assume that $\mathbf{H}$ is injective. According to (3.24), we have

$$\begin{aligned} \frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|} &\geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2}{\|\mathbf{x}\|} + \frac{h(\mathbf{x})}{\|\mathbf{x}\|} \\ &= \frac{\|\mathbf{H}\mathbf{x}\|^2}{\|\mathbf{x}\|} + O(1). \end{aligned} \tag{3.29}$$

Then, Assumption 3.2 is satisfied. Moreover, using (3.26), we have

$$\begin{aligned} \frac{\mathbf{x}^\top \nabla \mathcal{J}(\mathbf{x})}{\|\mathbf{x}\|\|\nabla \mathcal{J}(\mathbf{x})\|} &\geqslant \frac{\|\mathbf{H}\mathbf{x}\|^2 + h(\mathbf{x})}{\|\mathbf{x}\| \left(\|\mathbf{H}\|\|\mathbf{H}\mathbf{x} - \mathbf{z}\| + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\| \|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\| \right)} \\ &= \frac{\|\mathbf{H}\mathbf{x}\|^2}{\|\mathbf{x}\| \left(\|\mathbf{H}\|\|\mathbf{H}\mathbf{x} - \mathbf{z}\| + \sum_{s=1}^{S} \bar{\omega}_s \|\mathbf{V}_s\| \|\mathbf{V}_s\mathbf{x} - \mathbf{c}_s\| \right)} + o(1). \end{aligned} \tag{3.30}$$

So that, Assumption 3.3 also holds.

Similar arguments allow to derive the results in the case when $\mathbf{H}$ is not injective.

□

Assumption 3.4 For every $\mathbf{x} \in \mathbb{R}^Q$, the preconditionning matrix $\mathbf{Q}(\mathbf{x})$ has a bounded spectrum i.e., there exist two constants $\nu_{\min} > 0$ and $\nu_{\max} > 0$ independent of $\mathbf{x}$ such that

$$\left(\forall \mathbf{x} \in \mathbb{R}^Q\right) \ \nu_{\max} \mathbf{I}_Q \succeq \mathbf{Q}(\mathbf{x}) \succeq \nu_{\min} \mathbf{I}_Q. \tag{3.31}$$

Remark 3.3 Assumption 3.4 holds for all curvature matrices proposed in Section 2.3 provided that $\zeta > 0$. Furthermore, Assumption 3.4 together with (3.21), imply that the drift term $\mathbf{b}$ is bounded that is:

$$\left(\forall \mathbf{x} \in \mathbb{R}^Q\right) \quad \|\mathbf{b}(\mathbf{x})\| \leqslant \frac{\varepsilon^2}{2} \nu_{\min}^{-1} d. \tag{3.32}$$

Subsequently, under Assumptions 3.1-3.4 and using the expression in (3.21), we address the geometric ergodicity of the proposed algorithm based on the results concerning RW in [Jarner and Hansen, 2000] and by adapting the proofs in [Roberts and Tweedie, 1996; Atchadé, 2006; Schreck et al., 2016]. Since the algorithm appears as a special case of the MH algorithm, the chain $\left(\mathbf{x}^{(t)}\right)_{t\in\mathbb{N}}$ constructed by the 3MH algorithm has $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$ as an invariant distribution that is (2.62) is satisfied for $f^*(.) = \mathsf{p}(.|\mathbf{z})$ and $\mathsf{t}(.|\mathbf{z})$ equals to (2.78). The first important step of the proof of geometric ergodicity is to compare the proposal density g to Gaussian proposals.

Proposition 3.4 *There exist* $(k_1, k_2, \sigma_1, \sigma_2) \in (\mathbb{R}_+^*)^4$ *such that*

$$\left(\forall(\mathbf{x},\mathbf{y}) \in (\mathbb{R}^Q)^2\right) \; k_1 \mathsf{n}(\mathbf{y};\mathbf{x},\sigma_1^2\mathbf{I}_Q) \underset{(\mathbf{a})}{\leqslant} \mathsf{g}(\mathbf{x}|\mathbf{y}) \underset{(\mathbf{b})}{\leqslant} k_2 \mathsf{n}(\mathbf{y};\mathbf{x},\sigma_2^2\mathbf{I}_Q) \tag{3.33}$$

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

Proof. Let $\mathbf{x} \in \mathbb{R}^Q$ and $\boldsymbol{\mu}(\mathbf{x}) = \mathbf{x} + \dfrac{\varepsilon^2}{2}\mathbf{Q}^{-1}(\mathbf{x})\mathbf{D}(\mathbf{x})$, we have

$$-\log \mathsf{g}(\mathbf{x}|\mathbf{y}) = \frac{1}{2\varepsilon^2}\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2_{\mathbf{Q}(\mathbf{x})} - \frac{1}{2}\log|\mathbf{Q}(\mathbf{x})| + \frac{Q}{2}\log(2\pi\varepsilon^2). \tag{3.34}$$

From Assumption 3.4, we obtain

$$\nu_{\min}\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2 \leqslant \|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2_{\mathbf{Q}(\mathbf{x})} \leqslant \nu_{\max}\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2, \tag{3.35}$$

and

$$\nu_{\min}^Q \leqslant |\mathbf{Q}(\mathbf{x})| \leqslant \nu_{\max}^Q. \tag{3.36}$$

On the other hand, by using (3.21) and the triangle inequality, we have

$$\begin{aligned}\|\mathbf{y}-\mathbf{x}\| &\leqslant \|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\| + \|\boldsymbol{\mu}(\mathbf{x})-\mathbf{x}\|,\\ &\leqslant \|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\| + \frac{\varepsilon^2}{2}\nu_{\min}^{-1}d.\end{aligned} \tag{3.37}$$

By using Jensen inequality, it follows that

$$\|\mathbf{y}-\mathbf{x}\|^2 \leqslant 2\left(\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2 + \frac{\varepsilon^4}{4}\nu_{\min}^{-2}d^2\right). \tag{3.38}$$

Similarly, we have

$$\begin{aligned}\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\| &\leqslant \|\mathbf{y}-\mathbf{x}\| + \|\boldsymbol{\mu}(\mathbf{x})-\mathbf{x}\|,\\ &\leqslant \|\mathbf{y}-\mathbf{x}\| + \frac{\epsilon^2}{2}\nu_{\min}^{-1}d.\end{aligned} \tag{3.39}$$

Then

$$\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2 \leqslant 2\left(\|\mathbf{y}-\mathbf{x}\|^2+\frac{\varepsilon^4}{4}\nu_{\min}^{-2}d^2\right). \tag{3.40}$$

It follows from (3.35), (3.38) and (3.40) that

$$\frac{\nu_{\min}}{4\varepsilon^2}\|\mathbf{y}-\mathbf{x}\|^2-\frac{\varepsilon^2 d^2}{8\nu_{\min}} \leqslant \frac{1}{2\varepsilon^2}\|\mathbf{y}-\boldsymbol{\mu}(\mathbf{x})\|^2_{\mathbf{Q}(\mathbf{x})} \leqslant \frac{\nu_{\max}}{\varepsilon^2}\|\mathbf{y}-\mathbf{x}\|^2+\frac{\varepsilon^2\nu_{\max}d^2}{4\nu_{\min}^2}. \tag{3.41}$$

Then, using (3.34), (3.36) and (3.41), Proposition 3.4(a) holds for

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

$$\sigma_1^2=\frac{\varepsilon^2}{2\nu_{\max}}$$

and Proposition 3.4(b) is satisfied for

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

$$\sigma_2^2=\frac{2\varepsilon^2}{\nu_{\min}}.$$

□

Corollary 3.5 *For every* $(\mathbf{x},\mathbf{y})\in(\mathbb{R}^Q)^2$, *we have* $\mathsf{g}(\mathbf{x}|\mathbf{y})>0$ *and* $\mathsf{g}(\mathbf{y}|\mathbf{x})>0$.

Theorem 3.6 *Under Assumption* (3.1)-(3.4), *the Markov chain defined by the 3MH algorithm using the truncated gradient* (3.21) *is geometrically ergodic with stationary distribution* $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$.

Proof. From Algorithm 4 and Corollary 3.5, g is positive and continuous. Since $\mathsf{p}(.|\mathbf{z})$ is also positive and continuous, we can deduce that the chain is aperiodic and $\mathsf{p}(\mathbf{x}|\mathbf{z})$- irreducible with unique invariant distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$.

Assumptions 3.2 and 3.3 have already been introduced in [Jarner and Hansen, 2000] as sufficient conditions for the geometric ergodicity of the RW algorithm. It has also been shown that under these assumptions, the MALA algorithm with truncated gradient (3.21) is geometrically ergodic [Atchadé, 2006]. Since the drift term of the proposed algorithm is bounded with a truncated gradient, the geometric ergodicity property can be deduced by a straightforward adaptation of the proof in [Atchadé, 2006] for MALA algorithm with truncated drift. Note that this convergence proof is valid for any preconditioned MALA algorithm provided that the preconditioning metric has a bounded spectrum. □

§ 4 Experimental results

We are interested in the deconvolution of a seismic signal $\bar{\mathbf{x}}$ of length $Q = 784$. This sparse signal is composed of a sequence of spikes called primary reflection coefficients [Walden and Hosken, 1986; Repetti et al., 2015] as depicted in Figure 3.3. These coefficients give information about the travel time of seismic waves between two seismic reflectors, and the amplitude of the seismic events reflected back to the sensor. The signal is degraded according to the observation model defined in (2.2) where $\mathbf{H}$ is a circulant blur matrix and $\mathbf{w}$ is a Gaussian noise of variance σ^2.

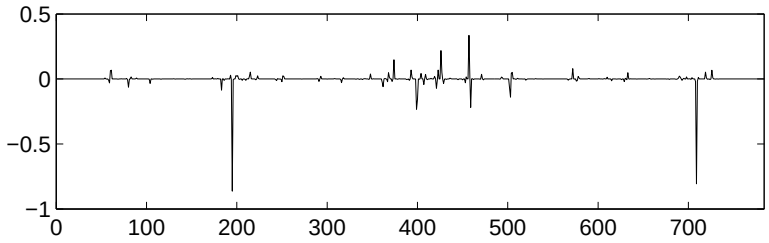


Figure 3.2: *Original signal.*

4.1 Prior and posterior distributions

In order to promote the signal sparsity, we suppose that its coefficients are independent and identically distributed according to a $\mathcal{ST}$ distribution. Then, we can identify the following functions:

$$\Phi(\mathbf{Hx} - \mathbf{z}) = \frac{1}{2\sigma^2}\|\mathbf{Hx} - \mathbf{z}\|^2 \tag{3.42}$$

and

$$\Psi(\mathbf{x}) = \frac{\nu + 1}{2} \sum_{i=1}^{Q} \log\left(\gamma^2 + \frac{(x_i - \mu)^2}{\nu}\right) \tag{3.43}$$

where $\nu > 0$ is the number of degrees of freedom determining the shape of the distribution, μ is the position parameter and $\gamma > 0$ is the scale parameter. Note that the Cauchy distribution is a particular case when $\nu = 1$. The $\mathcal{ST}$ distribution has been often used in image reconstruction to model the distribution of the wavelet coefficients [Chantas et al., 2008]. This penalty has already been introduced in [Hebert and Leahy, 1989] as a compromise between the ℓ_2 norm and the non-convex approximation of the semi-norm ℓ_0 presented in [Geman and McClure, 1987] to enforce the sparsity of the signal and better preserve discontinuities. Recall that the $\mathcal{ST}$ distribution can be written as a scale mixture of normal distribution where the hidden variable follows a gamma distribution with both parameters equal to $\nu/2$. In most

Bayesian methods, it is generally used in this form: the unknown signal $\mathbf{x}$ and the hidden variable are estimated from their posterior joint distribution. In this work, we propose to directly use the expression defined in (3.43).

In the following, we assume that ν is known, but we have only few prior information about the others parameters. Therefore, we use uniform distributions for μ and γ defined in $[-\mu_m, \mu_M]$ and $[\gamma_m, \gamma_M]$ respectively where μ_m, μ_M, γ_m et γ_M are positive constants. Thus, the posterior distributions of the parameters are given by

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

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

Since Φ and Ψ satisfy the properties in Section 2.3, we propose to use 3MH algorithm to sample from the posterior distribution of $\mathbf{x}$. More specifically, we aim to test the performance of 3MH algorithms using the different proposed curvature matrices namely $\mathbf{Q}_1$, the constant circulant matrix $\mathbf{Q}_2$ and the diagonal matrix $\mathbf{Q}_3$ defined in this application by

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_1(\mathbf{x}) = \frac{1}{\sigma^2}\mathbf{H}^\top\mathbf{H} + \mathrm{diag}\{\boldsymbol{\omega}(\mathbf{x})\} + \zeta\mathbf{I}_Q, \tag{3.44}$$

$$\mathbf{Q}_2 = \frac{1}{\sigma^2}\mathbf{H}^\top\mathbf{H} + \frac{\nu+1}{\nu\gamma^2}\,\mathbf{I}_Q, \tag{3.45}$$

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \mathbf{Q}_3(\mathbf{x}) = \mathrm{Diag}\left(\frac{1}{\sigma^2}\mathbf{P}^\top\mathbf{1}_Q + \boldsymbol{\omega}(\mathbf{x})\right), \tag{3.46}$$

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

$$(\forall i \in \{1,\ldots,Q\}) \quad \omega_i(\mathbf{x}) = \frac{\nu+1}{\nu\gamma^2 + (x_i-\mu)^2} \tag{3.47}$$

and

$$(\forall i \in \{1,\ldots,N\})\,(\forall j \in \{1,\ldots,Q\}) \quad P_{i,j} = |H_{i,j}| \sum_{k=1}^{N} |H_{k,j}|. \tag{3.48}$$

Note that $\zeta \geqslant 0$ is a constant added to ensure the positive definiteness of the matrix $\mathbf{Q}_1$. It is worth noting that if $\mathbf{H}$ is injective, the 3MH algorithm is geometrically ergodic.

The posterior laws of the $\mathcal{ST}$ prior parameters do not have usual forms. Then, it is not easy to directly generate samples of μ and γ. We propose therefore to estimate them using a RW algorithm whose scale parameter is tuned automatically during the burn-in period so as to reach an acceptance probability equals 0.33.

4.2 Results

The test signal is artificially degraded by a band-pass filter of size 41 with a frequency spectrum concentrated between 10 and 40 Hz and an additive Gaussian noise of variance $\sigma^2 = 2.5 \times 10^{-3}$ (see Figure 3.3). The initial signal-to-noise ratio (SNR) is -4.58 dB. We fix $\nu = 1$ which corresponds to the special case of the Cauchy prior. Figure 3.4 shows the error between the original signal and the degraded one as well as the error between the original signal and the restored one using the MMSE estimator which correspond to an SNR equal to 8.24 dB.

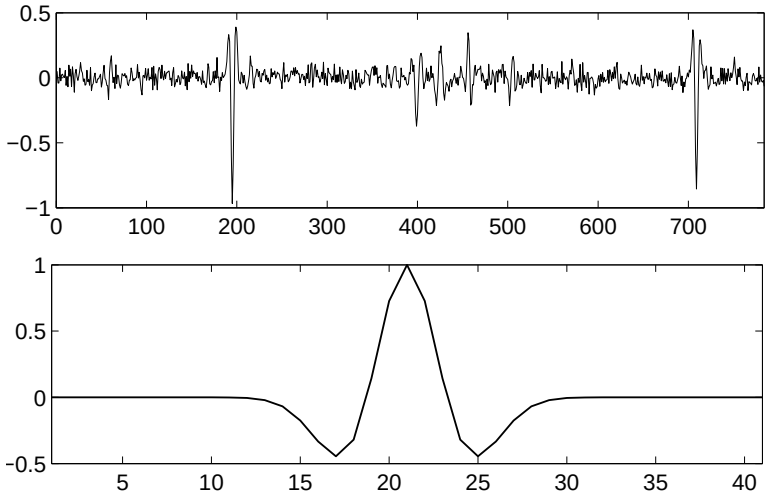


Figure 3.3: *Degraded signal (top). Blurring kernel (bottom).*

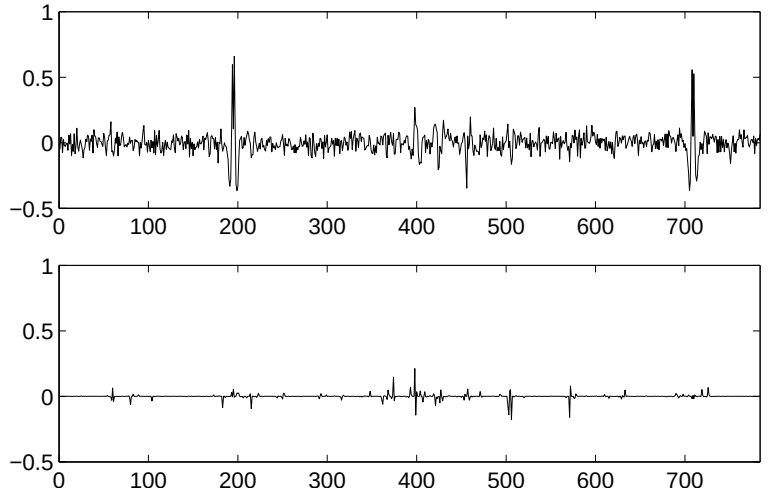


Figure 3.4: *Error* $\bar{\mathbf{x}} - \mathbf{z}$ *(top). Error* $\bar{\mathbf{x}} - \widehat{\mathbf{x}}$ *(bottom).*

We propose to compare the convergence speed of 3MH algorithm using the different curvatures matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ with the standard MALA algorithm. Thus, we run these algorithms until convergence. The discretization stepsize ε is adjusted for all these algorithms during the burn-in period to correspond to an acceptance probability between 0.3 and 0.6. Note that in order to reduce the complexity of each iteration when using $\mathbf{Q} = \mathbf{Q}_1$, the inversion of the curvature matrix is replaced by a conjugate gradient algorithm and the generation of random variables according to the proposal is ensured using the sampling method from [Orieux et al., 2012]. Figure 3.5 shows the evolution of $\mathcal{J}$ with respect to time. Following [Atchadé, 2006], we also compare the different methods in terms of the mean square jump (MSJ) in stationarity defined in (2.75) which indicates how much the Markov chain is exploring the whole target space in convergence. Note that MSJ has been estimated with an empirical average over $P = 5,000$ samples $\mathbf{x}^{(t_0+1)}, \ldots, \mathbf{x}^{(t_0+P)}$ generated after the burn-in period as follows

$$MSJ = \left(\frac{1}{P-1} \sum_{t=1}^{P-1} \|\mathbf{x}^{(t_0+t)} - \mathbf{x}^{(t_0+t+1)}\|^2 \right)^{1/2} . \tag{3.49}$$

In Table 3.1, we show estimates of the mean square jump per second in stationarity which is defined as the ratio of the mean square jump and the computational time per iteration. We also compare the statistical efficiency of the different samplers with respect to MALA defined as the mean square jump per second of each sampler over the mean square jump per second of MALA.

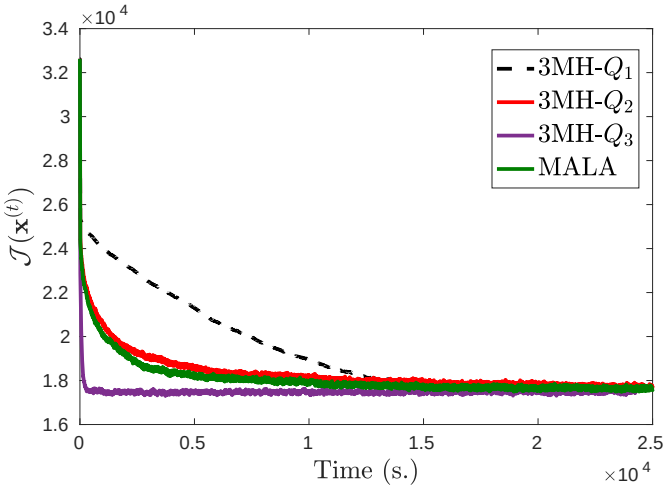


Figure 3.5: *Convergence speed of MALA, 3MH - $\mathbf{Q}_1$, 3MH - $\mathbf{Q}_2$ and 3MH - $\mathbf{Q}_3$.*

One can notice that the behavior of 3MH algorithm using the constant curvature matrix $\mathbf{Q}_2$ is close to that of MALA in terms of convergence speed

Table 3.1: *Mixing results for the different proposed algorithms. First row: Estimates of the mean square jump in stationarity. Second row: Time per iteration in stationarity. Third row: Estimates of the mean square jump per second in stationarity. Fourth row: Efficiency relatively to MALA.*

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

and MSJ. This can be explained by the low dispersion of the eigenvalues of $\mathbf{Q}_2$ in this example. The addition of an adaptive preconditioning matrix in MALA allows us to reach areas of high probability in fewer number of iterations and increase the MSJ which indicates better mixing of the Markov chain in convergence. Nevertheless, the use of the matrix $\mathbf{Q}_1$ at each iteration becomes more expensive as the problem dimension increases which deteriorates the efficiency of the algorithm. Thus, the choice of the diagonal adaptive matrix $\mathbf{Q}_3$ appears to achieve the best compromise between these different algorithms due to the low complexity that it induces at each iteration. It allows to reach stability much faster than the other algorithms while achieving mixing properties comparable to MALA at convergence.

Conclusion

In this work, we have proposed a new MCMC algorithm that can be considered as a scaled MALA where the scale matrix is adapted at each iteration with a MM strategy. We have shown that the geometric ergodicity property of the standard Langevin MH algorithms is maintained by introducing this scale matrix for the class of sub-exponential distributions. We have then applied this algorithm to compute the MMSE estimator of a sparse signal from its blurred version using a $\mathcal{ST}$ prior distribution. Experimental results indicate the satisfactory performance of this new MCMC method compared to the standard MALA algorithm. Note that another example, maybe more striking, of the good performance offered by the 3MH algorithm, will be provided in Section 3 in Chapter 4.

- Chapter 4 -

An Auxiliary Variable Method for MCMC algorithms

High dimensional models, often encountered in inverse problems, present a challenging task for Bayesian inferences. While many popular MCMC sampling algorithms have been widely used to fit complex multivariable models in small dimensional spaces, they generally fail to explore the target distribution efficiently when applied to large scale problems. This is mainly due to the poor mixing properties of the Markov chain or to the high cost of each iteration.

In this chapter, we propose a method for Bayesian sampling in large scale problems. Our approach is a special case of data augmentation type strategies [Van Dyk and Meng, 2012] allowing to overcome the limitations of standard sampling algorithms.

The remainder of this chapter is organized as follows. In Section 1, we discuss the main difficulties encountered in standard sampling algorithms for large scale problems and how adding auxiliary variables to the model can alleviate these issues. The core of our contribution is detailed in Section 2. We first give a complete description of the proposed approach in the case of a Gaussian noise and we study its extension to scale mixture 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 3 and 4 are devoted to the experimental validation of our method. First, 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. Second, we test the performance of our method in sampling from large scale Gaussian distributions with an application to image recovery under two-terms mixed Gaussian noise.

§ 1 Motivation

1.1 Sampling issues in high dimensional space

MCMC sampling methods may face two main difficulties when applied to large scale inverse problems. First, the structure of the observation model that links the unknown signal to the observations can make the evaluation of the parameters of the posterior distribution very expensive mainly because of the observation matrix. Second, even with common models, the resulting posterior distribution may be difficult to sample from directly or to explore efficiently using standard sampling algorithms. As specific cases, this problem emerges either in high dimensional Gaussian distribution sampling or in MH algorithms especially when constructing efficient proposals that cope with both the high dimensionality and the strong correlation existing between the target parameters.

Sampling from high dimensional Gaussian distribution: Suppose that we are interested in 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. In fact, let us consider the linear model in (2.2) and 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 $\mathcal{N}(\mathbf{m}_\mathbf{x}, \mathbf{G}_\mathbf{x}^{-1})$ respectively where $\mathbf{m}_\mathbf{x} \in \mathbb{R}^Q$ and $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ and $\mathbf{G}_\mathbf{x} \in \mathbb{R}^{Q\times Q}$ are semi-definite positive 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 $\mathbf{m}$ and precision matrix $\mathbf{G}$ where $\mathbf{m}$ and $\mathbf{G}$ are defined as follows:

$$\mathbf{G} = \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H} + \mathbf{G}_\mathbf{x} \tag{4.1}$$

$$\mathbf{m} = \mathbf{G}^{-1} \left(\mathbf{H}^\top \mathbf{\Lambda} \mathbf{z} + \mathbf{G}_\mathbf{x} \mathbf{m}_\mathbf{x} \right). \tag{4.2}$$

A common solution is to use the Cholesky factorization of the covariance or the precision matrix [Rue, 2001]. However, when implemented through a Gibbs sampler, this method is of a 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, spending such computational time at each iteration of the Gibbs sampler to compute the Cholesky decomposition of the updated matrix may reduce the convergence speed of

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

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 [Geman and Yang, 1995; Chellappa and Chatterjee, 1985] or sparse [Rue and Held, 2005]). 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 domain. Other recently proposed algorithms are based on two-step approaches named Perturbation-Optimization [Bardsley, 2012; Papandreou and Yuille, 2010; Orieux et al., 2012; Gilavert et al., 2015; Parker, 2012; Féron et al., 2016], 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 linear system can be computed using iterative methods such as conjugate gradient algorithms leading to an approximate sample [Papandreou and Yuille, 2010; Orieux et al., 2012]. This issue has been considered in [Gilavert et al., 2015] by adding a Metropolis step in the sampling algorithm. In [Parker, 2012; Féron et al., 2016], the authors propose to reduce the computational cost by sampling along mutually conjugate directions instead of the initial high dimensional space.

Designing efficient proposals in MH algorithms: Non-Gaussian models arise in numerous applications in inverse problems [Lasanen, 2012; Bach et al., 2012; Kamilov et al., 2012; Kolehmainen et al., 2012]. In this context, the posterior distribution is non-Gaussian and does not generally belong to common probability models. 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 the MH algorithm is to construct a proposal density that simultaneously 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, that is, at each iteration, the proposal density $\mathbf{g}(.|\mathbf{x}^{(t)})$ is a Gaussian centered at the current state $\mathbf{x}^{(t)}$ and with covariance matrix $\mathbf{Q}(\mathbf{x}^{(t)})$. Other sampling algorithms are improved by incorporating information about the derivative of the logarithm of the target distribution to guide the Markov chain toward the target space where most of samples should be concentrated. For example, in Langevin-based algorithms, the mean of the Gaussian proposal density is replaced with one iteration of a preconditioned gradient descent algorithm. However, it is worth noting that the choice of the scale matrix $\mathbf{Q}$ may deeply affect the efficiency of these algorithms. 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 preferred especially in high dimensional spaces. In contrast, in the case of low dimensional problems and when the coefficients of the signal are not highly correlated, the standard RW and MALA algorithms defined for $\mathbf{Q} \equiv \mathbf{I}_Q$ achieve generally 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 law, they can be either sampled independently using RW or jointly by resorting to MALA. However, these algorithms may be not accurate for large scale problems especially when the coefficients of the signal exhibit high correlations. In this case, the design of a good proposal often requires considering the curvature of the target distribution. More sophisticated (and thus more computationally intensive) scale matrices should be chosen to guide the chain in the directions that reflect the dependence structure. Optimally, the curvature matrix should be chosen such as 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. When the minus-log of the target density can be expressed as in (3.12), good candidates of the curvature matrix take the following form:

$$\mathbf{Q} = \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H} + \mathbf{V}^\top \mathbf{\Omega} \mathbf{V} \tag{4.3}$$

where $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are semi-definite positive matrices. Feasible numerical factorization of $\mathbf{Q}$ can be ensured if $\mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}$ and $\mathbf{V}^\top \mathbf{\Omega} \mathbf{V}$ are diagonalizable in the same domain. Otherwise, the use of such matrix remains generally of limited interest especially for large scale problems where the manipulation of the resulting proposal generally induces a high computational complexity at the expense of the convergence speed. Alternatively, under mild conditions about the posterior density, MM strategy offers a large flexibility for building curvatures matrices with a lower computational cost (e.g., diagonal matrices, bloc-diagonal matrices, circulant...) as it has already been presented in Chapter 3. However, MH algorithms with too simple preconditioning matrices resulted 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 it induced in the algorithm and the accuracy and the closeness of the proposal to the true distribution.

It can be noted that the main difficulty arising in the last two sampling problems is mainly 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 very wide neighborhood even if the coefficients of the signal are supposed to

be statistically independent 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 domain as $\mathbf{H}$ which is the case of most real problems. One should therefore treat these two sources of correlations separately. One appealing idea is to eliminate one of these sources of correlation directly related to $\mathbf{x}$ by adding some auxiliary variables.

1.2 Auxiliary variables and data augmentation strategies

Indeed, to improve the mixing of sampling algorithms, many works have proposed to add some auxiliary variables to the initial model 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 [Tanner and Wong, 1987] and, auxiliary variables strategies [Mira and Tierney, 1997]. It is worthwhile to note 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) [Van Dyk and Meng, 2012]. In the following, we will use the term Data Augmentation (DA) to refer to any method that constructs sampling algorithms via introducing auxiliary variables. Some DA algorithms have been proposed in [Robert, 2013; Doucet et al., 2005; Févotte et al., 2011; Giovannelli, 2008; David, 1997; Hurn, 1997; Damlen et al., 1999]. A specific attention has been turned towards the Hamiltonian MCMC (HMC) approach [Duane et al., 1987; Girolami and Calderhead, 2011], that defines auxiliary variables based on physically inspired dynamics.

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

$$\mathcal{J}(\mathbf{x},\mathbf{u}) = \mathcal{J}(\mathbf{u}|\mathbf{x}) + \mathcal{J}(\mathbf{x}) \tag{4.4}$$

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

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

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 (positive and with integral with respect to $\mathbf{x}$ equal to 1). In fact, the importance of the first condition is obvious because the latent variable is only introduced for computational purposes and should not alter the considered initial model. The need of the second requirement 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 (4.4) provided that $\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z})$ is a density of a proper distribution (positive and with integral with respect to $\mathbf{u}$ equal to 1). For the second condition, it is sufficient to choose $\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z})$ such that $\mathsf{p}(\mathbf{x}|\mathbf{u}, \mathbf{z})$ remains a valid probability density function.

Instead of simulating directly from $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$, we now draw alternatively samples from the conditional distributions of the two variables $\mathbf{x}$ and $\mathbf{u}$ of densities $\mathcal{P}_{\mathbf{x}|\mathbf{u},\mathbf{z}}$ and $\mathcal{P}_{\mathbf{u}|\mathbf{x},\mathbf{z}}$ in an arbitrary order. This simply reduces to a special case of an 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 the required conditions [Geman and Geman, 1984; Gilks et al., 1999], the constructed chain $\left(\mathbf{x}^{(t)}, \mathbf{u}^{(t)}\right)_{t\geqslant 0}$ has as stationary distribution $\mathcal{P}_{\mathbf{x},\mathbf{u}|\mathbf{z}}$.

The usefulness of 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 obtained conditional distribution $\mathcal{P}_{\mathbf{x}|\mathbf{u},\mathbf{z}}$ and $\mathcal{P}_{\mathbf{u}|\mathbf{x},\mathbf{z}}$ is much easier than sampling directly from the initial distribution $\mathcal{P}_{\mathbf{x}|\mathbf{z}}$. Moreover, 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 will be 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 will only intervene through the auxiliary variable $\mathbf{u}$ and $\mathbf{z}$. Note that, half quadratic approaches [Idier, 2001; Ciuciu and Idier, 2002; Geman and Yang, 1995; Geman and Reynolds, 1992; Champagnat and Idier, 2004; Nikolova and Ng, 2005] had already motivated the introduction of auxiliary variables in optimization or sampling algorithms. For instance, a similar approach had already been proposed in [Bect et al., 2004] in the case of uncorrelated Gaussian noise with covariance matrix $\sigma^2\mathbf{I}_N$ and was used in some variational applications for image restoration. Moreover, this technique has been adopted to facilitate sampling using classical MH algorithm and Gibbs sampler in the maximum likelihood estimation approach proposed in [Cavicchioli et al., 2013]. Similarly, in [Ciuciu, 2000], the prior distribution has been replaced with a new one involving additional variables based on half-quadratic

formulation and inferences have been deduced according to the new resulting posterior distribution.

In this work, we propose a new formulation of the method introduced in [Bect et al., 2004] and we extend it to more general models and sampling algorithms. In the following, we will consider some examples and discuss how this approach can be applied.

§ 2 Proposed approach

2.1 Correlated Gaussian noise

We consider the linear observation model in (2.2) and we focus on the case when the noise $\mathbf{w}$ is additive, independent from the signal and Gaussian that is $\mathbf{w} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{\Lambda}^{-1})$ where $\mathbf{\Lambda} \in \mathbb{R}^{N \times N}$ is a known semi-definite positive precision matrix.

Hence, the minus logarithm of the posterior density has typically the following form:

$$\mathcal{J}(\mathbf{x}) = \frac{1}{2} (\mathbf{H}\mathbf{x} - \mathbf{z})^\top \mathbf{\Lambda} (\mathbf{H}\mathbf{x} - \mathbf{z}) + \Psi(\mathbf{V}\mathbf{x}) \tag{4.5}$$

where $\Psi(\mathbf{V}\mathbf{x}) = -\log \mathsf{p}(\mathbf{x})$ and $\mathbf{V}$ is a linear transform operator that can correspond for example to a frame analysis or to a discrete gradient matrix (see Chapter 2).

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 [Girolami and Calderhead, 2011]. 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 $\Psi(\mathbf{x}) = \sum_{i=1}^{Q} \psi_i(x_i)$, MALA and RW algorithms may behave poorly as they do not take into account data fidelity dependencies while Langevin algorithms with complicated curvature matrices may have high computational load due the presence of heterogeneous dependencies [Marnissi et al., 2015] (See Chapter 3).

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

$$\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z}) = \frac{\det(\mathbf{C})^{-1/2}}{(2\pi)^{J/2}} \exp\left(-\frac{1}{2} \|\mathbf{C}^{-1/2} (\mathbf{u} - \mathbf{A}\mathbf{x})\|^2\right) \tag{4.6}$$

where $\mathbf{C} \in \mathbb{R}^{J \times J}$ is a positive definite covariance matrix and $\mathbf{A} \in \mathbb{R}^{J \times Q}$. Note that, since $\mathsf{p}(\mathbf{x}|\mathbf{z})$ is positive and integrable with respect to $\mathbf{x}$ and $\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z})$ is bounded with respect to $\mathbf{x}$, the product $\mathsf{p}(\mathbf{u}|\mathbf{x}, \mathbf{z})\ \mathsf{p}(\mathbf{x}|\mathbf{z})$ remains

integrable with respect to $\mathbf{x}$. Then, the joint distribution satisfies the two conditions defined in Section 1 and its minus logarithm has the following expression:

$$\begin{aligned}\mathcal{J}(\mathbf{x},\mathbf{u}) = &\frac{1}{2}\mathbf{x}^\top\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}+\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}\right)\mathbf{x}\\ &+\frac{1}{2}\left(\mathbf{z}^\top\mathbf{\Lambda}\mathbf{z}+\mathbf{u}^\top\mathbf{C}^{-1}\mathbf{u}-2\mathbf{x}^\top\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z}+\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}\right)\right)\\ &+\Psi(\mathbf{V}\mathbf{x}).\end{aligned} \tag{4.7}$$

From (4.7), we can identify two sources of correlations directly related to the target signal. The first one comes from the first term through $\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}+\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}$ and the second one comes from the prior information through the operator $\mathbf{V}$ (and possibly additional correlation related for example to the non-separability of Ψ).

Let us define

$$\mathbf{Y} = \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}+\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}. \tag{4.8}$$

The key point is to set $\mathbf{A}$ and $\mathbf{C}$ such that $\mathbf{Y}$ has a simple structure. Note also that $\mathbf{Y}-\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H} = \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}$ should be a semi-definite positive matrix. Actually, the choice of the auxiliary variable and, hence, $\mathbf{Y}$ is subjective and is related to specifying the source of heterogeneous dependencies that one wants to eliminate in the target distribution with respect to the properties of $\mathbf{H}$, $\mathbf{\Lambda}$, $\mathbf{V}$ and Ψ. More specifically, one should identify if the main difficulty comes either from the matrix $\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}$ or only from the precision matrix $\mathbf{\Lambda}$.

Alternative I: Eliminate the coupling induced by $\mathbf{\Lambda}$

This problem is encountered for example for Model (4.1) with circulant matrices $\mathbf{H}$ and $\mathbf{G}_\mathbf{x}$ and with $\mathbf{\Lambda} \neq \mathbf{I}_N$ which induces further correlation when passing to the Fourier domain. Hence, one can eliminate the correlations induced by $\mathbf{\Lambda}$ by setting

$$\mathbf{Y} = \frac{1}{\mu}\mathbf{H}^\top\mathbf{H} \tag{4.9}$$

where $\mu > 0$ is such that $\mu\|\mathbf{\Lambda}\| < 1$. This is equivalent to choosing $\mathbf{A}$ and $\mathbf{C}$ such that

$$\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A} = \mathbf{H}^\top\left(\frac{1}{\mu}\mathbf{I}_N-\mathbf{\Lambda}\right)\mathbf{H}. \tag{4.10}$$

Hence, the minus logarithm of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ and $\mathbf{u}$ reads up to an additive constant:

$$\mathcal{J}(\mathbf{x}|\mathbf{u}) = \frac{1}{2\mu}\|\mathbf{H}\mathbf{x}\|^2-\mathbf{x}^\top\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z}+\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}\right)+\Psi(\mathbf{V}\mathbf{x}). \tag{4.11}$$

In addition, it can be noted that we do not need to compute directly the auxiliary variable $\mathbf{u}$ as it is not the variable of interest. In the Gibbs sampling

algorithm, the auxiliary variable $\mathbf{u}^{(t)}$ at each iteration t only intervenes in the product $\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}^{(t)}$ in (4.11). According to (4.6), we have:

$$\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}^{(t)} = \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}\mathbf{x}^{(t)} + \mathbf{A}^\top\mathbf{C}^{-1/2}\mathbf{n}^{(t)} \tag{4.12}$$

where $\mathbf{n}^{(t)} \sim \mathcal{N}(\mathbf{0}_J, \mathbf{I}_J)$. Since $\mathbf{A}$ and $\mathbf{C}$ satisfy (4.10), we obtain

$$\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}^{(t)} = \mathbf{H}^\top\left(\frac{1}{\mu}\mathbf{I}_N - \mathbf{\Lambda}\right)\mathbf{H}\mathbf{x}^{(t)} + \mathbf{A}^\top\mathbf{C}^{-1/2}\mathbf{n}^{(t)}. \tag{4.13}$$

Note that $\mathbf{A}^\top\mathbf{C}^{-1/2}\mathbf{n}^{(t)}$ follows the centered Gaussian distribution with covariance matrix $\mathbf{H}^\top\left(\frac{1}{\mu}\mathbf{I}_N - \mathbf{\Lambda}\right)\mathbf{H}$. Let $\mathbf{\Gamma} = \frac{1}{\mu}\mathbf{I}_N - \mathbf{\Lambda}$. It follows that

$$\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}^{(t)} = \mathbf{H}^\top\mathbf{v}^{(t)} \tag{4.14}$$

where

$$\mathbf{v}^{(t)} \sim \mathcal{N}\left(\mathbf{\Gamma}\mathbf{H}\mathbf{x}^{(t)}, \mathbf{\Gamma}\right). \tag{4.15}$$

Then, the minus logarithm of the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ and the new auxiliary variable $\mathbf{v}$ is given by

$$\mathcal{J}(\mathbf{x}|\mathbf{v}) = \frac{1}{2\mu}\|\mathbf{H}\mathbf{x} - \mu\left(\mathbf{\Lambda}\mathbf{z} + \mathbf{v}\right)\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{4.16}$$

The main steps of the proposed Gibbs sampling algorithm are given in Algorithm 5. The appealing advantage of this algorithm with respect to a Gibbs sampler which would be applied directly to Model (4.1) when $\mathbf{H}$ and $\mathbf{G}_\mathbf{x}$ are diagonalizable in the same domain, is that it allows to easily handle the case when $\mathbf{\Lambda}$ is not equal to a diagonal matrix having identical diagonal elements.

Algorithm 5 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\mathbf{\Lambda}$.

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

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

Note that, minimizing (4.16) can be seen as a restoration problem with an uncorrelated noise of variance μ. Step 3 in Algorithm 5 can be more easily implemented in the transform domain where $\mathbf{H}$ and $\mathbf{V}$ are diagonalized (see Section 4 for example).

Alternative II: Eliminate the coupling induced by $\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}$

In most real problems, $\mathbf{H}$ and $\mathbf{V}$ have different properties. While $\mathbf{H}$ almost represents a blur, a projection or a decimation matrix, $\mathbf{V}$ may model a wavelet transform, discrete differences, which makes the posterior covariance matrix have a complicated form. In such cases, one can eliminate the source of correlations related to $\mathbf{x}$ through $\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}+\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A}$, by setting $\mathbf{Y} = \frac{1}{\mu}\mathbf{I}_Q$, so that $\mathbf{A}$ and $\mathbf{C}$ satisfy

$$\mathbf{A}^\top\mathbf{C}^{-1}\mathbf{A} = \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H} \tag{4.17}$$

where $\mu > 0$ is such that $\mu\|\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\| < 1$.

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 follows:

$$\mathcal{J}(\mathbf{x}|\mathbf{u}) = \frac{1}{2\mu}\|\mathbf{x}\|^2 - \mathbf{x}^\top\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}\right) + \Psi(\mathbf{V}\mathbf{x}). \tag{4.18}$$

Similarly, we propose to use the following change of variables:

$$\mathbf{v}^{(t)} = \mathbf{A}^\top\mathbf{C}^{-1}\mathbf{u}^{(t)}.$$

According to (4.17), we obtain

$$\mathbf{v}^{(t)} = \left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\right)\mathbf{x}^{(t)} + \mathbf{A}^\top\mathbf{C}^{-1/2}\mathbf{n}^{(t)} \tag{4.19}$$

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

$$\mathbf{v}^{(t)} \sim \mathcal{N}\left(\mathbf{\Gamma}\mathbf{x}^{(t)}, \mathbf{\Gamma}\right) \tag{4.20}$$

and the new target conditional distribution reads

$$\mathcal{J}(\mathbf{x}|\mathbf{v}) = \frac{1}{2\mu}\|\mathbf{x} - \mu\left(\mathbf{v} + \mathbf{H}^\top\mathbf{\Lambda}\mathbf{z}\right)\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{4.21}$$

The proposed Gibbs sampling algorithm in this case is summarized in Algorithm 6.

Note that in (4.21), 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 $\mathbf{z}$. The original problem reduces to solving a denoising problem where the variance of the Gaussian noise is μ. Thereby, the new target distribution

Algorithm 6 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**

(4.21) is generally simpler to sample from compared to the initial one. In the particular case when the coefficients of the signal are uncorrelated in the prior law, one can sample them independently. Otherwise, when Ψ is a smooth function, one can use a Langevin-based MCMC algorithm. For instance, it is possible to construct an efficient curvature matrix that takes into account the prior correlation and that can be easily manipulated.

It is worth noting that the auxiliary variable can be introduced in the data fidelity term as well as in the prior information. The derivation of the proposed method in (4.7) allows us to identify classes of models for which our approach can be extended. Obviously, the key 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 [Andrews and Mallows, 1974; West, 1987] 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 [Geman and Geman, 1984]. In Section 2.2, we will investigate the case of scale mixture of Gaussian models.

In particular, when both the likelihood and the prior distribution are Gaussian conditionally to some parameters, the proposed method can be applied to each term. In Section 2.3, we will propose a Gibbs algorithm to address the problem of sampling from high dimensional Gaussian distributions.

Note that another step should be added to the Gibbs algorithm to sample the auxiliary variable $\mathbf{v}$. In Algorithm 5, it suffices to sample from the Gaussian distribution with covariance matrix $\left(\frac{1}{\mu}\mathbf{I}_N - \mathbf{\Lambda}\right)$. In Algorithm 6, we should be able to sample from the Gaussian distribution whose covariance

matrix is of the form $\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top \mathbf{\Lambda} \mathbf{H}\right)$, which is possible for a large class of observation models as it will be discussed in Section 2.4.

2.2 Scale mixture of Gaussian noise

2.2.1 Problem formulation

Let us consider the following observation model:

$$(\forall i \in \{1, \ldots, N\}) \quad z_i = [\mathbf{H}\mathbf{x}]_i + w_i \tag{4.22}$$

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

$$\begin{cases} w_i = 0 & \text{if } \sigma_i = 0 \\ w_i \sim \mathcal{N}(0, \sigma_i^2) & \text{if } \sigma_i > 0 \end{cases} \tag{4.23}$$

where $(\sigma_1, \ldots, \sigma_N)$ are independent realizations of a random variable in $\mathbb{R}^+$ distributed 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 $\sigma_1^2, \ldots, \sigma_N^2$ are realizations of a random variable following an inverse Gamma distribution. Let $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_N]^\top$. The joint posterior density of $\mathbf{x}$ and $\boldsymbol{\sigma}$ is given by:

$$\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}|\mathbf{z}) = \mathsf{p}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{z})\mathsf{p}(\boldsymbol{\sigma}). \tag{4.24}$$

In such a Bayesian inverse problem, 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 $\mathbf{z}$, i.e.,

$$\int_{\mathcal{S}_0} \prod_{i=1}^{N} \mathsf{p}(\sigma_i|z_i)\mathrm{d}\sigma_i = 0. \tag{4.25}$$

Note that by imposing (4.25), we ensure that, at each iteration t of the Gibbs algorithm, $\boldsymbol{\sigma}^{(t)} \neq \mathbf{0}_N$.

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.

The first property is satisfied since $\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z}) = \mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}|\mathbf{z})\mathsf{p}(\mathbf{v}|\mathbf{x}, \boldsymbol{\sigma}, \mathbf{z})$ provided that $\mathsf{p}(\mathbf{v}|\mathbf{x}, \boldsymbol{\sigma}, \mathbf{z})$ is a density of a proper distribution (positive and with integral with respect to $\mathbf{v}$ equal to 1). The second property means that $\mathcal{P}(\mathbf{x}, \boldsymbol{\sigma}|\mathbf{v}, \mathbf{z})$ should define a proper distribution, that is, $\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z})$ has to be integrable with respect to $\mathbf{x}$ and $\boldsymbol{\sigma}$. 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 $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)},\mathbf{z}}$.

2.2.2 Proposed algorithms

At each iteration t of the Gibbs sampler, let $\mathbf{D}^{(t)}$ be the diagonal matrix whose elements are given by:

$$(\forall i \in \{1, \dots, N\}) \quad D_{ii}^{(t)} = \begin{cases} 0 & \text{if } \sigma_i^{(t)} = 0 \\ \left(\sigma_i^{(t)}\right)^{-2} & \text{if } \sigma_i^{(t)} > 0. \end{cases} \tag{4.26}$$

Note that from (4.25), we have:

$$\|\mathbf{D}^{(t)}\| > 0. \tag{4.27}$$

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

$$(\forall t \geqslant 0)\,(\forall i \in \{1, \dots, N\}) \quad \nu \leqslant \sigma_i^{(t)}. \tag{4.28}$$

  Then, results in Section 2.1 with a Gaussian noise can be extended to scale mixture of Gaussian noise by setting at each iteration t, $\boldsymbol{\Lambda} = \mathbf{D}^{(t)}$, $\mu < \nu^2$ in Algorithm 5 and $\mu\|\mathbf{H}^\top\mathbf{H}\| < \nu^2$ in Algorithm 6. 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 (4.28) does not exist, results in Section 2.1 remain also valid when, at each iteration t, for a given value of $\boldsymbol{\sigma}^{(t)}$, we replace $\boldsymbol{\Lambda}$ by $\mathbf{D}^{(t)}$. There are two differences. The first difference is that μ depends on the value of the mixing variable $\boldsymbol{\sigma}^{(t)}$ and hence takes different values throughout the algorithm. Subsequently, $\mu^{(t)}$ will denote the value of μ in each iteration t of the Gibbs sampler. The second difference is that another step is added to sample the mixing variables $\sigma_1, \dots, \sigma_N$ from their distributions conditioned to $\mathbf{x}$, $\mathbf{v}$ and $\mathbf{z}$.

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

At each iteration, $\mu^{(t)} > 0$ is chosen such that $\mu^{(t)}\|\mathbf{D}^{(t)}\| < 1$ and the auxiliary variable is drawn from

$$\mathbf{v}^{(t)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}^{(t)}\mathbf{H}\mathbf{x}^{(t+1)}, \boldsymbol{\Gamma}^{(t)}\right) \tag{4.29}$$

where $\mathbf{\Gamma}^{(t)} = \frac{1}{\mu^{(t)}}\mathbf{I}_N - \mathbf{D}^{(t)}$.

The minus logarithm of the posterior density $\mathsf{p}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}, \mathbf{z})$ is given by

$$\mathcal{J}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}) = \frac{1}{2\mu}\|\mathbf{H}\mathbf{x} - \mu\left(\mathbf{v} + \mathbf{D}\mathbf{z}\right)\|^2 + \Psi(\mathbf{V}\mathbf{x}), \tag{4.30}$$

where μ and $\mathbf{D}$ are related to $\boldsymbol{\sigma}$.

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

Similarly, in order to eliminate the coupling induced by the matrix $\mathbf{H}^\top\mathbf{D}^{(t)}\mathbf{H}$, $\mu^{(t)}$ is chosen at each iteration t so as to satisfy $\mu^{(t)}\|\mathbf{H}^\top\mathbf{D}^{(t)}\mathbf{H}\| < 1$. Then, the auxiliary variable is drawn from

$$\mathbf{v}^{(t)} \sim \mathcal{N}\left(\mathbf{\Gamma}^{(t)}\mathbf{x}^{(t+1)}, \mathbf{\Gamma}^{(t)}\right) \tag{4.31}$$

where $\mathbf{\Gamma}^{(t)} = \frac{1}{\mu^{(t)}}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{D}^{(t)}\mathbf{H}$.

The minus logarithm of the posterior density $\mathsf{p}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}, \mathbf{z})$ is given by

$$\mathcal{J}(\mathbf{x}|\boldsymbol{\sigma}, \mathbf{v}) = \frac{1}{2\mu}\|\mathbf{x} - \mu\left(\mathbf{v} + \mathbf{H}^\top\mathbf{D}\mathbf{z}\right)\|^2 + \Psi(\mathbf{V}\mathbf{x}). \tag{4.32}$$

It is worth noting that in (4.29) and (4.31), the mixing variable $\boldsymbol{\sigma}^{(t)}$ at each iteration t is presented implicitly through $\mathbf{D}^{(t)}$ and also $\mu^{(t)}$ which makes $\boldsymbol{\sigma}$ and $\mathbf{v}$ two random variables dependent conditionally to $\mathbf{x}$ and $\mathbf{z}$. In the following, we will give a sufficient condition on the choice of $\mu^{(t)}$, under which, $\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z})$ still defines the density of a proper distribution.

Proposition 2.1 *Suppose that, at each iteration t, the auxiliary variable follows (4.29) with $\mu^{(t)} = \epsilon\|\mathbf{D}^{(t)}\|^{-1}$ (respectively (4.31) with $\mu^{(t)} = \frac{\epsilon}{\|\mathbf{H}\|^2\|\mathbf{D}^{(t)}\|}$) where ϵ is a constant chosen such that $0 < \epsilon < 1$. Then, $\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z})$ is integrable with respect to $\mathbf{x}$ and $\boldsymbol{\sigma}$.*

Proof. We focus on the general case when the auxiliary variable follows (4.31). We have then

$$\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{v}|\mathbf{z}) = \mathsf{p}(\mathbf{v}|\mathbf{x}, \boldsymbol{\sigma}, \mathbf{z})\mathsf{p}(\mathbf{x}, \boldsymbol{\sigma}|\mathbf{z}). \tag{4.33}$$

Since the matrix $\mathbf{H}^\top\mathbf{D}\mathbf{H}$ is positive, the spectrum of $\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{D}\mathbf{H}$ satisfies

$$\text{Spec}\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{D}\mathbf{H}\right) \subset [a, b] \tag{4.34}$$

where

$$a = \frac{1}{\mu} - \|\mathbf{H}\|^2\|\mathbf{D}\| = \left(\frac{1}{\epsilon} - 1\right)\|\mathbf{H}\|^2\|\mathbf{D}^{(t)}\|$$

and

$$b = \frac{1}{\mu}.$$

We have

$$\|\mathbf{D}\| = \left(\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}\right)^{-2}. \tag{4.35}$$

Note that the existence of $\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}$ in (4.35) follows from (4.25).

As $\mathsf{p}(\mathbf{v}|\mathbf{x},\boldsymbol{\sigma})$ is the density of a proper Gaussian distribution, we can show that

$$\begin{aligned}\mathsf{p}(\mathbf{v}|\mathbf{x},\boldsymbol{\sigma}) &\leqslant (2\pi)^{-Q/2}\det\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{D}\mathbf{H}\right)^{-1/2} \\ &\leqslant (2\pi a)^{-Q/2} \\ &= C\left(\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}\right)^{Q}\end{aligned} \tag{4.36}$$

where C is a positive constant that only depends on $\mathbf{H}$ and ϵ (i.e., independent of $\mathbf{v}$, $\mathbf{x}$ and $\boldsymbol{\sigma}$).

Then,

$$\begin{aligned}\int_{\mathbb{R}^N}\int_{\mathbb{R}^Q}\mathsf{p}(\mathbf{x},\boldsymbol{\sigma},\mathbf{v}|\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\sigma} &\leqslant C\int_{\mathbb{R}^N}\left(\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}\right)^{Q}\int_{\mathbb{R}^Q}\mathsf{p}(\mathbf{x},\boldsymbol{\sigma}|\mathbf{z})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\sigma} \\ &= C\int_{\mathbb{R}^N}\left(\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}\right)^{Q}\prod_{i=1}^{N}\mathsf{p}(\sigma_i|z_i)\mathrm{d}\sigma_i \\ &= C\mathsf{E}_{\boldsymbol{\sigma}|\mathbf{z}}\left[\left(\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}\right)^{Q}\right] \\ &\leqslant C\mathsf{E}_{\boldsymbol{\sigma}|\mathbf{z}}\left[\left(\min(\sigma_i)_{\substack{1\leqslant i\leqslant N \\ s.t.\ \sigma_i>0}}\right)\right]^{Q}.\end{aligned} \tag{4.37}$$

Since the moments of the posterior distribution $\mathsf{p}(\boldsymbol{\sigma}|\mathbf{z})$ are assumed to be finite, then Proposition 2.1 holds. □

The resulting Gibbs samplers are summarized in Algorithms 7 and 8.

2.2.3 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 μ and $\mathbf{D}$ in the conditional distribution of $\mathbf{v}$. One

Algorithm 7 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\mathbf{D}$ in the case of a scale mixture of Gaussian noise.

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^N$, $\boldsymbol{\sigma}^{(0)} \in \mathbb{R}_+^N$, $0 < \epsilon < 1$, $\mu^{(0)} = \epsilon \left(\min(\sigma_i^{(0)})_{\substack{1 \leqslant i \leqslant N \\ s.t.\ \sigma_i^{(0)} > 0}} \right)^2$

1: **for** $t = 0, 1, \ldots$ **do**
2: Generate $\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}^{(t)}\mathbf{H}\mathbf{x}^{(t)}, \boldsymbol{\Gamma}^{(t)}\right)$ where $\boldsymbol{\Gamma}^{(t)} = \dfrac{1}{\mu^{(t)}}\mathbf{I}_N - \mathbf{D}^{(t)}$
3: Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\boldsymbol{\sigma}^{(t)},\mathbf{z}}$
4: For all $i \in \{1, \ldots, N\}$, generate $\sigma_i^{(t+1)} \sim \mathcal{P}_{\sigma_i|\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)},\mathbf{z}}$
5: Set $\mu^{(t+1)} = \epsilon \|\mathbf{D}^{(t+1)}\|^{-1}$
6: **end for**

Algorithm 8 Gibbs sampler with auxiliary variables in order to eliminate the coupling induced by $\mathbf{H}^\top\mathbf{D}\mathbf{H}$ in the case of a scale mixture of Gaussian noise.

Initialize: $\mathbf{x}^{(0)} \in \mathbb{R}^Q$, $\mathbf{v}^{(0)} \in \mathbb{R}^Q$, $\boldsymbol{\sigma}^{(0)} \in \mathbb{R}_+^N$, $0 < \epsilon < 1$, $\mu^{(0)} = \epsilon \, \|\mathbf{H}\|^{-2} \left(\min(\sigma_i^{(0)})_{\substack{1 \leqslant i \leqslant N \\ s.t.\ \sigma_i^{(0)} > 0}} \right)^2$

1: **for** $t = 0, 1, \ldots$ **do**
2: Generate $\mathbf{v}^{(t+1)} \sim \mathcal{N}\left(\boldsymbol{\Gamma}^{(t)}\mathbf{x}^{(t)}, \boldsymbol{\Gamma}^{(t)}\right)$ where $\boldsymbol{\Gamma}^{(t)} = \dfrac{1}{\mu^{(t)}}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{D}^{(t)}\mathbf{H}$
3: Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\boldsymbol{\sigma}^{(t)},\mathbf{z}}$
4: For all $i \in \{1, \ldots, N\}$, generate $\sigma_i^{(t+1)} \sim \mathcal{P}_{\sigma_i|\mathbf{x}^{(t+1)},\mathbf{v}^{(t+1)},\mathbf{z}}$
5: Set $\mu^{(t+1)} = \epsilon \|\mathbf{H}^\top\mathbf{H}\|^{-1} \|\mathbf{D}^{(t+1)}\|^{-1}$
6: **end for**

may 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. However, as $\boldsymbol{\sigma}$ is sampled independently from $\mathbf{v}$, the constructed Markov chain $\left(\mathbf{x}^{(t)}, \boldsymbol{\sigma}^{(t)}, \mathbf{v}^{(t)}\right)_{t \geqslant 0}$ may have a transition kernel with an unknown stationary distribution [Van Dyk and Park, 2008]. 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 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) [Van Dyk and Park, 2008]. Although this strategy can significantly decrease the complexity of the sampling process, it must be implemented with care to be sure that the desired stationary distribution is preserved. Applications of PCGS algorithms can be found in [Park and van Dyk, 2009; Costa et al., 2016; Kail et al., 2012].

Assume that, in addition to $\mathbf{x}$, $\boldsymbol{\sigma}$, $\mathbf{v}$, we have a set of unknown parameters $\boldsymbol{\Theta} \in \mathbb{R}^P$ 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 [Van Dyk and Park, 2008], 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 9 shows the main steps of the proposed sampler. More details can be found in Appendix A. However, 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 9 PCGS in the case of a scale mixture of Gaussian noise

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

1: **for** $t = 0, 1, \ldots$ **do**
2: For all $i \in \{1, \ldots, N\}$, generate $\sigma_i^{(t+1)} \sim \mathcal{P}_{\sigma_i|\mathbf{x}^{(t)},\boldsymbol{\Theta}^{(t)},\mathbf{z}}$
3: Generate $\boldsymbol{\Theta}^{(t+1)} \sim \mathcal{P}_{\boldsymbol{\Theta}|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t+1)},\mathbf{z}}$.
4: Set $\mu^{(t+1)}$ and generate $\mathbf{v}^{(t+1)} \sim \mathcal{P}_{\mathbf{v}|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t+1)},\boldsymbol{\Theta}^{(t+1)},\mathbf{z}}$
5: Generate $\mathbf{x}^{(t+1)} \sim \mathcal{P}_{\mathbf{x}|\mathbf{v}^{(t+1)},\boldsymbol{\sigma}^{(t+1)},\boldsymbol{\Theta}^{(t+1)},\mathbf{z}}$
6: **end for**

2.3 High dimensional Gaussian distribution

The proposed method can 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 (4.1) and (4.2).

In the following, we will give some examples where the introduction of auxiliary variables facilitates the sampling process.

- If the prior precision matrix $\mathbf{G}_\mathbf{x}$ and the observation matrix $\mathbf{H}$ can be diagonalized in the same domain, we introduce the auxiliary variable $\mathbf{v}_1$ in the data fidelity term. Following Algorithm 5, let $\mu_1 > 0$ such that $\mu_1 \|\boldsymbol{\Lambda}\| < 1$ and

$$\mathbf{v}_1 \sim \mathcal{N}\left(\left(\frac{1}{\mu_1}\mathbf{I}_N - \boldsymbol{\Lambda}\right)\mathbf{H}\mathbf{x}, \frac{1}{\mu_1}\mathbf{I}_N - \boldsymbol{\Lambda}\right). \tag{4.38}$$

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:

$$\tilde{\mathbf{G}} = \frac{1}{\mu_1}\mathbf{H}^\top\mathbf{H} + \mathbf{G}_\mathbf{x}. \tag{4.39}$$

$$\tilde{\mathbf{m}} = \tilde{\mathbf{G}}^{-1}\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{G}_\mathbf{x}\mathbf{m}_\mathbf{x} + \mathbf{H}^\top\mathbf{v}_1\right). \tag{4.40}$$

Then, sampling from the target signal can be performed by passing to the transform domain where $\mathbf{H}$ and $\mathbf{G}_\mathbf{x}$ are diagonalizable (a.e., Fourier domain when $\mathbf{H}$ and $\mathbf{G}_\mathbf{x}$ are circulant) .
Similarly, if it is possible to write $\mathbf{G}_\mathbf{x} = \mathbf{V}^\top\mathbf{\Omega}\mathbf{V}$, so as $\mathbf{H}$ and $\mathbf{V}$ can be diagonalized in the same domain, we introduce an extra auxiliary variable in the prior term to eliminate the coupling introduced by $\mathbf{\Omega}$ when passing to the transform domain. Let $\mu_2 > 0$ such that $\mu_2\|\mathbf{D}\| < 1$ and

$$\mathbf{v}_2 \sim \mathcal{N}\left(\left(\frac{1}{\mu_2}\mathbf{I}_N - \mathbf{\Omega}\right)\mathbf{V}\mathbf{x}, \frac{1}{\mu_2}\mathbf{I}_N - \mathbf{\Omega}\right). \tag{4.41}$$

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

$$\tilde{\mathbf{G}} = \frac{1}{\mu_1}\mathbf{H}^\top\mathbf{H} + \frac{1}{\mu_2}\mathbf{V}^\top\mathbf{V} \tag{4.42}$$

and

$$\tilde{\mathbf{m}} = \tilde{\mathbf{G}}^{-1}\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{G}_\mathbf{x}\mathbf{m}_\mathbf{x} + \mathbf{H}^\top\mathbf{v}_1 + \mathbf{V}^\top\mathbf{v}_2\right). \tag{4.43}$$

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

$$\mathbf{v}_1 \sim \mathcal{N}\left(\left(\frac{1}{\mu_1}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\right)\mathbf{x}, \frac{1}{\mu_1}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\right). \tag{4.44}$$

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

$$\tilde{\mathbf{G}} = \frac{1}{\mu_1}\mathbf{I}_Q + \mathbf{G}_\mathbf{x}. \tag{4.45}$$

$$\tilde{\mathbf{m}} = \tilde{\mathbf{G}}^{-1}\left(\mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{G}_\mathbf{x}\mathbf{m}_\mathbf{x} + \mathbf{v}_1\right). \tag{4.46}$$

Note that if $\mathbf{G}_\mathbf{x}$ has some simple structure (e.g,. diagonal, block diagonal, sparse, circulant,...), the precision matrix (4.45) will inherit this simple structure.
Otherwise, if $\mathbf{G}_\mathbf{x}$ does not present any specific structure, we can apply the proposed method to both data fidelity and prior terms. Additionally to

the auxiliary variable $\mathbf{v}_1$ in (4.44), we introduce an extra auxiliary variable $\mathbf{v}_2$ in the prior law. Let $\mu_2 > 0$ such that $\mu_2\|\mathbf{G}_\mathbf{x}\| < 1$ and

$$\mathbf{v}_2 \sim \mathcal{N}\left(\left(\frac{1}{\mu_2}\mathbf{I}_Q - \mathbf{G}_\mathbf{x}\right)\mathbf{x}, \frac{1}{\mu_2}\mathbf{I}_Q - \mathbf{G}_\mathbf{x}\right). \tag{4.47}$$

The joint distribution of the unknown parameters is given by

$$\mathsf{p}(\mathbf{x}, \mathbf{v}_1, \mathbf{v}_2|\mathbf{z}) = \mathsf{p}(\mathbf{x}|\mathbf{z})\mathsf{p}(\mathbf{v}_1|\mathbf{x}, \mathbf{z})\mathsf{p}(\mathbf{v}_2|\mathbf{x}, \mathbf{z}). \tag{4.48}$$

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:

$$\tilde{\mathbf{G}} = \frac{1}{\mu}\mathbf{I}_Q \tag{4.49}$$

and

$$\tilde{\mathbf{m}} = \mu\left(\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{H}^\top\mathbf{\Lambda}\mathbf{z} + \mathbf{G}_\mathbf{x}\mathbf{m}_\mathbf{x}\right) \tag{4.50}$$

where

$$\mu = \frac{\mu_1\mu_2}{\mu_1 + \mu_2}. \tag{4.51}$$

Note that the sampling steps of $\mathbf{x}$, $\mathbf{v}_1$ and $\mathbf{v}_2$ can be merged to an equivalent but more direct step as follows:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \mu\mathbf{G}(\mathbf{m} - \mathbf{x}^{(t)}) + \mu\mathbf{n} \tag{4.52}$$

where

$$\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}_Q, \frac{2}{\mu}\mathbf{I}_Q - \mathbf{G}\right). \tag{4.53}$$

2.4 Sampling the auxiliary variable

It is clear that the main issue in the implementation of the proposed Gibbs algorithms is the sampling of the auxiliary variable $\mathbf{v}$ from a multivariate Gaussian distribution with covariance matrix of the form $\mathbf{\Gamma} = \frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}$ where $\mu > 0$ is chosen to satisfy $\mu\|\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\| < 1$.

Let $\mu > 0$ and $\beta > 0$ be such that

$$\mu\|\mathbf{H}\|^2 < \beta < \frac{1}{\|\mathbf{\Lambda}\|}. \tag{4.54}$$

For example, we can set $\mu \leqslant \frac{\epsilon}{\|\mathbf{H}\|^2\|\mathbf{\Lambda}\|}$ and $\beta = \frac{\sqrt{\epsilon}}{\|\mathbf{\Lambda}\|}$ where $0 < \epsilon < 1$.

We have then $0 < \mu\|\mathbf{H}^\top\mathbf{\Lambda}\mathbf{H}\| < 1$ and

$$\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{\Lambda}\mathbf{H} = \frac{1}{\beta}\left(\frac{\beta}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right) + \mathbf{H}^\top\left(\frac{1}{\beta}\mathbf{I}_N - \mathbf{\Lambda}\right)\mathbf{H}. \tag{4.55}$$

As a result, the sampling step of the auxiliary variable in all the previously proposed algorithms can be replaced by the three following steps:

1. Generate $\mathbf{n}^{(t+1)} \sim \mathcal{N}\left(\mathbf{0}_N, \frac{1}{\beta}\mathbf{I}_N - \mathbf{\Lambda}\right)$.
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 $\lambda = \frac{\mu}{\beta} \leqslant \frac{\sqrt{\epsilon}}{\|\mathbf{H}\|^2}$.
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. Then, the sampling problem of the auxiliary variables is separated into two independent subproblems. When $\mathbf{\Lambda}$ is diagonal (e.g., when the model is a scale mixture of Gaussian variables), coefficients $n_i^{(t+1)}$, $i \in \{1, \ldots, N\}$, can be drawn separately. Note that, $\mathbf{\Lambda}$ has often a simple structure even when it is not diagonal. Then, direct sampling from the centered Gaussian distribution with covariance matrix $\frac{1}{\beta}\mathbf{I}_N - \mathbf{\Lambda}$ is usually often easy. In the following, we address the problem of sampling 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, when $\mathbf{H}^\top\mathbf{H}$ is diagonalizable in a known domain i.e, there exists an orthogonal matrix $\mathbf{N}$ such that $\mathbf{N}\mathbf{H}^\top\mathbf{H}\mathbf{N}^\top$ is a diagonal matrix with positive elements, sampling from the Gaussian distribution with covariance matrix $\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}$ can be fulfilled easily in the transform domain 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 for tight frame synthesis operators and decimation matrices. Note that $\nu\lambda \leqslant \sqrt{\epsilon} < 1$. We have then:

$$\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H} = \left(\frac{1}{\sqrt{\lambda}}\mathbf{I}_Q - \sqrt{\lambda}\mathbf{H}^\top\mathbf{H}\right)^2 + (1 - \lambda\nu)\,\mathbf{H}^\top\mathbf{H}. \tag{4.56}$$

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:

$$\mathbf{y}^{(t+1)} = \left(\frac{1}{\sqrt{\lambda}}\mathbf{I}_Q - \sqrt{\lambda}\mathbf{H}^\top\mathbf{H}\right)\mathbf{y}_1^{(t+1)} + \sqrt{1 - \lambda\nu}\mathbf{H}^\top\mathbf{y}_2^{(t+1)} \tag{4.57}$$

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

- Suppose that $\mathbf{H} = \mathbf{MP}$ with $\mathbf{M} \in \mathbb{R}^{N\times K}$ and $\mathbf{M} \in \mathbb{R}^{K\times Q}$. Hence, we propose to set $\lambda > 0$ and $\tilde{\lambda} > 0$ such that

$$\lambda\|\mathbf{P}\|^2 < \tilde{\lambda} < \frac{1}{\|\mathbf{M}\|^2}. \tag{4.58}$$

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

$$\frac{1}{\lambda}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H} = \frac{1}{\tilde{\lambda}}\left(\frac{\tilde{\lambda}}{\lambda}\mathbf{I}_Q - \mathbf{P}^\top\mathbf{P}\right) + \mathbf{P}^\top\left(\frac{1}{\tilde{\lambda}}\mathbf{I}_K - \mathbf{M}^\top\mathbf{M}\right)\mathbf{P}. \tag{4.59}$$

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

$$\mathbf{y}^{(t+1)} = \frac{1}{\sqrt{\tilde{\lambda}}}\mathbf{y}_1^{(t+1)} + \mathbf{P}^\top\mathbf{y}_2^{(t+1)}. \tag{4.60}$$

§ 3 Application to multichannel image recovery in the presence of Gaussian noise

Multichannel Images (MCI) are widely used in many application areas such as medical imaging and remote sensing. The multiple components are obtained by 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 times, MCI are corrupted with noise and blurred during the acquisition process and transmission steps. Therefore, restoring MCI is of primary importance for several applications such as classification, segmentation and object recognition [Delp and Mitchell, 1979]. In this work, we propose a Bayesian method for MCI recovery in the wavelet domain by jointly processing the spectral components.

3.1 Problem formulation

We are interested in the recovery of 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 a zero-mean additive Gaussian noise with known variance σ^2. We address the 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 $\mathbf{F}_1^*, \ldots, \mathbf{F}_B^*$ such as

$$(\forall b \in \{1, \ldots, B\}) \quad \bar{\mathbf{y}}_b = \mathbf{F}_b^* \bar{\mathbf{x}}_b \tag{4.61}$$

where $\mathbf{F}_b^*$ 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 $\bar{\mathbf{y}}_b$. Note that, 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 $\sum_{m=1}^{M} K_m = K$. We have

$$\begin{aligned}(\forall b \in \{1, \ldots, B\}) \quad \bar{\mathbf{x}}_b = \; & (\bar{x}_{b,1,1}, \ldots, \bar{x}_{b,1,K_1}, \ldots, \\ & \bar{x}_{b,m,1}, \ldots, \bar{x}_{b,m,K_m}, \ldots, \\ & \bar{x}_{b,M,1}, \ldots, \bar{x}_{b,M,K_M})^\top.\end{aligned} \tag{4.62}$$

Then, the problem can be formulated as in (2.2), that is:

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{4.63}$$

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

$$\mathbf{F}^* = \begin{pmatrix} \mathbf{F}_1^* & \mathbf{0} & \ldots & \mathbf{0} \\ 0 & \mathbf{F}_2^* & \mathbf{0} & \mathbf{0} \\ \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{F}_B^* \end{pmatrix} \tag{4.64}$$

and

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ 0 & \mathbf{B}_2 & \mathbf{0} & \mathbf{0} \\ \ddots & \ddots & \ddots & \ddots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_B \end{pmatrix}. \tag{4.65}$$

In this work, we propose to exploit the cross-component similarities by estimating jointly the frame coefficients of a specific orientation and scale through all the components B. In this respect, for all $m \in \{1, \ldots, M\}$, for all $k \in \{1, \ldots, K_m\}$, 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 (see Figure 4.1). Note that such vector can be easily obtained as follows $\mathbf{x}_{m,k} = \mathbf{P}_{m,k}\mathbf{x}$ where $\mathbf{P}_{m,k} \in \mathbb{R}^{B \times Q}$ is a sparse matrix containing B lines of a permutation matrix. To promote the sparsity of the wavelet coefficients and the inter-component dependency, following [Marnissi et al., 2013], we assume that for every $m \in \{1, \ldots, M\}$, the vectors $\mathbf{x}_{m,1}$, …, $\mathbf{x}_{m,K_m}$ are realizations of a random vector following a $\mathcal{GMEP}$ distribution

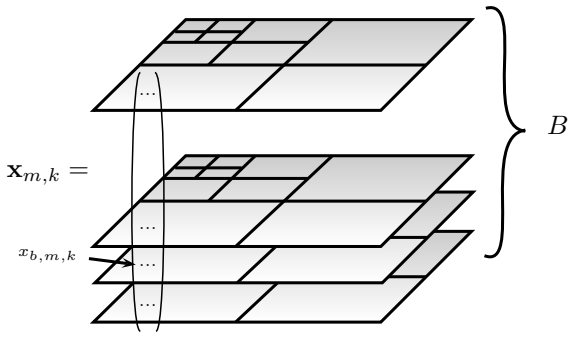


Figure 4.1: *Vector of wavelet coefficients in a multiscale wavelet basis decomposition.*

(see Chapter 2, Section 2.3.1) 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

$$-\log \mathsf{p}(\mathbf{x}|\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_M) = \sum_{m=1}^{M} \sum_{k=1}^{K_m} \psi_m(\|\mathbf{\Sigma}_m^{-1/2}(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|) \qquad (4.66)$$

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

In this work, we aim to compute the posterior mean estimate of the target image as well as the unknown regularization parameters using MCMC sampling algorithms. In the following, we denote by $\mathbf{\Theta}$ the set of the unknown regularization parameters to be estimated jointly with $\mathbf{x}$ in the Gibbs algorithm.

3.2 Sampling from the posterior distribution of the wavelet coefficients

We can expect that the standard sampling algorithms may 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 shapes (i.e., $\beta_m \leqslant \frac{1}{2}$) 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.

Note that for every $m \in \{1, \dots, M\}$, the function ψ_m is differentiable, and $t \mapsto \psi_m(\sqrt{t})$ is concave on $\mathbb{R}_+$ provided that $\beta_m \leqslant 1$. Thus, we propose to use 3MH algorithms to generate samples according to the posterior law. The sampling performance can be improved by adding a curvature matrix. The resulting proposal better strides the support of the target distribution by taking into account the correlation existing between the coefficients coming either from the likelihood or the prior law.

The curvature matrix is ideally the sum of two matrices (see Chapter 3). While the first matrix is related to the data fidelity term and addresses the spatial correlation between the coefficients belonging to the same spectral channel, the second one is related to the prior and deals with the spectral correlation existing between coefficients at the same spatial position and belonging to different spectral channels. Nevertheless, the manipulation of such high dimensional full curvature matrix may induce a high computational burden in the algorithm. One can use instead diagonal curvature matrices which result on rough local approximations of the posterior law. However, even if the computational cost of each iteration is reduced, the mixing properties of the chain may be deteriorated compared to the full curvature matrix.

We propose to tackle this preconditioning problem by adding auxiliary variables to the data fidelity term. More specifically, following Algorithm 6, we propose to introduce an auxiliary variable $\mathbf{v} \in \mathbb{R}^Q$ such that:

$$\mathbf{v} \sim \mathcal{N}\left(\frac{1}{\sigma^2}\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right)\mathbf{x}, \frac{1}{\sigma^2}\left(\frac{1}{\mu}\mathbf{I}_Q - \mathbf{H}^\top\mathbf{H}\right)\right) \tag{4.67}$$

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

Since the set of hyperparameters $\boldsymbol{\Theta}$ is independent of the auxiliary variable $\mathbf{v}$ when conditioned to $\mathbf{x}$, each iteration t 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 tight frame transform, sampling can be performed using results in (4.57) and (4.60).

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:

$$\mathsf{p}(\mathbf{x}|\mathbf{z},\mathbf{v},\boldsymbol{\Theta}) \propto \prod_{m=1}^{M}\prod_{k=1}^{K_m} \exp\left(-\mathcal{J}_{m,k}(\mathbf{P}_{m,k}\mathbf{x}|\mathbf{v})\right) \tag{4.68}$$

where

$$\begin{aligned}(\forall \mathbf{c}\in\mathbb{R}^B)\quad \mathcal{J}_{m,k}(\mathbf{c}|\mathbf{v}) = \frac{1}{2\mu\sigma^2}\|\mathbf{c}-\mu\mathbf{P}_{m,k}\mathbf{v}-\frac{\mu}{\sigma^2}\mathbf{P}_{m,k}\mathbf{H}^\top\mathbf{z}\|^2 \\ + \psi_m(\|\boldsymbol{\Sigma}_m^{-1/2}(\mathbf{c}-\mathbf{a}_m)\|).\end{aligned} \tag{4.69}$$

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. Note that we have generally $K\gg B$.

The advantage of the proposed method is twofold. First, instead of processing all the different wavelet coefficients at the same time, each subproblem can be fulfilled independently. 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 the 3MH algorithm with a curvature matrix constructed for each subband $m\in\{1,\ldots,M\}$ using a MM strategy, which has the following form:

$$(\forall \mathbf{c}\in\mathbb{R}^B)\quad \mathbf{Q}_m(\mathbf{c}) = \frac{1}{\mu}\mathbf{I}_B + \boldsymbol{\Sigma}_m^{-1}\psi_m'\left(\|\boldsymbol{\Sigma}_m^{-1/2}(\mathbf{c}-\mathbf{a}_m)\|\right). \tag{4.70}$$

The second advantage is that, since the problem dimension is reduced, we may expect that the discretization stepsize in the 3MH algorithm takes larger values compared with standard algorithms without auxiliary variables so that the chain makes larger moves and explores the target space faster and more efficiently.

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

3.3 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 may take 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 belonging to the first level of decomposition to relatively higher values ($0.5 < \beta_m < 1$) for details subband in 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 (4.66) is differentiable. We propose to estimate the scale matrices $\mathbf{\Sigma}_m$. Let $\mathcal{P}_{\mathbf{\Sigma}_m}$ be the prior distribution of the scale matrix for each subband $m \in \{1, \ldots, M\}$ and let $\mathsf{p}(\mathbf{\Sigma}_m)$ be the related density. Then, its posterior density reads:

$$\begin{aligned}\mathsf{p}(\mathbf{\Sigma}_m|\mathbf{x}) \propto \mathsf{p}(\mathbf{\Sigma}_m)\det(\mathbf{\Sigma}_m)^{-K_m/2} \\ \times \exp\left(-\sum_{k=1}^{K_m} \psi_m(\|\mathbf{\Sigma}_m^{-1/2}(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|)\right).\end{aligned} \tag{4.71}$$

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 (4.71) is also an inverse Wishart distribution [Murphy, 2007]. However, when $0 < \beta_m < 1$, (4.71) 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 correlations independently. Let us decompose the scale matrix for each subband $m \in \{1, \ldots, M\}$ as follows [Barnard et al., 2000]:

$$\mathbf{\Sigma}_m = C_{\beta_m,\delta_m}\mathrm{Diag}(\mathbf{s}_m)^{-1}\mathbf{R}_m\mathrm{Diag}(\mathbf{s}_m)^{-1} \tag{4.72}$$

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 [Marnissi et al., 2013]. 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:

$$\mathbf{s}_m = \gamma_m^{1/(2\beta_m)}\mathbf{n}_m \tag{4.73}$$

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.

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 all

$m \in \{1, \ldots, M\}$. Furthermore, let us suppose that $\mathbf{n}$ is known. We have then

$$\boldsymbol{\Theta} = \{\mathbf{R}, \gamma_1, \ldots, \gamma_M\}. \tag{4.74}$$

Prior and posterior distribution of 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 $\mathbf{R} = \boldsymbol{\Delta}\mathbf{C}\boldsymbol{\Delta}$ where $\boldsymbol{\Delta}$ is the diagonal matrix whose elements are given by: $\boldsymbol{\Delta}_{ii} = C_{ii}^{-1/2}$, for every $i \in \{1, \ldots, B\}$. Following [Barnard et al., 2000], we can show that the prior density of $\mathbf{R}$ reads:

$$\mathsf{p}(\mathbf{R}) \propto \det(\mathbf{R})^{-\frac{1}{2}(B+1+c)} \prod_{i=1}^{B} (\mathbf{R}^{-1}\mathbf{A})_{ii}^{-\frac{\nu}{2}}. \tag{4.75}$$

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 $\mathsf{p}(\rho_{ij}) = (1-\rho_{ij}^2)^{\frac{c-B-1}{2}}$ for all $i, j \in \{1, \ldots, B\}$ 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 either high correlation values around the extremity of the intervals (or near zero values), respectively [Barnard et al., 2000]. Thus, the posterior distribution of $\mathbf{R}$ is given by

$$\begin{aligned}\mathsf{p}(\mathbf{R}|\mathbf{x}, \gamma_1, \ldots, \gamma_M) \propto \det(\mathbf{R})^{-\frac{1}{2}(B+1+c+Q)} \exp\left(-\Psi(\mathbf{x})\right) \\ \times \prod_{i=1}^{B} (\mathbf{R}^{-1}\mathbf{A})_{ii}^{-\frac{c}{2}}.\end{aligned} \tag{4.76}$$

In this work, we propose to sample from (4.76) at each iteration t 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 $b_{\gamma_m} > 0$ [Fink, 1997]. Then, the posterior distribution of γ_m is given by:

$$\begin{aligned}&\mathsf{p}(\gamma_m|\mathbf{x}, \mathbf{R}) \propto \gamma_m^{a_{\gamma_m} + \frac{K_m B}{2\beta_m} - 1} \exp\left(-b_{\gamma_m}\gamma_m\right) \\ &\times \exp\left(-\frac{1}{2}\sum_{k=1}^{K_m}\left(\gamma_m^{\frac{1}{\beta_m}} C_{\beta_m,\delta_m}^{-1} \|\mathbf{R}^{-\frac{1}{2}}\mathrm{Diag}(\mathbf{n})(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|^2 + \delta_m\right)^{\beta_m}\right).\end{aligned} \tag{4.77}$$

Remark that if $\delta_m = 0$, (4.77) reduces to a Gamma distribution of parameters:

$$\tilde{a}_{\gamma_m} = a_{\gamma_m} + \frac{K_m B}{2\beta_m}, \tag{4.78}$$

$$\tilde{a}_{\gamma_m} = b_{\gamma_m} + C_{\beta_m,\delta_m}^{-\beta_m} \frac{1}{2} \sum_{k}^{K_m} \|\mathbf{R}^{-\frac{1}{2}} \mathbf{N}(\mathbf{P}_{m,k}\mathbf{x} - \mathbf{a}_m)\|^{2\beta_m}. \tag{4.79}$$

When $\delta_m > 0$, sampling from (4.77) can be performed using an independent Metropolis Hastings algorithm with a Gamma proposal of parameters (4.78) and (4.79).

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 in a given channel b and a subband m are approximately related to those of the degraded image through:

$$\text{var}([\mathbf{F}_b \mathbf{z}_b]_m) = \alpha_m \text{var}([\mathbf{x}_b]_m) + \sigma^2 \tag{4.80}$$

where $[.]_m$ denotes the wavelet coefficients belonging to the subband m and α_m is a positive constant that depend on m (the scale and the orientation corresponding to m) and on the blur matrix. Expression (4.80) is derived from the considered observation model (4.63) 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: For 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 the blur 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, estimators of $\text{var}([\mathbf{x}_b]_m)$ can be reliably computed from α_m and $\text{var}([\mathbf{F}_b \mathbf{z}_b]_m)$ using (4.80). In this work, we propose to use this method to compute estimators of the variances in subbands corresponding to the highest levels of decomposition and then 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 follow a power law with respect to the scale of the subband which can be expressed as follows [Flandrin, 1992]:

$$\log \text{var}([\mathbf{x}_b]_m) = \varrho_l j + \varpi_l \tag{4.81}$$

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 (4.80), 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 using (4.81).

We can then deduce from these variances an empirical estimator of $\mathbf{n}$. Moreover, we can set the parameters of the prior distributions of $\gamma_1, \ldots, \gamma_M$.

3.4 Experimental results

In these experiments, we consider the Hydice hyperspectral[2] 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 $B = 6$ 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 depth 3, and 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 $\delta_m = 0.0001$. Moreover, we set $\beta_m = 0.2$ for the subbands corresponding to the lower level of decomposition, $\beta_m = 0.4$ for the second level of decomposition and $\beta_m = 0.5$ for the third level of decomposition.

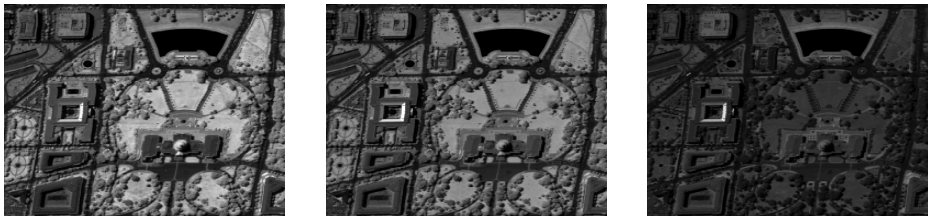
We run the Gibbs sampling algorithm 6 for a sufficient number of iterations to reach stability. The obtained samples of the wavelet coefficients after the burn-in period are then used to compute the empirical MMSE estimator for the original image. Table 4.1 reports the results obtained for the different components in terms of SNR, PSNR, BSNR and SSIM. It can be noted that the MMSE estimator yields good numerical results. This can also be observed in Figure 4.2 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 instead of 3MH algorithm. We also compare the speed of our proposed approaches with standard RW, MALA and 3MH algorithms without use of auxiliary variables. Figures 4.3 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 different algorithms. The results associated with the proposed algorithms appear in

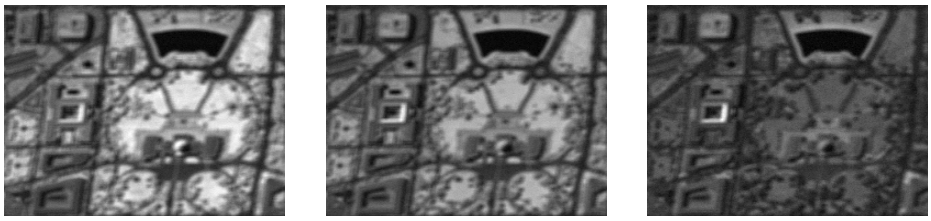
[2] https://engineering.purdue.edu/ biehl/MultiSpec/hyperspectral.html

Table 4.1: *Restoration results.*

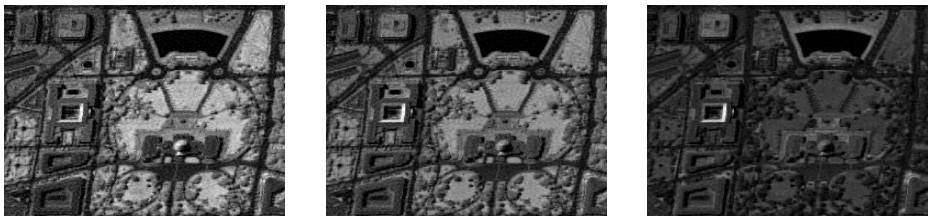
		$b=1$	$b=2$	$b=3$	$b=4$	$b=5$	$b=6$	Average
Initial	BSNR	24.27	30.28	31.73	28.92	26.93	22.97	27.52
	PSNR	25.47	21.18	19.79	22.36	23.01	26.93	23.12
	SNR	11.65	13.23	13.32	13.06	11.81	11.77	12.47
	SSIM	0.6203	0.5697	0.5692	0.5844	0.5558	0.6256	0.5875
MMSE	BSNR	32.04	38.33	39.21	38.33	35.15	34.28	36.22
	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 4.2: *From top to bottom: Original images-Degraded images-Restored images.*

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 smaller allowing large moves to be accepted with high probability values. Note that 3MH 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 4.2. It can be noted that all proposed algorithms provide similar estimation results.

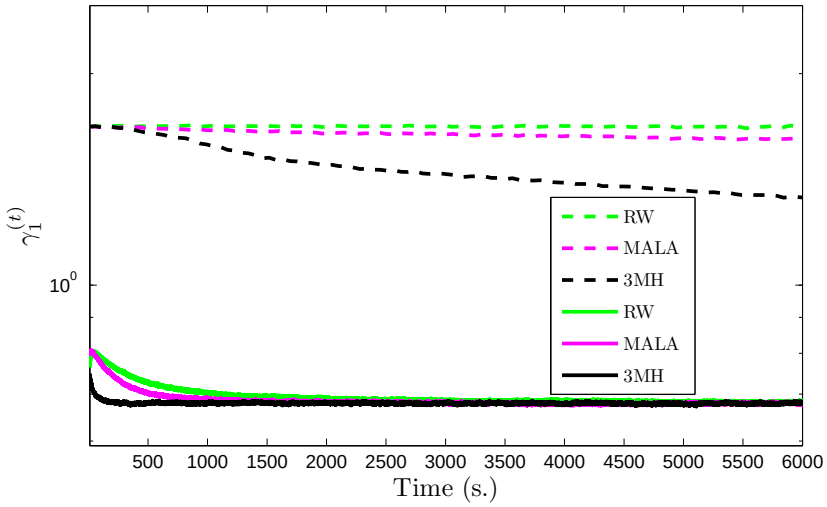


Figure 4.3: *Trace plot of the scale parameter in subband $m = 1$ as time (horizontal subband in the first level of decomposition).*

Following [Atchadé, 2006], we also compare the different proposed methods in terms of mixing properties based on the mean square jump in stationarity (MSJ) defined in (2.75). More precisely, MSJ is estimated as in (3.49), using $P = 20.000$ samples generated after the burn-in period. In Table 4.3, we show estimates of the mean square jump per second in stationarity which is defined as the ratio of the mean square jump and the computational time per iteration. We also compare the statistical efficiency of the different samplers with respect to RW defined as the mean square jump per second of each sampler over the mean square jump per second of RW. It can be noted that 3MH algorithm achieves the best results in terms of mixing properties whereas RW yields the poorest ones. A huge improvement in the efficiency of Langevin based algorithms is particularly observed when introducing the curvature matrix (4.70).

Table 4.2: *Mean and variance estimates of hyperparameters.*

		RW	MALA	3MH
$\hat{\gamma_1}$	Mean	0.67	0.67	0.67
(γ_1=0.71)	Std.	(1.63 e-3)	(1.29 e-3)	(1.48 e-3)
$\hat{\gamma_2}$	Mean	0.83	0.83	0.83
(γ_2=0.99)	Std.	(1.92 e-3)	(2.39 e-3)	(2.01 e-3)
$\hat{\gamma_3}$	Mean	0.62	0.61	0.61
(γ_3=0.72)	Std.	(1.33 e-3)	(1.23 e-3)	(1.28 e-3)
$\hat{\gamma_4}$	Mean	0.24	0.24	0.24
(γ_4=0.0.24)	Std.	(1.30 e-3)	(1.39 e-3)	(1.34 e-3)
$\hat{\gamma_5}$	Mean	0.37	0.37	0.37
(γ_5=0.40)	Std.	(2.10 e-3)	(2.42 e-3)	(2.35 e-3)
$\hat{\gamma_6}$	Mean	0.21	0.21	0.21
(γ_6=0.22)	Std.	(1.19 e-3)	(1.25 e-3)	(1.20 e-3)
$\hat{\gamma_7}$	Mean	0.08	0.08	0.08
(γ_7=0.0.07)	Std.	(0.91 e-3)	(1.08 e-3)	(1 e-3)
$\hat{\gamma_8}$	Mean	0.13	0.13	0.13
(γ_8=0.13)	Std.	(1.60 e-3)	(1.64 e-3)	(1.62 e-3)
$\hat{\gamma_9}$	Mean	0.07	0.07	0.07
(γ_9=0.07)	Std.	(0.83 e-3)	(1 e-3)	(0.88 e-3)
$\hat{\gamma_1 0}$	Mean	7.80 e-4	7.87 e-4	7.86 e-4
(γ_{10}=7.44 e-4)	Std.	(1.34 e-5)	(2.12 e-5)	(1.66 e-5)
det($\hat{\mathbf{R}}$)	Mean	1.89 e-8	2.10 e-8	2.08 e-8
det($\mathbf{R}$)= 5.79 e-8	Std.	(9.96 e-10)	(2.24 e-9)	(1.72 e-9)

Table 4.3: *Results for the different proposed algorithms. First row: Estimates of the mean square jump in stationarity. Second row: Time per iteration. Third row: Estimates of the mean square jump per second in stationarity. Fourth row: Relative efficiency compared to RW.*

	RW	MALA	3MH
MSJ	0.74	1.14	2.19
T(s.)	0.16	0.18	0.22
MSJ/T	4.65	6.01	9.85
Efficiency	1	1.29	2.11

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.

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

4.1 Problem formulation

In this second experiment, we consider the observation problem defined in (4.22) 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 $i \in \{1, \ldots, N\}$, $w_i \sim \mathcal{N}(0, \sigma_i^2)$ such that

$$\sigma_i \sim (1-\beta)\delta_{\kappa_1} + \beta\delta_{\kappa_2} \tag{4.82}$$

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 (4.82) can approximate for example mixed impulse Gaussian noise arising in radar, acoustic, and mobile radio applications [Velayudhan and Paul, 2016; Chang et al., 2016]. 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.,

$$\mathsf{p}(\mathbf{x}|\gamma) \propto \gamma^{-Q/2} \exp\left(-\frac{\gamma}{2}\|\mathbf{L}\mathbf{x}\|^2\right). \tag{4.83}$$

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 $\mathbf{L} = \delta\mathbf{I}_Q - \nabla$ where ∇ 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 posterior 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 4.4.

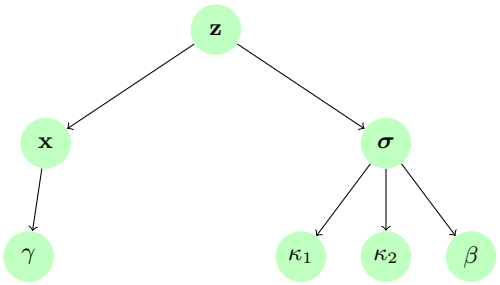


Figure 4.4: *Hierarchical model for image deblurring under two term mixed Gaussian noise.*

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:

$$\mathbf{G} = \mathbf{H}^\top \mathbf{D}\mathbf{H} + \gamma \mathbf{L}^\top \mathbf{L}, \tag{4.84}$$

$$\mathbf{m} = \mathbf{G}^{-1}\mathbf{H}^\top \mathbf{D}\mathbf{y}, \tag{4.85}$$

where $\mathbf{D}$ is the diagonal matrix defined by $D_{ii} = \sigma_i^{-2}$.

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

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

$$\eta_i = \frac{\beta}{1-\beta} \exp\left(-\frac{1}{2}\left(\kappa_2^{-2} - \kappa_1^{-2}\right) \left([\mathbf{H}\mathbf{x}]_i - z_i\right)^2 \right) \frac{\kappa_1}{\kappa_2}, \tag{4.86}$$

- $\beta | \mathbf{x}, \mathbf{z}, \boldsymbol{\sigma}, \kappa_1^2, \kappa_2^2 \sim \mathcal{B}\left(n_2 + 1, n_1 + 1\right)$, where $\mathcal{B}$ is the Beta distribution and n_1 and n_2 are the cardinals of the sets $\{i \in \{1, \ldots, N\}, \mid \sigma_i = \kappa_1, \}$ and $\{i \in \{1, \ldots, N\}, \mid \sigma_i = \kappa_2, \}$ respectively so that $n_1 + n_2 = N$,

- $\kappa_1^2 | \mathbf{x}, \boldsymbol{\sigma}, \beta, \mathbf{z} \sim \mathcal{IG}\left(a_1 + \frac{n_1}{2}, b_1 + \sum_{i|\sigma_i = \kappa_1} \dfrac{\left([\mathbf{H}\mathbf{x}]_i - z_i\right)^2}{2} \right)$,

- $\kappa_2^2 | \mathbf{x}, \boldsymbol{\sigma}, \beta, \mathbf{z} \sim \mathcal{IG}\left(a_2 + \frac{n_2}{2}, b_2 + \sum_{i|\sigma_i = \kappa_2} \dfrac{\left([\mathbf{H}\mathbf{x}]_i - z_i\right)^2}{2} \right)$,

- $\gamma | \mathbf{x} \sim \mathcal{G}\left(\dfrac{Q}{2} + a_\gamma, \dfrac{1}{2}\|\mathbf{L}\mathbf{x}\|^2 + b_\gamma \right)$.

4.2 Sampling from the posterior distribution of x

In the Gibbs algorithm, we need to draw samples from the multivariate Gaussian distribution of parameters (4.84) and (4.85) 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 proposed method in Section 2.3 to sample from this multivariate Gaussian distribution. More specifically, we will test 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 method: We introduce the variable $\mathbf{v}$ whose conditional distribution given the set of main parameters of the problem, is the Gaussian distribution of mean $\left(\frac{1}{\mu}\mathbf{I}_N - \mathbf{D}\right)\mathbf{H}\mathbf{x}$ and covariance matrix $\left(\frac{1}{\mu}\mathbf{I}_N - \mathbf{D}\right)$ where $\mu = \epsilon\|\mathbf{D}\|^{-1}$ and $0 < \epsilon < 1$. In practice, we set $\epsilon = 0.99$.

It follows that the new conditional distribution of the target signal is

$$\mathbf{x}|\boldsymbol{\sigma}, \beta, \kappa_1^2, \kappa_2^2, \gamma, \mathbf{v}, \mathbf{z} \sim \mathcal{N}(\tilde{\mathbf{m}}, \tilde{\mathbf{G}}^{-1}) \tag{4.87}$$

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

$$\tilde{\mathbf{G}} = \frac{1}{\mu}\mathbf{H}^\top\mathbf{H} + \gamma\mathbf{L}^\top\mathbf{L}, \tag{4.88}$$

$$\tilde{\mathbf{m}} = \tilde{\mathbf{G}}^{-1}\mathbf{H}^\top\left(\mathbf{H}^\top\mathbf{D}\mathbf{z} + \mathbf{v}\right). \tag{4.89}$$

It is worth noting that the auxiliary variable $\mathbf{v}$ depends on $\mathbf{x}$ and also on $\boldsymbol{\sigma}$ through μ and $\mathbf{D}$ and does not depend on β, κ_1, κ_2, γ when conditioned to $\mathbf{x}$, $\boldsymbol{\sigma}$ and $\mathbf{z}$. In this work, 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, the proposed algorithm goes through the following steps in an ordered manner:

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

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

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 $\mu^{(t+1)} = \epsilon \min\left(\sigma_i^{(t+1)}\right)_{1\leqslant i\leqslant N}^{-2}$ and sample $\mathbf{v}^{(t+1)}$ from $\mathcal{P}_{\mathbf{v}|\mathbf{x}^{(t)},\boldsymbol{\sigma}^{(t+1)},\mathbf{z}}$.

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

In the following, this algorithm will be denoted as AuxV1.

Second method: We introduce two independent auxiliary variables $\mathbf{v}_1$ and $\mathbf{v}_2$ in $\mathbb{R}^Q$ following Gaussian distributions of means $\boldsymbol{\Gamma}_1\mathbf{x}$ and $\boldsymbol{\Gamma}_2\mathbf{x}$ and covariance matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ respectively where

$$\boldsymbol{\Gamma}_1 = \frac{1}{\mu_1} - \mathbf{H}^\top\mathbf{D}\mathbf{H} \tag{4.90}$$

and

$$\boldsymbol{\Gamma}_2 = \frac{1}{\mu_2} - \mathbf{L}^\top\mathbf{L}. \tag{4.91}$$

In practice, we set $\mu_1 = \epsilon\|\mathbf{H}\|^{-2}\|\mathbf{D}\|^{-1}$ and $\mu_2 = \epsilon\|\mathbf{L}\mathbf{L}\|^{-2}$ 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 $\tilde{\mathbf{m}}$ and precision matrix $\tilde{\mathbf{G}}$ defined by:

$$\tilde{\mathbf{G}} = \left(\frac{\gamma}{\mu_2} + \frac{1}{\mu_1}\right)\mathbf{I}_Q \tag{4.92}$$

and

$$\tilde{\mathbf{m}} = \mu_1\mu_2\left(\gamma\mu_1 + \mu_2\right)^{-1}\mathbf{H}^\top\left(\mathbf{D}\mathbf{y} + \mathbf{v}_1 + \sqrt{\gamma}\mathbf{v}_2\right). \tag{4.93}$$

The auxiliary variable $\mathbf{v}_1$ depends implicitly of $\boldsymbol{\sigma}$ 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 of the proposed PCGS algorithm is

composed of the following arranged sampling steps.

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

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

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 $\mathcal{P}_{\mathbf{v}_2|\mathbf{x}^{(t)},\mathbf{z}}$.

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

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

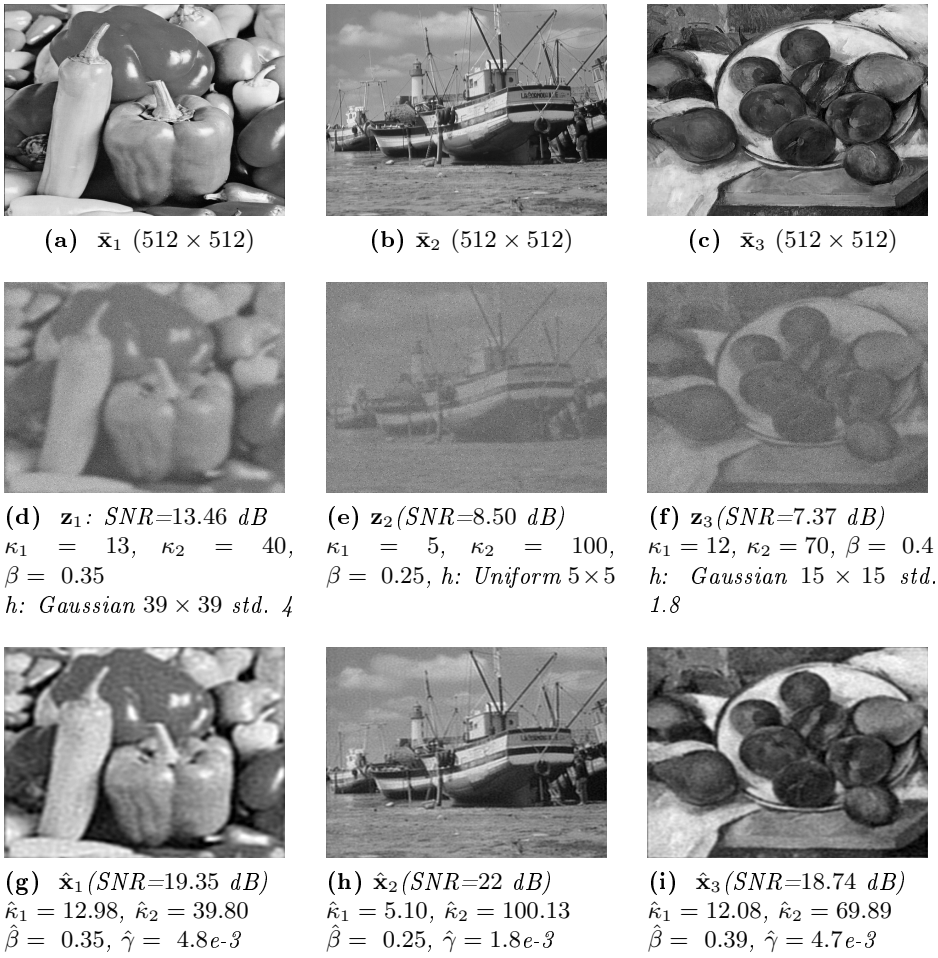
In the following, this algorithm will be denoted by AuxV2.

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 2.4.

4.3 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 4.5 as well as estimates of hyperparameters using AuxV1.

We consider the image $\bar{\mathbf{x}}_1$ and we propose to compare the two variants of our proposed method with the Reversible Jump Perturbation Optimization (RJPO) algorithm [Gilavert et al., 2015]. For this method, we use the conjugate gradient algorithm as linear solver at each iteration whose maximal number of iterations and tolerance are adjusted to corresponds to an acceptance probability of around 0.9. We use the same initialization for all compared algorithms. Figures 4.6-4.9 display samples of hyperparameters



(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)* $\hat{\kappa}_1 = 12.98$, $\hat{\kappa}_2 = 39.80$ $\hat{\beta} = 0.35$, $\hat{\gamma} = 4.8e\text{-}3$

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

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

Figure 4.5: *Visual results. From top to bottom: Original images. Degraded images. Restored images.*

as a function of iteration or time. Table 4.4 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 4.6-4.9 show that all the tested algorithms reach the same stationary distribution. In particular, it can be noted from Table 4.4 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 behavior, AuxV2 needs more iterations to reach stability. However, the proposed algorithms need less time to converge compared to the RJPO algorithm since the cost of each iteration is highly reduced.

Table 4.4: *Mean and variance estimates.*

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

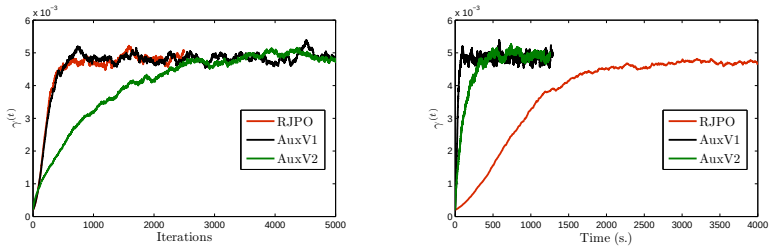


Figure 4.6: *Chains of γ as iteration/time.*

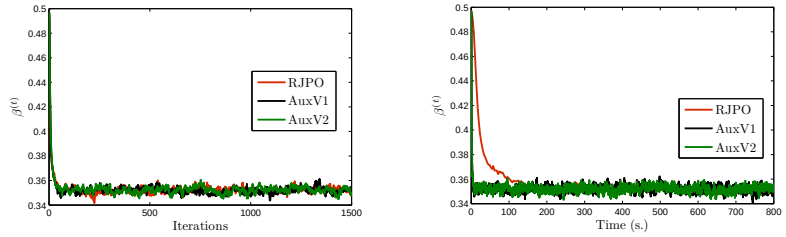


Figure 4.7: *Chains of β as iteration/time.*

Similarly to Section 3, we also report comparisons in terms of mixing properties in convergence. Table 4.5 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.

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

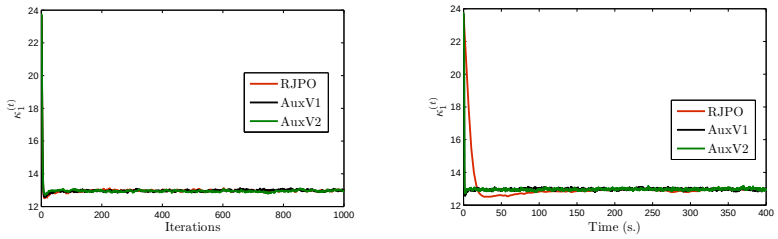


Figure 4.8: *Chains of κ_1 as iteration/time.*

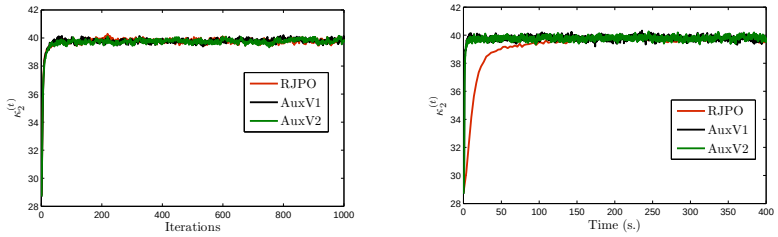


Figure 4.9: *Chains of κ_2 as iteration/time.*

Table 4.5: *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

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

Conclusion

In this chapter, we have proposed a method that addresses sampling from 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. In the first experiment, we have shown the good performance of this new approach in terms of convergence speed and mixing properties when applied to the recovery of multispectral images from their blurred version. 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. In the second set of experiments, we have applied the proposed method to the recovery of signals corrupted with mixed Gaussian noise. When compared to state-of-the-art methods 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.

- Chapter 5 -

A variational Bayesian approach for restoring data corrupted with non-Gaussian noise

Noise arising in real signal processing problems may originate from various sources. In particular, the signal of interest may suffer from noise with complex characteristics where signal-independent additive Gaussian hypothesis fails to properly describe it. For example, noise may be signal-dependent [Moser, 2012], multiplicative [Aubert and Aujol, 2008] and with non-Gaussian characteristics [Salmon et al., 2014; Altmann et al., 2016]. However, most of the existing denoising methods only consider the noise as independent Gaussian, mainly because of the difficulties raised in handling other noise sources than the Gaussian one.

In this work, we focus on signal recovery beyond the standard additive independent Gaussian noise assumption. We propose to resort to VBA methods to restore signals degraded by an arbitrary linear operator and corrupted with non-Gaussian noise. One of the main advantages of our proposed method is that it allows us to jointly estimate the original signal and the required regularization parameter from the observed data by providing good approximations of the MMSE estimators for the problem of interest.

This chapter is organized as follows. In Section 1, we formulate the considered signal recovery problem in the Bayesian framework and we present a short overview on the state-of-the-art methods. In Section 2, we present our proposed estimation method based on VBA. Finally, in Section 3, we provide simulation results together with comparisons with state-of-the-art methods in terms of image restoration performance and computation time.

§ 1 Problem statement

1.1 Model

In this chapter, we consider a wide range of applications where the degradation model can be formulated as an inverse possibly ill-posed problem as in (2.1). We further assume that the coefficients of the vector of observations

$\mathbf{z} = (z_i)_{1 \leqslant i \leqslant N} \in \mathbb{R}^N$ are independents, then the observation model in (2.1) reads:

$$(\forall i \in \{1, \ldots, N\}), \quad z_i = \mathcal{D}([\mathbf{H}\overline{\mathbf{x}}]_i) \tag{5.1}$$

where $[\mathbf{H}\overline{\mathbf{x}}]_i$ denotes the i-th component of $\mathbf{H}\overline{\mathbf{x}}$ and $\mathcal{D}$ is the noise model that may depend on the signal of interest $\mathbf{x}$. The objective is to find an estimator $\hat{\mathbf{x}}$ of $\overline{\mathbf{x}}$ from $\mathbf{H}$ and $\mathbf{z}$. The neg-log-likelihood ϕ of the observations reads

$$(\forall \mathbf{x} \in \mathbb{R}^Q) \quad \Phi(\mathbf{x}) = -\log \mathsf{p}(\mathbf{z}|\mathbf{x}) = \sum_{i=1}^{N} \phi_i([\mathbf{H}\mathbf{x}]_i ; z_i). \tag{5.2}$$

Depending on the noise statistical model $\mathcal{D}$, ϕ_i may take various forms [Janesick, 2007; Mäkitalo and Foi, 2012; Cai et al., 2010; Yan, 2013]. In particular, it reduces to a least squares function for additive Gaussian noise.

1.2 Related work

Most of the existing strategies in the literature propose to tackle the problem of restoration for signals corrupted with non-Gaussian noise using minimization approaches. As pointed out in Chapter 2, the cost function is the sum of two terms: the neg-log-likelihood related to the noise statistics and the regularization term that incorporates prior information about the target signal so as to ensure the stability of the solution [Demoment, 1989]. For example, in [Repetti et al., 2012], a method is proposed to restore signals degraded by a linear operator and corrupted with an additive Gaussian noise having a signal-dependent variance. An early work in [Snyder et al., 1993] and more recent developments in [Lantéri and Theys, 2005; Benvenuto et al., 2008; Jezierska et al., 2012; Li et al., 2015; Roberts and Kingsbury, 2014; Chouzenoux et al., 2015; Bajić et al., 2016] have proposed to restore signals corrupted with mixed PG noise using different approximations of the PG data fidelity term. In all these approaches, the regularization parameter allows a tradeoff to be performed between fidelity to the observations and the prior information. Too small values of this parameter may lead to noisy estimates while too large values yield oversmoothed solutions. Consequently, the problem of setting a proper value of the regularization parameter should be addressed carefully and may depend on both the properties of the observations and the statistics of the target signal. When ground truth is available, one can choose the value of the regularization parameter that gives the minimal residual error evaluated through some suitable metric. However, in real-world applications where no ground truth is available, the problem of selecting the regularization parameter remains an open issue especially in situations where the images are acquired under poor conditions i.e., when the noise level is very high. Among existing approaches dealing with regularization parameter estimation, the works in [Ramani et al., 2008; Eldar, 2009; Pesquet et al., 2009; Deledalle et al., 2014; Giryes et al., 2011; Almeida

and Figueiredo, 2013; Hansen et al., 2006] have to be mentioned. However, most of the presented methods were developed under the assumption of a Gaussian noise and their extension to the context of non-Gaussian noise is tricky. One can however mention the works in [Luisier et al., 2011; Mäkitalo and Foi, 2012] proposing efficient estimators in the context of denoising i.e., problems that do not involve linear degradation. Other approaches can be found in [Bertero et al., 2010; Zanni et al., 2015] proposing efficient estimates in the specific case of a Poisson likelihood. To address the shortcomings of these methods, one may resort to the Bayesian framework where regularization is applied by assigning a prior distribution to the data $\mathbf{x}$ to be recovered. In particular, Bayesian estimation methods based on MCMC sampling algorithms have been recently extended to inverse problems involving non-Gaussian noise [Ying et al., 2012; Altmann et al., 2015; Murphy and Godsill, 2011; Chaâri et al., 2013]. However, despite good estimation performance that has been obtained, such methods remain computationally expensive for large scale problems. In this chapter, we propose to tackle this problem by resorting to VBA approaches.

1.3 Bayesian formulation

We propose to adopt the following flexible expression of the prior density of $\mathbf{x}$:

$$\mathsf{p}(\mathbf{x}|\gamma) = \tau\gamma^{\frac{Q}{2\kappa}} \exp\Big(-\gamma \sum_{j=1}^{J} \|\mathbf{D}_j\mathbf{x}\|^{2\kappa}\Big) \tag{5.3}$$

where κ is a constant in $(0,1]$, $\|\cdot\|$ denotes the ℓ_2-norm and $(\mathbf{D}_j)_{1\leqslant j\leqslant J} \in (\mathbb{R}^{S\times Q})^J$ where $\mathbf{D} = [\mathbf{D}_1^\top, \ldots, \mathbf{D}_J^\top]^\top$ is a linear operator. For instance, $\mathbf{D}$ may be a matrix computing the horizontal and vertical discrete difference between neighboring pixels so that $J = Q$ and $S = 2$. A sparsity prior in an analysis frame can also be modeled by setting $S = 1$ and $\mathbf{D}$ equals to a frame operator with decomposition size $J \geqslant Q$ [Pustelnik et al., 2016]. Other examples will be given in Section 3. Note that the constant $\gamma \in (0, +\infty)$ can be viewed as a regularization parameter that plays a prominent role in the restoration process and $\tau \in (0, +\infty)$ is a constant independent of γ. The form of the partition function for such a prior distribution, i.e. the normalizing factor $\tau\gamma^{Q/2k}$, follows from the fact that the associated potential is 2κ-homogeneous [Pereyra et al., 2015].

Generally, κ is the shape parameter of the prior law which determines the type of prior information introduced by the user and hence can be fixed according to the prior knowledge while γ expresses the compromise between data fidelity and prior information. In this work, we aim at estimating parameter γ together with $\mathbf{x}$. To this end, we choose a Gamma prior for γ i.e., $\mathsf{p}(\gamma) \propto \gamma^{\alpha-1}\exp(-\beta\gamma)$ where α and β are positive constants (set in practice to small values to ensure a weakly informative prior).

Using the Bayes' rule, we can obtain the posterior distribution of the set of unknown variables $\boldsymbol{\Theta} = (\mathbf{x}, \gamma)$ given the vector of observations $\mathbf{z}$:

$$\mathsf{p}(\boldsymbol{\Theta}|\mathbf{z}) \propto \mathsf{p}(\mathbf{z}|\mathbf{x})\mathsf{p}(\mathbf{x}|\gamma)\mathsf{p}(\gamma). \tag{5.4}$$

However, this distribution has an intricate form. In particular, its normalization constant does not have a closed form expression. To cope with this problem, we resort to the variational Bayesian framework. The rationale of this work is to find a simple approximation to the true posterior distortion, leading to a tractable computation of the posterior mean estimate.

§ 2 Proposed approach

In this work, we assume the following separable form for q:

$$\mathsf{q}(\boldsymbol{\Theta}) = \mathsf{q}_{\mathbf{X}}(\mathbf{x})\mathsf{q}_{\Gamma}(\gamma). \tag{5.5}$$

Unfortunately, by using directly (2.87), we cannot obtain an explicit expression of $\mathsf{q}_{\mathbf{X}}(\mathbf{x})$ due to the intricate form of both the prior distribution and the likelihood when the statistics of the noise are no longer Gaussian. In this work, we propose to use deterministic methods to construct quadratic upper bounds for the negative logarithms of both the likelihood and the prior density [Seeger and Bouchard, 2012]. This allows us to derive an upper bound of the desired cost function in (2.81) as will be described in the following.

2.1 Construction of the majorizing approximation

2.1.1 Likelihood

One popular approach in signal recovery is the half-quadratic formulation [Geman and Yang, 1995]. Under some mild assumptions and by introducing some auxiliary variables, a complicated criterion can be written as the infimum of a surrogate half-quadratic function i.e., the latter is quadratic with respect to the original variables and the auxiliary variables appear decoupled. This half-quadratic criterion can be then efficiently minimized using classical optimization algorithms. Furthermore, we have recently extended this technique to sampling algorithms [Marnissi et al., 2016a]. The initial intractable posterior distribution to sample from is replaced by the conditional distribution of the target signal given the auxiliary variables. The obtained distribution has been shown to be much simpler to explore by using standard sampling algorithms. This formulation has been widely used in energy-minimization approaches [Idier, 2001; Nikolova and Ng, 2005; Champagnat and Idier, 2004] where the initial optimization problem is replaced by the minimization of the constructed surrogate function. In this work, we propose to use half-quadratic approaches to construct an upper bound for the objective function in (2.81).

Table 5.1: *Examples of differentiable functions satisfying Assumption 2.1. The Anscombe transform provides a differentiable approximation of the exact Poisson data fidelity term, while the three last functions can be employed to approximate the exact mixed Poisson-Gaussian log-likelihood. Note that alternative expressions for the Anscombe-based approaches can be found in [Mäkitalo and Foi, 2011, 2013]. ϕ_i' denotes the first derivative of function ϕ_i and $\beta_i(z_i)$ is the Lipschitz constant of ϕ_i' (for functions in lines 3-6, we assume that ϕ_i is replaced on $\mathbb{R}_-$ by its quadratic extension (5.7).) The expression for the Lipschitz constant of the gradient of the weighted least squares likelihood was established in [Repetti, 2015, Chap. IV].*

Name	$\phi_i(v; z_i)$	$\phi_i'(v; z_i)$	$\beta_i(z_i)$	Domain of validity	Noise model
Gaussian	$\frac{1}{2\sigma^2}(v - z_i)^2$	$\frac{1}{\sigma^2}(v - z_i)$	$\frac{1}{\sigma^2}$	$z_i \in \mathbb{R}$, $\sigma > 0$	Gaussian
Cauchy	$\ln\left(1 + \frac{(v - z_i)^2}{\sigma^2}\right)$	$\frac{2(v - z_i)}{\sigma^2 + (v - z_i)^2}$	$\frac{2}{\sigma^2}$	$z_i \in \mathbb{R}$, $\sigma > 0$	Cauchy
Anscombe transform	$2\left(\sqrt{z_i + \frac{3}{8}} - \sqrt{v + \frac{3}{8}}\right)^2$	$2 - \frac{2\sqrt{z_i + \frac{3}{8}}}{\sqrt{v + \frac{3}{8}}}$	$\left(\frac{3}{8}\right)^{-3/2}\sqrt{z_i + \frac{3}{8}}$	$z_i \geqslant -\frac{3}{8}$	Poisson
Generalized Anscombe transform	$2\left(\sqrt{z_i + \sigma^2 + \frac{3}{8}} - \sqrt{v + \sigma^2 + \frac{3}{8}}\right)^2$	$2 - \frac{2\sqrt{z_i + \frac{3}{8} + \sigma^2}}{\sqrt{v + \frac{3}{8} + \sigma^2}}$	$\left(\frac{3}{8} + \sigma^2\right)^{-3/2}\sqrt{z_i + \frac{3}{8} + \sigma^2}$	$z_i \geqslant -\frac{3}{8} - \sigma^2$	Poisson-Gaussian
Shifted Poisson	$(v + \sigma^2) - (z_i + \sigma^2)\ln(v + \sigma^2)$	$1 - \frac{z_i + \sigma^2}{v + \sigma^2}$	$\frac{z_i + \sigma^2}{\sigma^4}$	$z_i \geqslant -\sigma^2$, $\sigma > 0$	Poisson-Gaussian
Weighted least squares	$\frac{(z_i - v)^2}{2(\sigma^2 + v)} + \frac{1}{2}\ln(\sigma^2 + v)$	$\frac{1}{2} - \frac{(z_i + \sigma^2)^2}{2(v + \sigma^2)^2} + \frac{1}{2(\sigma^2 + v)}$	$\max\left\{\frac{(z_i + \sigma^2)^2}{\sigma^6} - \frac{1}{2\sigma^4}, \frac{1}{54(z_i + \sigma^2)^4}\right\}$	$z_i \in \mathbb{R}\backslash\{-\sigma^2\}$, $\sigma > 0$	Poisson-Gaussian

We assume that the likelihood satisfies the following property:

Assumption 2.1 For every $i \in \{1, \dots, N\}$, ϕ_i is differentiable on $\mathbb{R}$ and there exists $\mu_i(z_i) > 0$ such that the function defined by $v \mapsto \frac{v^2}{2} - \frac{\phi_i(v;z_i)}{\mu_i(z_i)}$ is convex on $\mathbb{R}$.

In particular, this assumption is satisfied when, for every $i \in \{1, \dots, N\}$, ϕ_i is $\beta_i(z_i)$-Lipschitz differentiable on $\mathbb{R}$, i.e.,

$$(\forall u \in \mathbb{R})\,(\forall v \in \mathbb{R}) \quad |\phi_i'(v;z_i) - \phi_i'(u;z_i)| \leqslant \beta_i(z_i)|v-u| \tag{5.6}$$

as soon as $\mu_i(z_i) \geqslant \beta_i(z_i)$.

Table 5.1 shows some examples of useful functions satisfying the desired property (up to an additive constant). Note that, since the functions in lines 3-6 of Table 5.1 are $\beta_i(z_i)$-Lipschitz differentiable only on $\mathbb{R}_+$, we propose to use on $\mathbb{R}_-$ a quadratic extension of them defined as follows:

$$(\forall v \in \mathbb{R}_-) \quad \phi_i(v;z_i) = \phi_i(0;z_i) + \phi_i'(0;z_i)v + \frac{1}{2}\beta_i(z_i)v^2 \tag{5.7}$$

so that the extended version of $\phi_i(.;z_i)$ is now differentiable on $\mathbb{R}$ with $\beta_i(z_i)$-Lipschitzian gradient.

For every $i \in \{1, \dots, N\}$ and $v \in \mathbb{R}$, let us define the following function:

$$\varsigma_i(v;z_i) = \sup_{t \in \mathbb{R}} \left(-\frac{1}{2}(v-t)^2 + \frac{\phi_i(t;z_i)}{\mu_i(z_i)} \right). \tag{5.8}$$

Then, the following property holds:

Proposition 2.1 *Upper bound for minus log likelihood*
For every $i \in \{1, \dots, N\}$,

$$(\forall v \in \mathbb{R}) \quad \phi_i(v;z_i) = \inf_{w_i \in \mathbb{R}} \; T_i(v, w_i; z_i) \tag{5.9}$$

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

$$T_i(v, w_i; z_i) = \mu_i(z_i) \left(\frac{1}{2}(v-w_i)^2 + \varsigma_i(w_i;z_i) \right). \tag{5.10}$$

Moreover, the unique minimizer of $w_i \mapsto T_i(v, w_i; z_i)$ reads

$$\widehat{w}_i(v) = v - \frac{1}{\mu_i(z_i)}\phi_i'(v;z_i). \tag{5.11}$$

Proof. See Appendix B. □

It follows from this result that

$$\left(\forall \mathbf{x} \in \mathbb{R}^Q\right) \quad \phi(\mathbf{x};\mathbf{z}) = \inf_{\mathbf{w}\in\mathbb{R}^N} T(\mathbf{x},\mathbf{w};\mathbf{z}) \tag{5.12}$$

where $T(\mathbf{x},\mathbf{w};\mathbf{z}) = \sum_{i=1}^{N} T_i([\mathbf{H}\mathbf{x}]_i, w_i; z_i)$.

Note that (5.9) shows that, for every $i \in \{1,\dots,N\}$, $\phi_i(\cdot;z_i)$ is a so-called Moreau envelope of the function $\mu_i(z_i)\varsigma_i(\cdot;z_i)$. A more direct proof of Proposition 2.1 can thus be derived from the properties of the proximity operator [Combettes and Pesquet, 2010] when the functions $(\phi_i)_{1\leqslant i\leqslant N}$ are convex. The proof we provide in the appendix however does not make such a restrictive assumption.

2.1.2 Prior

Similarly, we construct a surrogate function for the prior distribution. More precisely, we follow the same idea as in [Chen et al., 2014] and we use the following convexity inequality to derive a majorant for the ℓ_κ-norm with $\kappa \in (0,1]$:

$$(\forall \nu > 0)(\forall \upsilon \geqslant 0) \quad \upsilon^\kappa \leqslant (1-\kappa)\nu^\kappa + \kappa\nu^{\kappa-1}\upsilon.$$

Hence, we obtain the following majorant function for the negative logarithm of the prior distribution:

$$\gamma \sum_{j=1}^{J} \|\mathbf{D}_j\mathbf{x}\|^{2\kappa} \leqslant \gamma \sum_{j=1}^{J} \frac{\kappa\|\mathbf{D}_j\mathbf{x}\|^2 + (1-\kappa)\lambda_j}{\lambda_j^{1-\kappa}}. \tag{5.13}$$

where $(\lambda_j)_{1\leqslant j\leqslant J}$ are positive variables. In the following, we will denote by $Q(\mathbf{x},\boldsymbol{\lambda};\gamma) = \sum_{j=1}^{J} Q_j(\mathbf{D}_j\mathbf{x},\lambda_j;\gamma)$, the function in the right-hand side of the above inequality where, for every $j\in\{1,\dots,J\}$,

$$Q_j(\mathbf{D}_j\mathbf{x},\lambda_j;\gamma) = \gamma\frac{\kappa\|\mathbf{D}_j\mathbf{x}\|^2 + (1-\kappa)\lambda_j}{\lambda_j^{1-\kappa}}. \tag{5.14}$$

2.1.3 Proposed majorant

Thus, we can derive the following lower bound for the posterior distribution:

$$\mathsf{p}(\boldsymbol{\Theta} \mid \mathbf{z}) \geqslant L(\boldsymbol{\Theta}|\mathbf{z};\mathbf{w},\boldsymbol{\lambda}), \tag{5.15}$$

where function L is defined as

$$L(\boldsymbol{\Theta}|\mathbf{z};\mathbf{w},\boldsymbol{\lambda}) = C(\mathbf{z})\exp\left[-T(\mathbf{x},\mathbf{w};\mathbf{z}) - Q(\mathbf{x},\boldsymbol{\lambda};\gamma)\right]\mathsf{p}(\gamma)$$

with $C(\mathbf{z}) = \mathsf{p}(\mathbf{z})^{-1}(2\pi)^{-N/2}\tau\gamma^{\frac{Q}{2\kappa}}$. The minorization of the distribution leads to an upper bound for the $\mathcal{KL}$ divergence:

$$\mathcal{KL}(\mathsf{q}(\boldsymbol{\Theta})\|\mathsf{p}(\boldsymbol{\Theta} \mid \mathbf{z})) \leqslant \mathcal{KL}(\mathsf{q}(\boldsymbol{\Theta})\|L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda})). \tag{5.16}$$

Note that, although the constructed lower bound in (5.15) is tangent to the posterior distribution i.e.

$$\mathsf{p}(\boldsymbol{\Theta} \mid \mathbf{z}) = \sup_{\mathbf{w}\in\mathbb{R}^N, \boldsymbol{\lambda}\in\mathbb{R}^J} L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}),$$

the tangency property may not be generally satisfied in (5.16). Thus, the tightness of the constructed majorant of the $\mathcal{KL}$ divergence may have a significant impact on the accuracy of the method. By minimizing the constructed bound (5.16) with respect to $\mathbf{w}$ and $\boldsymbol{\lambda}$, we make this bound as tight as possible. Note that, for every $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, J\}$, $\lambda_j \mapsto \mathcal{KL}(\mathsf{q}(\boldsymbol{\Theta})\|L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}))$ and $w_i \mapsto \mathcal{KL}(\mathsf{q}(\boldsymbol{\Theta})\|L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}))$ can be minimized separately. Hence, Problem (2.81) can be solved by the following four-step alternating optimization scheme:

- Minimizing the upper bound in (5.16) w.r.t. $\mathsf{q}_{\mathbf{X}}(\mathbf{x})$;
- Updating the auxiliary variables w_i in order to minimize $\mathcal{KL}(\mathsf{q}(\boldsymbol{\Theta})\|L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}))$, for every $i \in \{1, \ldots, N\}$;
- Updating the auxiliary variable λ_j in order to minimize $\mathcal{KL}(\mathsf{q}(\boldsymbol{\Theta})\|L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}))$, for every $j \in \{1, \ldots, J\}$;
- Mimimizing the upper bound in (5.16) w.r.t. $\mathsf{q}_{\Gamma}(\gamma)$.

The main benefit of this majorization strategy is to guarantee that the optimal approximate posterior distribution for $\mathbf{x}$ belongs to the Gaussian family and the optimal approximate posterior distribution for γ belongs to the Gamma one, i.e.

$$\mathsf{q}_{\mathbf{X}}(\mathbf{x}) \equiv \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}), \quad \mathsf{q}_{\Gamma}(\gamma) \equiv \mathcal{G}(a, b).$$

Therefore, the distribution updates can be performed by updating their parameters, namely $\mathbf{m}$, $\boldsymbol{\Sigma}$, a, and b.

2.2 Iterative algorithm

Subsequently, at a given iteration k of the proposed algorithm, the corresponding estimated variables will be indexed by k.

2.2.1 Updating $\mathsf{q}_{\mathbf{X}}(\mathbf{x})$

Because of the majorization step, we need to minimize the upper bound on the $\mathcal{KL}$ divergence. The standard solution (2.87) can still be used by replacing the joint distribution by a lower bound $L(\boldsymbol{\Theta}, \mathbf{z}; \mathbf{w}, \boldsymbol{\lambda})$ chosen proportional to $L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda})$:

$$\begin{aligned}
\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x}) &\propto \exp\left(\left\langle \log L(\mathbf{x}, \gamma, \mathbf{z}; \mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)})\right\rangle_{\mathsf{q}_{\Gamma}^{(k)}(\gamma)}\right) \\
&\propto \exp\left(\int \log L(\mathbf{x}, \gamma, \mathbf{z}; \mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)}) \mathsf{q}_{\Gamma}^{(k)}(\gamma) \mathrm{d}\gamma\right) \\
&\propto \exp\left(-\sum_{i=1}^{N} \frac{1}{2}\mu_i(z_i)\left([\mathbf{H}\mathbf{x}]_i - w_i^{(k)}\right)^2\right. \\
&\qquad \left. - \frac{a_k}{b_k}\sum_{j=1}^{J} \frac{\kappa\|\mathbf{D}_j\mathbf{x}\|^2 + (1-\kappa)\lambda_j^{(k)}}{(\lambda_j^{(k)})^{1-\kappa}}\right).
\end{aligned} \tag{5.17}$$

The above distribution can be identified as a multivariate Gaussian distribution whose covariance matrix and mean parameter are given by

$$\boldsymbol{\Sigma}_{k+1}^{-1} = \mathbf{H}^{\top}\mathrm{Diag}(\boldsymbol{\mu}(\mathbf{z}))\mathbf{H} + 2\frac{a_k}{b_k}\mathbf{D}^{\top}\boldsymbol{\Lambda}^{(k)}\mathbf{D}, \tag{5.18}$$

$$\mathbf{m}_{k+1} = \boldsymbol{\Sigma}_{k+1}\mathbf{H}^{\top}\mathbf{u}, \tag{5.19}$$

where $\boldsymbol{\mu}(\mathbf{z}) = [\mu_1(z_1), \ldots, \mu_M(z_N)]^{\top}$, $\mathbf{u}$ is a $N \times 1$ vector whose i-th component is given by $u_i = \mu_i(z_i)w_i^{(k)}$ and $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal elements are $\left(\kappa(\lambda_j^{(k)})^{\kappa-1}\mathbf{I}_S\right)_{1\leqslant j\leqslant J}$.

2.2.2 Updating $\mathbf{w}$

The auxiliary variable $\mathbf{w}$ is determined by minimizing the upper bound of $\mathcal{KL}$ divergence with respect to this variable:

$$\begin{aligned}
\mathbf{w}^{(k+1)} &= \arg\min_{\mathbf{w}} \int \mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma) \log \frac{\mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma)}{L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}^{(k)})} \mathrm{d}\mathbf{x}\mathrm{d}\gamma \\
&= \arg\min_{\mathbf{w}} \int \mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma) \left(-\log L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}, \boldsymbol{\lambda}^{(k)})\right) \mathrm{d}\mathbf{x}\mathrm{d}\gamma \\
&= \arg\min_{\mathbf{w}} \int \mathsf{q}_{\mathbf{X}}^{(k+1)}(\mathbf{x}) \sum_{i=1}^{N} T_i([\mathbf{H}\mathbf{x}]_i, w_i; z_i) \mathrm{d}\mathbf{x} \qquad (5.20) \\
&= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} T_i([\mathbf{H}\mathbf{m}_{k+1}]_i, w_i; z_i), \qquad (5.21)
\end{aligned}$$

where the equality in (5.20) follows from the expression in (5.10). Interestingly, it follows from Property 2.1 that

$$\begin{aligned} w_i^{(k+1)} &= \arg\min w_i T_i([\mathbf{H}\mathbf{m}_{k+1}]_i, w_i; z_i) \\ &= [\mathbf{H}\mathbf{m}_{k+1}]_i - \frac{1}{\mu_i(z_i)}\phi_i'([\mathbf{H}\mathbf{m}_{k+1}]_i; z_i). \end{aligned} \tag{5.22}$$

2.2.3 Updating $\boldsymbol{\lambda}$

The variable $\boldsymbol{\lambda}$ is determined in a similar way: for every $j \in \{1, \ldots, J\}$,

$$\begin{aligned} \lambda_j^{(k+1)} &= \arg\min \lambda_j \in [0, +\infty) \mathcal{KL}(q_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma) \| L(\boldsymbol{\Theta}|\mathbf{z}; \mathbf{w}^{(k+1)}, \boldsymbol{\lambda})) \\ &= \arg\min \lambda_j \in [0, +\infty) \sum_{i=1}^{Q} \int q_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma) Q_i(\mathbf{D}_i\mathbf{x}, \lambda_i; \gamma) \mathrm{d}\mathbf{x}\mathrm{d}\gamma \\ &= \arg\min \lambda_j \in [0, +\infty) \int q_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma) Q_j(\mathbf{D}_j\mathbf{x}, \lambda_j; \gamma) \mathrm{d}\mathbf{x}\mathrm{d}\gamma \\ &= \arg\min \lambda_j \in [0, +\infty) \int q_{\mathbf{X}}^{(k+1)}(\mathbf{x}) q_{\Gamma}^{(k)}(\gamma) \\ &\qquad \times \gamma \, \frac{\kappa\|\mathbf{D}_j\mathbf{x}\|^2 + (1-\kappa)\lambda_j}{\lambda_j^{1-\kappa}} \mathrm{d}\mathbf{x}\mathrm{d}\gamma \\ &= \arg\min \lambda_j \in [0, +\infty) \frac{\kappa \, \mathsf{E}_{q_{\mathbf{X}}^{(k+1)}(\mathbf{x})}\left[\|\mathbf{D}_j\mathbf{x}\|^2\right] + (1-\kappa)\lambda_j}{\lambda_j^{1-\kappa}}. \end{aligned} \tag{5.23}$$

The minimum is achieved at

$$\begin{aligned} \lambda_j^{(k+1)} &= \mathsf{E}_{q_{\mathbf{X}}^{(k+1)}(\mathbf{x})}\left[\|\mathbf{D}_j\mathbf{x}\|^2\right] \\ &= \|\mathbf{D}_j\mathbf{m}_{k+1}\|^2 + \mathrm{trace}\left[\mathbf{D}_j^\top\mathbf{D}_j\boldsymbol{\Sigma}_{k+1}\right]. \end{aligned} \tag{5.24}$$

2.2.4 Updating $\mathsf{q}_\Gamma(\gamma)$

Using (2.87) where the joint distribution is replaced by its lower bound function, we obtain

$$\begin{aligned} \mathsf{q}_\Gamma^{(k+1)}(\gamma) &\propto \exp\left(\left\langle \log L(\mathbf{x}, \gamma, \mathbf{z}; \mathbf{w}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)})\right\rangle_{q_{\mathbf{X}}^{(k+1)}(\mathbf{x})}\right) \\ &\propto \exp\left(\int \log L(\mathbf{x}, \gamma, \mathbf{z}; \mathbf{w}^{(k+1)}, \boldsymbol{\lambda}^{(k+1)}) q_{\mathbf{X}}^{(k+1)}(\mathbf{x}) d\mathbf{x}\right) \\ &\propto \gamma^{\frac{Q}{2\kappa}+\alpha-1}\exp(-\beta\gamma) \\ &\quad \times \exp\left(-\gamma\sum_{j=1}^{J}\frac{\kappa\mathsf{E}_{q_{\mathbf{X}}^{(k+1)}(\mathbf{x})}\left[\|\mathbf{D}_j\mathbf{x}\|^2\right] + (1-\kappa)\lambda_j^{(k+1)}}{(\lambda_j^{(k+1)})^{1-\kappa}}\right) \\ &\equiv \mathcal{G}(a_{k+1}, b_{k+1}). \end{aligned} \tag{5.25}$$

Using (5.24), one can recognize that the above distribution is a Gamma one with parameters

$$a_{k+1} = \frac{Q}{2\kappa} + \alpha = a, \qquad b_{k+1} = \sum_{j=1}^{J} (\lambda_j^{(k+1)})^{\kappa} + \beta. \tag{5.26}$$

2.2.5 Resulting algorithm

The proposed method is outlined in Algorithm 10. It alternates between the update of the auxiliary variables and the distribution of the unknown parameters.

Algorithm 10 VBA approach for recovery of signals corrupted with non-Gaussian noise.

Initialize: $\mathbf{w}^{(0)}, \boldsymbol{\lambda}^{(0)}, b_0$. Compute a with (5.26).

1: **for** $k = 0, 1, \ldots$ **do**
2: Update parameters $\boldsymbol{\Sigma}_{k+1}$ and $\mathbf{m}_{k+1}$ of $q_{\mathbf{X}}^{(k+1)}(\mathbf{x})$ using (5.18) and (5.19).
3: Update $\mathbf{w}^{(k+1)}$ using (5.22).
4: Update $\boldsymbol{\lambda}^{(k+1)}$ using (5.24).
5: Update parameter b_{k+1} of $q_{\Gamma}^{(k+1)}(\gamma)$ using (5.26).
6: **end for**

2.3 Implementation issues

An additional difficulty arising in the implementation of Algorithm 10 is that the determination of $\boldsymbol{\Sigma}_{k+1}$ requires inverting the matrix given by (5.18), which is computationally expensive in high dimension. To bypass this operation, we propose to compare two approaches. The first one follows the idea in [Babacan et al., 2011]: we make use of the linear conjugate gradient method to approximate $\mathbf{m}_{k+1}$ iteratively and in (5.24), where an explicit form of $\boldsymbol{\Sigma}_{k+1}$ cannot be sidestepped, this matrix is approximated by a diagonal one whose diagonal entries are equal to the inverse of the diagonal elements of $\boldsymbol{\Sigma}_{k+1}^{-1}$. The second technique uses Monte-Carlo sample averaging to approximate $\mathbf{m}_{k+1}$ and $\lambda_j^{(k+1)}$: specifically, we generate samples $(\mathbf{n}_s)_{1 \leqslant s \leqslant N_s}$ from Gaussian distribution with mean $\mathbf{m}_{k+1}$ and covariance matrix $\boldsymbol{\Sigma}_{k+1}$ using [Orieux et al., 2012], as summarized in Algorithm 11. This estimator has two desirable properties. First, its accuracy is independent of the problem size, its relative error only depends on the number of samples and it decays as $\sqrt{2/N_s}$ (only $N_s = 2/\rho^2$ samples are required to reach a desired relative error ρ) [Papandreou and Yuille, 2010]. Second, for the simulation of N_s independent Gaussian samples, one can take advantage of a

multiprocessor architecture by resorting to parallel implementation allowing us to reduce the computation time.

Algorithm 11 Stochastic approach for computing the parameters of $\mathsf{q}(\mathbf{x})$.

1: **for** $s = 1, 2, \ldots, N_s$ **do**
2: **Perturbation:** Generate

$$\boldsymbol{\nu}_s \sim \mathcal{N}\left(\mathbf{u}, \left(\mathrm{Diag}\big(\boldsymbol{\mu}(\mathbf{z})\big)\right)^{1/2}\right)$$
$$\boldsymbol{\eta}_s \sim \mathcal{N}\left(0, \sqrt{2\gamma^{(k)}}\boldsymbol{\Lambda}_k^{1/2}\right)$$

with $\gamma^{(k)} = a_k/b_k$.
3: **Optimization**: Compute $\mathbf{n}_s$ as the minimizer of

$$\mathcal{J}(\mathbf{v}) = \|\boldsymbol{\nu}_s - \mathrm{Diag}\big(\boldsymbol{\mu}(\mathbf{z})\big)\mathbf{H}\mathbf{v}\|^2_{(\mathrm{Diag}(\boldsymbol{\mu}(\mathbf{z})))^{-1}} + \frac{1}{2\gamma^{(k)}}\|\boldsymbol{\eta}_s - 2\gamma^{(k)}\boldsymbol{\Lambda}_k\mathbf{D}\mathbf{v}\|^2_{\boldsymbol{\Lambda}_k^{-1}}$$

which is equivalent to minimizing

$$\tilde{\mathcal{J}}(\mathbf{v}) = \mathbf{v}^\top\boldsymbol{\Sigma}_{k+1}^{-1}\mathbf{v} - 2\mathbf{v}^\top\mathbf{z}_s$$

where

$$\mathbf{z}_s = \mathbf{H}^\top\boldsymbol{\nu}_s + \mathbf{D}^\top\boldsymbol{\eta}_s.$$

The minimizer is computed using the conjugate gradient algorithm.
4: **end for**
5: Update

$$\mathbf{m}_{k+1} = \frac{1}{N_s}\sum_{s=1}^{N_s}\mathbf{n}_s.$$
$$(\forall j \in \{1, \ldots, J\}) \quad \lambda_j^{(k+1)} = \frac{1}{N_s}\sum_{s=1}^{N_s}\|\mathbf{D}_j\mathbf{n}_s\|^2.$$

§ 3 Application to PG image restoration

Let us now illustrate the usefulness of our algorithm via experiments in the context of image restoration when the noise follows a mixed PG model. Recently, there has been a growing interest for the PG noise model as it arises in many real imaging systems in astronomy [Benvenuto et al., 2008; Snyder et al., 1993], medicine [Nichols et al., 2002], photography [Julliand et al., 2015], and biology [Delpretti et al., 2008]. Numerous efficient restoration

methods exist in the limit case when one neglects either the Poisson or the Gaussian component. However, such approximation may be rough, and lead to poor restoration results, especially in the context of low count imaging and/or high level electronic noise. On the opposite, restoration methods that specifically address mixed PG noise remain scarce, especially when the observation operator $\mathbf{H}$ differs from identity. The aim of this section is to show the applicability of the proposed VBA method in this context.

3.1 Problem formulation

The vector of observations $\mathbf{z} = (z_i)_{1 \leq i \leq N} \in \mathbb{R}^N$ is related to the original image $\overline{\mathbf{x}}$ through

$$\mathbf{z} = \mathbf{y} + \mathbf{w}, \tag{5.27}$$

where $\mathbf{z}$ and $\mathbf{w}$ are assumed to be mutually independent random vectors and

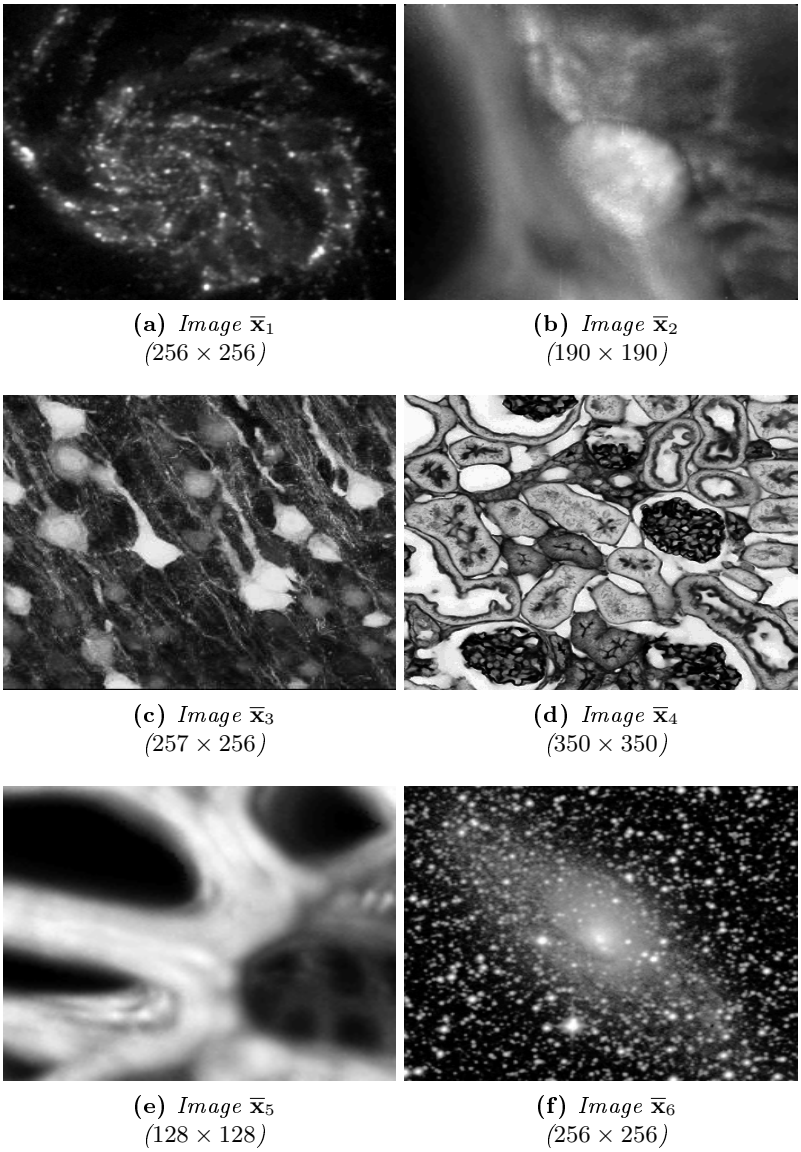
$$\mathbf{y} \mid \overline{\mathbf{x}} \sim \mathcal{P}(\mathbf{H}\overline{\mathbf{x}}), \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2 \mathbf{I}_N),$$

$\mathcal{P}$ denoting the independent Poisson distribution, and $\sigma \geqslant 0$. When $\mathbf{w} = \mathbf{0}_N$ (i.e., $\sigma = 0$), the model reduces to a pure Poisson image recovery. Otherwise, when $\sigma > 0$, the noise is a mixture of Poisson and Gaussian noise and the associated likelihood function reads [Chouzenoux et al., 2015]:

$$\mathsf{p}(\mathbf{z} \mid \mathbf{x}) = \prod_{i=1}^{N} \left(\sum_{n=1}^{+\infty} \frac{e^{-[\mathbf{H}\mathbf{x}]_i} \left([\mathbf{H}\mathbf{x}]_i\right)^n}{n!} \frac{e^{-\frac{1}{2\sigma^2}(z_i - n)^2}}{\sqrt{2\pi\sigma^2}} \right). \tag{5.28}$$

The expression of the PG likelihood (5.28) involves an infinite sum which makes its exact computation impossible in practice. In [Chouzenoux et al., 2015], the infinite sum was replaced by a finite summation with bounds depending on the current estimate of $\overline{\mathbf{x}}$. However, this strategy implies a higher computational burden in the reconstruction process when compared with other likelihoods proposed in the literature as accurate approximations of (5.28). In [Marnissi et al., 2016b], we have applied VBA inference techniques to the restoration of data corrupted with PG noise using the generalized Anscombe transform (GAST) likelihood [Murtagh et al., 1995; Starck and Murtagh, 1998; Mäkitalo and Foi, 2013, 2012]. Following these promising preliminary results, we will consider here the GAST approximation, as well as the shifted Poisson (SPoiss) [Chakrabarti and Zickler, 2012] and the weighted least squares (WL2) [Benvenuto et al., 2008; Li et al., 2015; Repetti et al., 2012] approximations, defined respectively in lines 4, 5 and 6 of Table 5.1. In order to satisfy Assumption 2.1, we will use $\mu_i(z_i) \equiv \max\{\beta_i(z_i), \varepsilon\}$ where $\varepsilon > 0$ for the GAST and the SPoiss approximations. For the WL2 approximation, we set $\mu_i(z_i) = \max\left\{(z_i + \sigma^2)^2/\sigma^6, \varepsilon\right\}$. Note that in all our experiments, a data truncation is performed as a pre-processing step on the observed image $\mathbf{y}$ in order to satisfy the domain condition given in the fifth column of Table 5.1.

3.2 Numerical results



(a) *Image* $\overline{\mathbf{x}}_1$ *(*256×256*)*
(b) *Image* $\overline{\mathbf{x}}_2$ *(*190×190*)*
(c) *Image* $\overline{\mathbf{x}}_3$ *(*257×256*)*
(d) *Image* $\overline{\mathbf{x}}_4$ *(*350×350*)*
(e) *Image* $\overline{\mathbf{x}}_5$ *(*128×128*)*
(f) *Image* $\overline{\mathbf{x}}_6$ *(*256×256*)*

Figure 5.1: *Original images.*

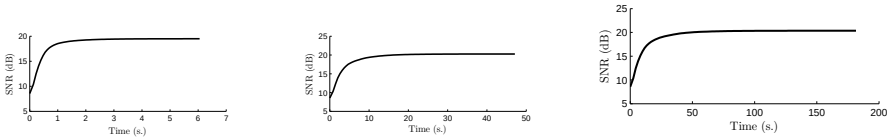
We evaluate the performance of the proposed approach for the restoration of images degraded by both blur and PG noise. We consider six test images, displayed in Figure 5.1, whose intensities have been rescaled so that pixel values belong to a chosen interval $[0, x^+]$. Images $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_6$ are HST astronomical images while images $\overline{\mathbf{x}}_2$, $\overline{\mathbf{x}}_3$, $\overline{\mathbf{x}}_4$ and $\overline{\mathbf{x}}_5$ correspond to the set of confocal microscopy images considered in [Chouzenoux et al., 2015]. These

images are then artificially degraded by an operator $\mathbf{H}$ modeling spatially invariant blur with point spread function h and by PG noise with variance σ^2.

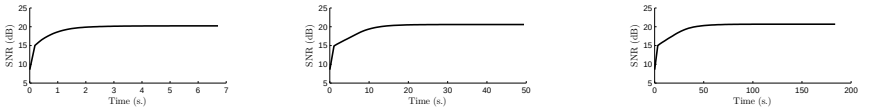
3.2.1 Comparison with MAP approaches

In this first set of experiments, we choose a standard total variation prior, i.e. $\kappa = 1/2$ and for every pixel $j \in \{1, \ldots, Q\}$, $\mathbf{D}_j\mathbf{x} = \left[[\nabla^h \mathbf{x}]_j, [\nabla^v \mathbf{x}]_j\right]^\top \in \mathbb{R}^2$ where ∇^h and ∇^v are the discrete gradients computed in the horizontal and vertical directions. As a result, $J = Q$ and $S = 2$. The goal of our experiments is twofold. First, for each likelihood, we compare the accuracy of the two proposed approximations of the covariance matrix described in Section 2.3 namely the diagonal approximation (denoted as approximation 1) and the Monte Carlo averaging strategy (designated as approximation 2) with different number of samples N_s, namely $N_s = 160$ or 640. Second, the proposed method is compared with state-of-the-art algorithms that compute the MAP estimate for the considered likelihoods. More specifically, as GAST and SPoiss data fidelity terms are convex and Lipschitz differentiable, we use the method presented in [Chouzenoux et al., 2015] where a primal-dual splitting algorithm was proposed to minimize convex penalized criteria in the context of Poisson-Gaussian image restoration. For the WL2 approximation, the corresponding data fidelity function is not convex so the previous method could not be applied anymore. We thus consider the variable metric forward-backward algorithm proposed in [Repetti et al., 2012] for the minimization of penalized WL2 functionals. For the aforementioned MAP approaches, it is necessary to set the regularization parameter γ that balances the fidelity to the observation model and the considered prior. In this respect, we test two variants. In the first variant, we estimate the regularization parameter using an approach based on the discrepancy principle [Bardsley and Goldes, 2009; Bertero et al., 2010; Zanni et al., 2015]. In the second variant, γ is adjusted empirically to achieve the maximum SNR value, which requires the availability of the true image.

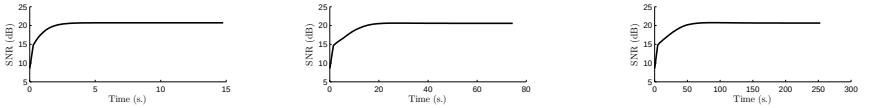
Tables 5.2-5.7 report the results obtained with the different images in terms of SNR, SSIM [Wang et al., 2004], and approximate computation time needed for convergence. For each likelihood, we emphasize in bold the approximation of the covariance matrix that achieves the best quantitative result in the shortest computational time. Simulations were performed on an Intel(R) Xeon(R) CPU E5-2630, @ 2.40 GHz, using a Matlab7 implementation. All tested methods were initialized with the degraded image. Moreover, the initial value of the regularization parameter results from a maximum likelihood estimation performed on the degraded image. The Monte Carlo averaging approximation was computed using parallel implementation with 16 cores by means of the command *PARFOR* of the Matlab® Parallel Computing Toolbox™. The iterations of VBA were run until the following



(a) *GAST - Approx. 1* **(b)** *GAST - Approx. 2* $(N_s = 160)$ **(c)** *GAST - Approx. 2* $(N_s = 640)$



(d) *SPoiss - Approx. 1* **(e)** *SPoiss - Approx. 2* $(N_s = 160)$ **(f)** *SPoiss - Approx. 2* $(N_s = 640)$



(g) *WL2 - Approx. 1* **(h)** *WL2 - Approx. 2* $(N_s = 160)$ **(i)** *WL2 - Approx. 2* $(N_s = 640)$

Figure 5.2: *Evolution of SNR with respect to time for image* $\bar{\mathbf{x}}_5$ *using different data-fidelity terms and covariance approximations.*

stopping criterion is satisfied: $\dfrac{\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|}{\|\mathbf{x}^{(t)}\|} \leqslant \varepsilon$. We have set $\varepsilon = 10^{-6}$, as it was observed to lead to a practical stabilization in terms of restoration quality. This can be checked by inspecting Figure 5.2 illustrating the evolution of the SNR of the restored image along time, until the achievement of the stopping criterion, in the test case from Table 5.6. For the MAP-based approaches, the computational time includes the search of the regularization parameter.

One can observe that in most studied situations (see Tables 5.4-5.7), the diagonal approximation of the covariance matrix appears to give satisfactory qualitative results after a small computation time. However, in few other situations (see Tables 5.2 and 5.3), it fails to capture the real qualitative structures of the covariance matrix leading to a poorer performance. The latter issue is well alleviated by using the Monte Carlo approximation where good results, in terms of image quality, are achieved within $N_s = 160$ samples which is equivalent to a relative approximation error equal to 11%. A few improvements are observed by decreasing the approximation error to 5% using $N_s = 640$ samples.

We also notice that the GAST approximation does not seem to be suitable for very low count images (see Tables 5.2 and 5.3), whereas, the other likelihoods lead to competitive results in all the experiments. The best trade-

Table 5.2: *Restoration results for image* $\overline{\mathbf{x}}_1$ *with* $x^+ = 10$ *and* $\sigma^2 = 4$. *Uniform kernel with size* 5×5. *Initial SNR= -2.55 dB.*

			GAST	SPoiss	WL2
VBA	Approx. 1	SNR	8.13	9.36	9.90
		SSIM	0.3987	0.4790	0.5140
		Time (s.)	55	62	67
	Approx. 2 $N_s = 160$	SNR	9.57	10.17	10.22
		SSIM	0.5260	0.6017	0.6058
		Time (s.)	688	601	1011
	Approx. 2 $N_s = 640$	SNR	**9.61**	**10.20**	**10.27**
		SSIM	**0.5308**	**0.6088**	**0.6112**
		Time (s.)	**3606**	**3507**	**3510**
MAP	Discrepancy principle	SNR	-1.13	5.24	10.17
		SSIM	0.0980	0.2961	0.6131
		Time (s.)	3326	2215	3053
	Best parameter	SNR	9.46	10.40	10.39
		SSIM	0.5078	0.6029	0.5920
		Time (s.)	4380	2560	13740

Table 5.3: *Restoration results for the image* $\overline{\mathbf{x}}_2$, $x^+ = 12$ *and* $\sigma^2 = 9$. *Gaussian kernel with size* 25×25, *std 1.6. Initial SNR= 2.21 dB.*

			GAST	SPoiss	WL2
VBA	Approx. 1	SNR	13.97	15.19	16.41
		SSIM	0.3544	0.4167	0.4959
		Time (s.)	58	64	70
	Approx. 2 $N_s = 160$	SNR	18.05	19.07	19.11
		SSIM	0.6664	0.6930	0.7066
		Time (s.)	524	498	491
	Approx. 2 $N_s = 640$	SNR	**18.11**	**19.12**	**19.13**
		SSIM	**0.6778**	**0.7034**	**0.7152**
		Time (s.)	**2048**	**1828**	**1735**
MAP	Discrepancy principle	SNR	16.52	17.41	18.09
		SSIM	0.5484	0.7570	0.6732
		Time (s.)	594	583	2286
	Best parameter	SNR	17.83	18.73	19.09
		SSIM	0.6519	0.6646	0.6702
		Time (s.)	674	705	4164

off between restoration quality and small computational time seems to be achieved by the WL2 approximation.

Table 5.4: *Restoration results for the image* $\overline{\mathbf{x}}_3$ *with* $x^+ = 15$ *and* $\sigma^2 = 9$. *Uniform kernel with size* 5×5. *Initial SNR= 3.14 dB.*

			GAST	SPoiss	WL2
VBA	Approx. 1	SNR	11.42	11.94	12.25
		SSIM	0.4184	0.4403	0.4588
		Time (s.)	45	47	53
	Approx. 2 $N_s = 160$	SNR	12.04	12.31	12.28
		SSIM	0.4555	0.4624	0.4627
		Time (s.)	328	332	396
	Approx.2 $N_s = 640$	SNR	**12.09**	**12.36**	**12.33**
		SSIM	**0.4617**	**0.4684**	**0.4683**
		Time (s.)	**1965**	**2051**	**2019**
MAP	Discrepancy principle	SNR	12.08	12.38	12.08
		SSIM	0.4523	0.4582	0.4314
		Time (s.)	6252	3865	1929
	Best parameter	SNR	12.17	12.45	12.37
		SSIM	0.4531	0.4576	0.4565
		Time (s.)	3348	2441	2525

Table 5.5: *Restoration results for the image* $\overline{\mathbf{x}}_4$ *with* $x^+ = 20$ *and* $\sigma^2 = 9$. *Uniform kernel with size* 5×5. *Initial SNR= 7.64 dB.*

			GAST	SPoiss	WL2
VBA	Approx. 1	SNR	**13.80**	**13.90**	**13.66**
		SSIM	**0.5752**	**0.5769**	**0.5582**
		Time (s.)	**29**	**34**	**88**
	Approx. 2 $N_s = 160$	SNR	13.72	13.76	13.56
		SSIM	0.5667	0.5641	0.5491
		Time (s.)	555	580	757
	Approx. 2 $N_s = 640$	SNR	13.78	13.81	13.61
		SSIM	0.5715	0.5687	0.5534
		Time (s.)	1897	2170	2719
MAP	Discrepancy principle	SNR	13.48	13.60	13.39
		SSIM	0.5348	0.5393	0.5103
		Time (s.)	3049	769	2644
	Best parameter	SNR	13.60	13.71	13.75
		SSIM	0.5568	0.5605	0.5602
		Time (s.)	8390	8477	2397

Finally, it can be observed that, in Tables 5.2 and 5.4, our VBA method yields comparable performance in terms of SNR to the MAP estimate when the latter is computed with the optimal regularization parameter, while our

Table 5.6: *Restoration results for image* $\overline{\mathbf{x}}_5$ *with* $x^+ = 20$ *and* $\sigma^2 = 9$. *Gaussian kernel with size* 7×7, *std 1. Initial SNR= 8.55 dB.*

			GAST	SPoiss	WL2
VBA	Approx. 1	SNR	19.5	20.23	**20.71**
		SSIM	0.6649	0.7135	**0.7793**
		Time (s.)	16	17	**34**
	Approx. 2 $N_s = 160$	SNR	20.27	20.59	20.56
		SSIM	0.7473	0.7660	0.7877
		Time (s.)	61	64	94
	Approx.2 $N_s = 640$	SNR	**20.35**	**20.67**	20.64
		SSIM	**0.7563**	**0.7798**	0.7989
		Time (s.)	**195**	**197**	272
MAP	Discrepancy principle	SNR	19.39	19.50	18.70
		SSIM	0.7458	0.7550	0.7448
		Time (s.)	717	1201	1087
	Best parameter	SNR	20.15	20.41	20.44
		SSIM	0.7535	0.7594	0.7628
		Time (s.)	559	125	253

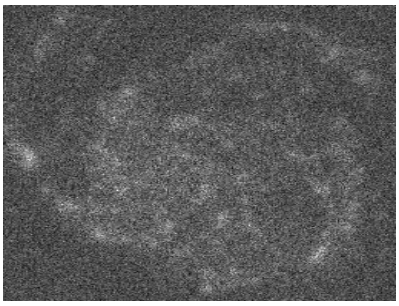
Table 5.7: *Restoration results for the image* $\overline{\mathbf{x}}_6$ *with* $x^+ = 100$ *and* $\sigma^2 = 36$. *Uniform kernel with size* 3×3. *Initial SNR= 10.68 dB.*

			GAST	SPoiss	WL2
VBA	Approx. 1	SNR	**14.17**	14.13	13.90
		SSIM	**0.7655**	0.7647	0.7569
		Time (s.)	**9**	8	26
	Approx.2 $N_s = 160$	SNR	14.1	14.13	14.09
		SSIM	0.7605	0.7619	0.7620
		Time (s.)	104	148	246
	Approx. 2 $N_s = 640$	SNR	14.16	**14.19**	**14.16**
		SSIM	0.7639	**0.7650**	**0.7658**
		Time (s.)	332	**479**	**913**
MAP	Discrepancy principle	SNR	13.23	13.29	13.32
		SSIM	0.7104	0.7126	0.7117
		Time (s.)	2796	4900	1045
	Best parameter	SNR	13.77	13.79	13.84
		SSIM	0.7565	0.7570	0.7591
		Time (s.)	10084	10005	821

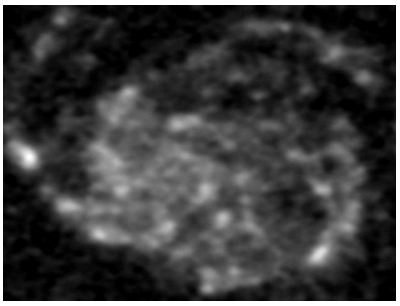
approach requires less time to converge. In the other experiments, our approach leads to the best qualitative results. For instance, in Table 5.3, the gain in terms of SNR reaches up to 0.2 dB compared with the MAP estimator

using the best regularization parameter, but our approach needs more time to converge. In Tables 5.5, 5.6 and 5.7, we achieve both the best quantitative results and the smallest computational time. It should be noted that for most tested scenarii, discrepancy based approaches perform relatively poorly compared with the other methods, especially in the case of low count images (see Table 5.2).

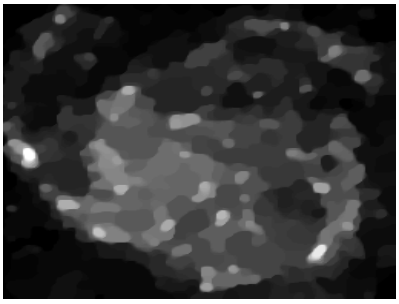
In Figures 5.3 - 5.8, we show some examples of visual results obtained with the different approaches, when the best approximation strategy for the covariance matrix is retained in the VBA method. It can be noticed that, unlike the other methods, the reconstructed images with the proposed VBA algorithm exhibit very few artifacts, without over-smoothed regions.



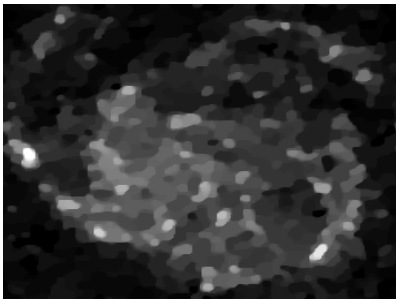
(a) *Degraded image with SNR= -2.55 dB (Uniform kernel* 5×5*,* $x^+ = 10$ *and* $\sigma^2 = 4$*).*



(b) *Restored image with VBA approach using the Monte Carlo approximation with* 640 *samples: SNR= 10.27 dB*



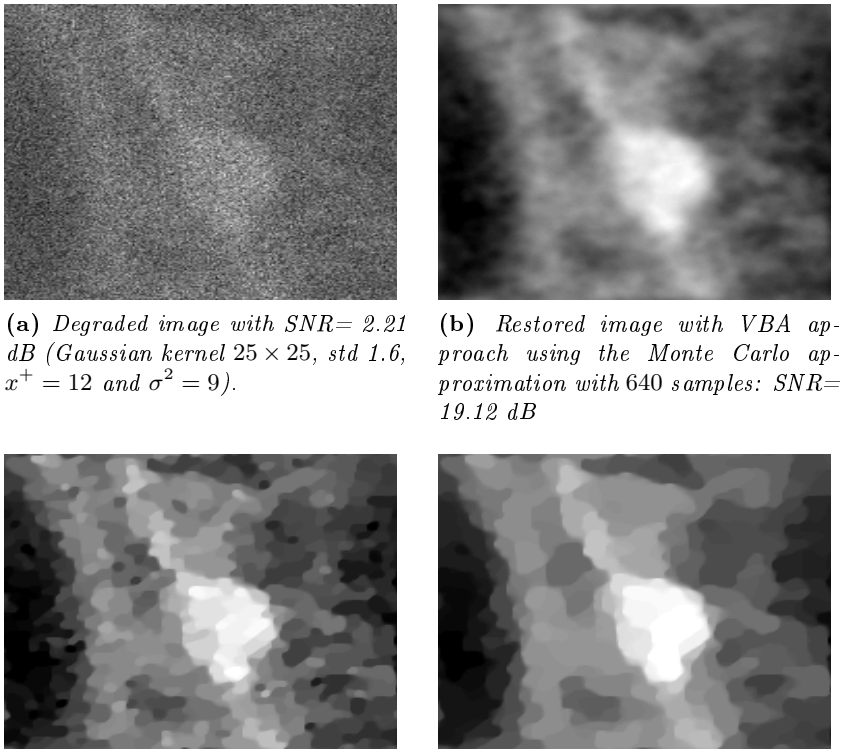
(c) *Restored image with discrepancy principle: SNR= 10.17 dB*



(d) *Restored image with best parameter: SNR= 10.39 dB*

Figure 5.3: *Restoration results for image* $\bar{\mathbf{x}}_1$ *using WL2 approximation.*

It should be emphasized that the problem of setting the regularization parameter for MAP-based algorithms must be carefully addressed as it highly impacts the quality of the restored image. The main advantage of our approach is that this parameter is tuned automatically without the need of the ground truth, while also often being the most competitive in terms of computation time. Furthermore, the performance of the proposed method could



(a) *Degraded image with SNR= 2.21 dB (Gaussian kernel* 25×25*, std 1.6,* $x^+ = 12$ *and* $\sigma^2 = 9$*).*

(b) *Restored image with VBA approach using the Monte Carlo approximation with* 640 *samples: SNR= 19.12 dB*

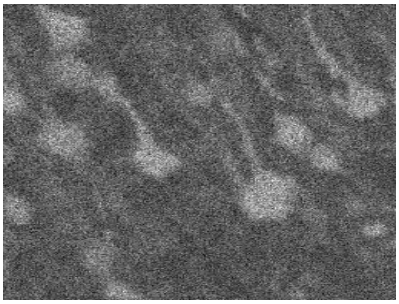
(c) *Restored image with discrepancy principle: SNR= 17.41 dB*

(d) *Restored image with best parameter: SNR= 18.73 dB*

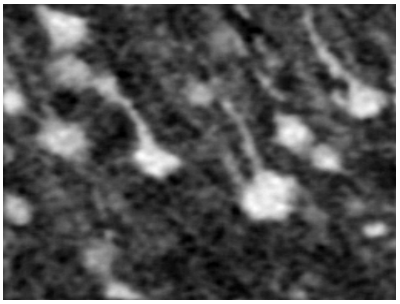
Figure 5.4: *Restoration results for image* $\bar{\mathbf{x}}_2$ *using SPoiss approximation.*

be further improved by using parallel implementation with more than 16 cores for the Monte Carlo approximation of the covariance matrix allowing either generating a higher number of samples (i.e. an improved estimation error) or a reduction of the computation time.

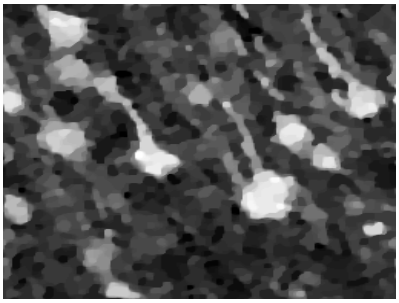
Comparisons with image deblurring methods dedicated to a pure Poisson noise model have also been conducted. However, in our examples, they were observed to lead to poor results in terms of restoration quality, and to present a high computational time. For instance, the application of the proximal method from [Pustelnik et al., 2011] using a TV prior and an empirical search for the regularization parameter, leads to an image with SNR equal to 12.88 dB (computation time: 3489 s.) on the test problem from Table 5.5, and a SNR of 18.37 dB (computation time: 986 s.) for the example from Table 5.6. The Plug and Play ADMM strategy from [Rond et al., 2016] also leads to unsatisfactory results with a final SNR of 9.11 dB (computation time: 1618 s.) and 10.31 dB (computation time: 204 s.) for the examples from Table 5.5 and Table 5.6, respectively. These numerical tests clearly highlight that image restoration in the presence of Poisson-Gaussian noise



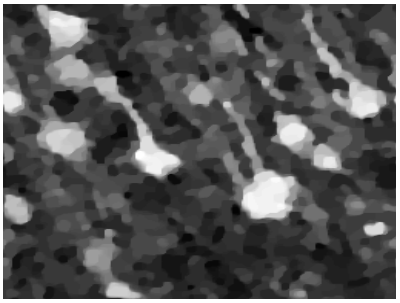
(a) *Degraded image with SNR= 3.14 dB (Uniform kernel* 5×5*,* $x^+ = 15$ *and* $\sigma^2 = 9$*).*



(b) *Restored image with VBA approach using the Monte Carlo approximation with* 640 *samples: SNR= 12.36 dB*



(c) *Restored image with discrepancy principle: SNR= 12.38 dB*



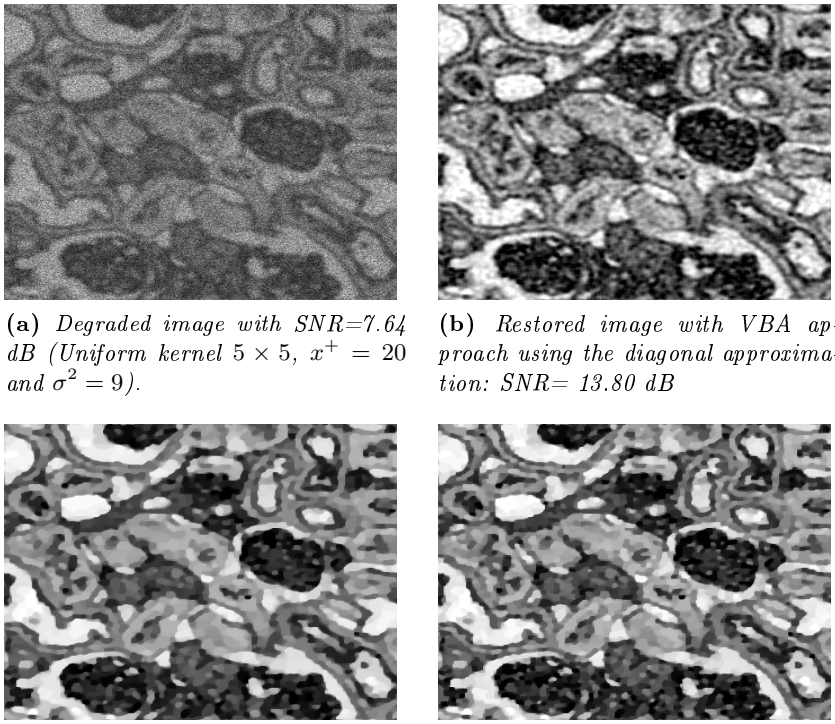
(d) *Restored image with best parameter: SNR= 12.45 dB*

Figure 5.5: *Restoration results for image* $\bar{\mathbf{x}}_3$ *using SPoiss approximation.*

is challenging, and should be treated with specific methods that take into account the mixed noise model in an explicit manner.

3.2.2 Influence of the regularization term

The versatility of the proposed VBA method allows us to consider a large variety of regularization strategies, by defining appropriate prior operators $\mathbf{D}$. In the previous experiments, the TV prior has led to satisfactory results in terms of SNR, but a visual inspection of the restored versions of images $\bar{\mathbf{x}}_4$ and $\bar{\mathbf{x}}_6$ shows an undesirable starcasing effect. In this new set of experiments, we propose to compare these TV-based restoration results to those obtained with priors that have been recently shown to better preserve the natural features in images. Namely, we will consider the Hessian-based penalization [Lefkimmiatis et al., 2012], the semi-local total variation (SLTV) [Condat, 2014], and the non-local total variation (NLTV) [Gilboa and Osher, 2008; Chierchia et al., 2014]. The Hessian prior operator is given, for every $j \in \{1, \ldots, Q\}$, by $\mathbf{D}_j\mathbf{x} = \left[[\nabla^{hh}\mathbf{x}]_j, \sqrt{2}[\nabla^{hv}\mathbf{x}]_j, [\nabla^{vv}\mathbf{x}]_j\right]^\top \in \mathbb{R}^3$



(a) *Degraded image with SNR=7.64 dB (Uniform kernel* 5×5*,* $x^+ = 20$ *and* $\sigma^2 = 9$*).*

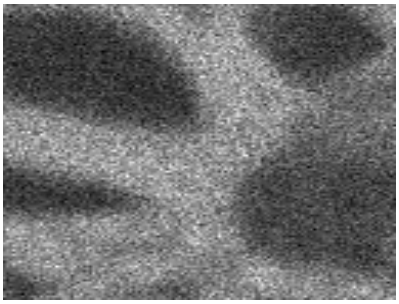
(b) *Restored image with VBA approach using the diagonal approximation: SNR= 13.80 dB*

(c) *Restored image with discrepancy principle: SNR= 13.48 dB*

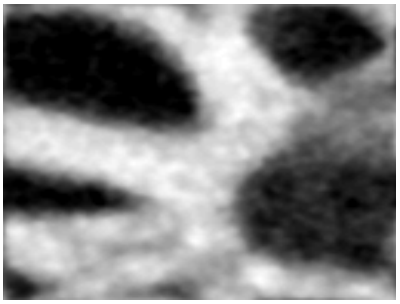
(d) *Restored image with best parameter: SNR= 13.60 dB*

Figure 5.6: *Restoration results for image* $\bar{\mathbf{x}}_4$ *using GAST approximation.*

where ∇^{hh}, ∇^{hv} and ∇^{vv} model the second-order finite difference operators between neighbooring pixels, so that $S = 3$ and $J = Q$. The SLTV is based on differences of neighboring gradient values and is computed here using a 6-pixels neighborhood, hence $S = 12$ and $J = Q$. The NLTV prior operator is defined at every pixel position by a collection of weighted discrete gradient differences operators across a large set of directions, the weights being calculated according to a rough estimate of the target image. In our experiments, 49 different directions are chosen and the corresponding weights are precomputed from the restored images using VBA with the TV prior and the diagonal approximation of the covariance matrix. As a result, $S = 98$ and $J = Q$ in that case. The SPoiss likelihood is chosen for the data fidelity term as it was observed to lead to the best tradeoff in terms of image quality and computational time in the previous set of tests. Table 5.8 summarizes the obtained results for all the six test images, using the different considered priors. Complementary to these numerical results, Figures 5.9 and 5.10 show the visual improvements resulting from the different priors. One can observe that the NLTV prior gives in most experiments the best results in terms of SNR while the other priors perform quite similarly. In particular,

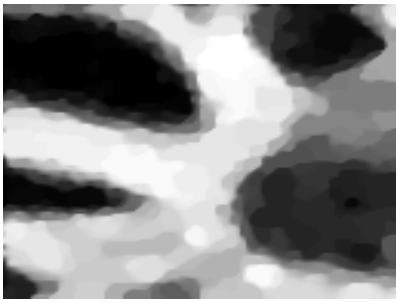


(a) *Degraded image with SNR= 8.55 dB (Gaussian kernel 7 × 7, std 1. $x^+ = 20$ and $\sigma^2 = 9$).*



(b) *Restored image with VBA approach using the diagonal approximation: SNR= 20.71 dB*

(c) *Restored image with discrepancy principle: SNR= 18.70 dB*



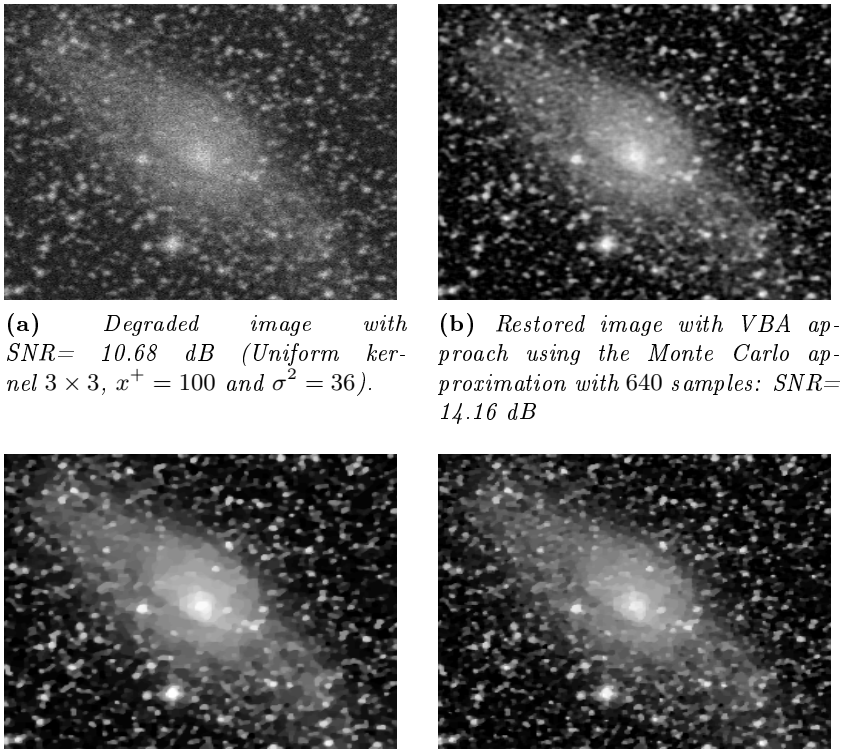
(d) *Restored image with best parameter: SNR= 20.44 dB*

Figure 5.7: *Restoration results for image $\bar{\mathbf{x}}_5$ using WL2 approximation.*

Table 5.8: *Restoration results for the considered test images using the SPoiss likelihood and different regularization functions.*

		$\overline{\mathbf{x}}_1$	$\overline{\mathbf{x}}_2$	$\overline{\mathbf{x}}_3$	$\overline{\mathbf{x}}_4$	$\overline{\mathbf{x}}_5$	$\overline{\mathbf{x}}_6$
TV	SNR	10.20	19.12	12.36	13.90	20.67	**14.19**
	SSIM	0.6088	0.6930	0.4684	0.5769	0.7790	**0.7650**
	Time (s.)	3507	1828	2051	34	184	**479**
Hessian	SNR	10.17	**19.41**	12.21	13.56	20.57	14.05
	SSIM	0.6016	**0.7300**	0.4618	0.5501	0.8392	0.7643
	Time (s.)	8600	**5404**	6974	5058	744	1332
SLTV	SNR	10.32	19.26	12.26	13.53	20.62	13.93
	SSIM	0.6006	0.7189	0.4656	0.5478	0.8368	0.7578
	Time (s.)	6359	2923	3497	1003	375	738
NLTV	SNR	**10.35**	19.10	**12.46**	**14.09**	**22.89**	13.95
	SSIM	**0.4644**	0.7075	**0.4704**	**0.5812**	**0.7972**	0.7530
	Time (s.)	**7821**	338	**4602**	**8595**	**807**	1547

for the image $\overline{\mathbf{x}}_5$, the gain in terms of SNR exceeds 2 dB when using the



(a) *Degraded image with SNR= 10.68 dB (Uniform kernel* 3×3*,* $x^+ = 100$ *and* $\sigma^2 = 36$*).*

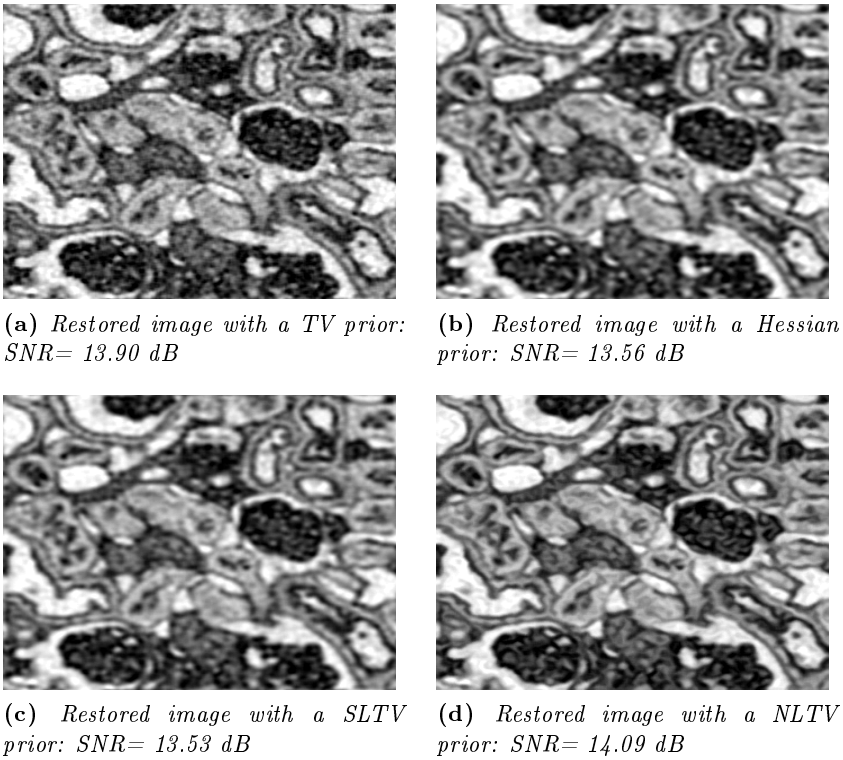
(b) *Restored image with VBA approach using the Monte Carlo approximation with* 640 *samples: SNR= 14.16 dB*

(c) *Restored image with discrepancy principle: SNR= 13.32 dB*

(d) *Restored image with best parameter: SNR= 13.84 dB*

Figure 5.8: *Restoration results for image* $\bar{\mathbf{x}}_6$ *using WL2 approximation.*

NLTV prior, compared to the other regularization strategies. Note that despite small differences in SNR between the results obtained with the TV, SLTV and the Hessian regularizers, the Hessian and the SLTV appear to offer good alternatives in terms of visual quality to the TV prior for images that consist mostly of ridges and smooth transition of intensities. Indeed, it can be seen in Figure 5.10 that the smooth piecewise constant areas are better reconstructed and the sharpness of edges is better maintained using these two priors. For textured images, Figure 5.9 shows that the NLTV prior gives rise to less blurry images than the SLTV and Hessian priors and seems to reduce again the undesired staircase effect arising from TV regularization. However, as shown in Table 5.8, the approaches based on Hessian, SLTV and NLTV take much more computation time than the TV based approach in most test cases. Our suggestion would be to use the VBA approach with the TV prior and the diagonal approximation of the covariance matrix to obtain a satisfactory result in a low computational cost, and to use VBA with NLTV prior, using the former TV-based result to approximate the NLTV weights, in order to further improve the visual quality of the restored image.



(a) *Restored image with a TV prior: SNR= 13.90 dB*

(b) *Restored image with a Hessian prior: SNR= 13.56 dB*

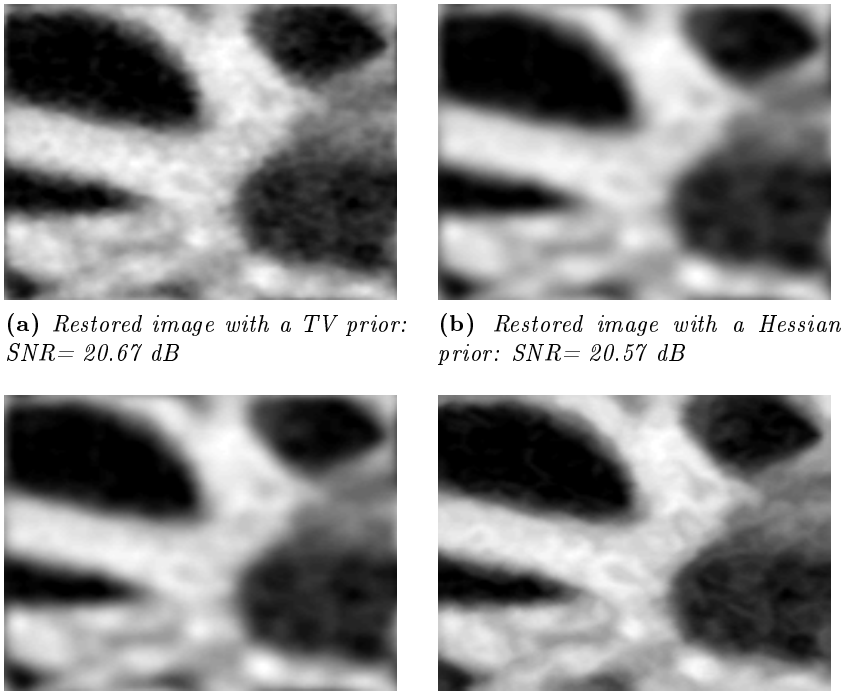
(c) *Restored image with a SLTV prior: SNR= 13.53 dB*

(d) *Restored image with a NLTV prior: SNR= 14.09 dB*

Figure 5.9: *Restoration results for image $\bar{\mathbf{x}}_4$ using SPoiss likelihood and different regularization functions.*

3.2.3 VBA performance in the case of pure Poisson noise

We now consider two test problems where the image is degraded by pure Poisson noise (i.e., $\sigma = 0$). Since the exact expression of the Poisson likelihood is not differentiable, we have ran our VBA method using its Anscombe-based approximation defined in line 3 of Table 5.1. The NLTV prior with weights precomputed from images restored using VBA with the TV prior, was employed as regularization term. We consider the test images $\overline{\mathbf{x}}_2$ and $\overline{\mathbf{x}}_5$ whose intensities are rescaled to achieve different intensity levels. The observed images are then generated by degrading the clean ones with a Gaussian blur of size 7×7 and variance 1 with symmetric boundary conditions and then by applying Poisson noise. Comparisons have been conducted between our VBA method, the Plug-and-Play scheme [Rond et al., 2016] and the variational approach using the Parallel Proximal algorithm (PPXA-TV) with a TV regularization [Pustelnik et al., 2011]. The regularization parameters for these two variational methods have been either fixed empirically to achieve



(a) *Restored image with a TV prior: SNR= 20.67 dB*

(b) *Restored image with a Hessian prior: SNR= 20.57 dB*

(c) *Restored image with a SLTV prior: SNR= 20.62 dB*

(d) *Restored image with a NLTV prior: SNR= 22.89 dB*

Figure 5.10: *Restoration results for image* $\bar{\mathbf{x}}_5$ *using SPoiss likelihood and different regularization functions.*

the best numerical results in terms of SNR or estimated using the discrepancy principle for Poisson likelihood [Bertero et al., 2010]. Results are shown in Tables 5.9 and 5.10. While the PPXA-TV yields very competitive results, the Plug-and-Play is quite efficient for very low count images ($x^+ < 5$) but its efficiency highly deteriorates for moderate values of x^+. The proposed method achieves the best restoration results for ($x^+ > 5$) but at the price of a larger computational cost.

Table 5.9: *Restoration results for the test image $\overline{\mathbf{x}}_2$ under pure Poisson noise (Anscombe transform likelihood and NLTV prior).*

		Plug-and-Play		PPXA-TV		Proposed
		Discrepancy principle	Best parameter	Discrepancy principle	Best parameter	
$x^+ = 3$ Initial SNR=1.37 dB	SNR	15.75	19.96	17.36	17.43	14.24
	SSIM	0.4789	0.7247	0.6090	0.6242	0.7121
	Time	1883	1172	7430	4421	24135
$x^+ = 5$ Initial SNR=3.58 dB	SNR	16.21	20.56	18.05	18.06	16.44
	SSIM	0.4632	0.7491	0.6440	0.6486	0.7372
	Time	2251	961	7824	4103	12629
$x^+ = 10$ Initial SNR=6.96 dB	SNR	16.97	17.03	10.39	19.13	18.40
	SSIM	0.7430	0.7434	0.5775	0.6739	0.7522
	Time	3217	316	14857	2603	11349
$x^+ = 15$ Initial SNR=8.57 dB	SNR	14.33	14.65	15.84	20.00	19.79
	SSIM	0.7620	0.7607	0.6705	0.6949	0.7634
	Time	1959	1238	7824	2676	21499
$x^+ = 20$ Initial SNR=9.76 dB	SNR	11.83	12.40	18.15	20.13	20.24
	SSIM	0.7612	0.7538	0.6992	0.7073	0.7696
	Time	1557	558	3718	2158	30885
$x^+ = 30$ Initial SNR=11.44 dB	SNR	10.04	10.10	19.95	20.66	21.08
	SSIM	0.6661	0.7320	0.7247	0.7212	0.8709
	Time	6211	568	6121	2233	55284

Table 5.10: *Restoration results for the test image $\overline{\mathbf{x}}_5$ under pure Poisson noise (Anscombe transform likelihood and NLTV prior).*

		Plug-and-Play		PPXA-TV		Proposed
		Discrepancy principle	Best parameter	Discrepancy principle	Best parameter	
$x^+ = 3$ Initial SNR=3.14 dB	SNR	17.36	19.71	14.91	16.47	14.89
	SSIM	0.7307	0.7749	0.5993	0.6201	0.7940
	Time	1101	782	7529	1617	3223
$x^+ = 5$ Initial SNR=5.27 dB	SNR	19.62	19.94	17.06	17.28	17.21
	SSIM	0.7952	0.7930	0.6650	0.6555	0.8172
	Time	950	518	6861	1707	3021
$x^+ = 10$ Initial SNR=8.52 dB	SNR	12.30	14.05	12.44	18.64	20.9
	SSIM	0.7703	0.7553	0.6123	0.6842	0.8406
	Time	996	316	2482	922	5523
$x^+ = 15$ Initial SNR=11.23 dB	SNR	11.68	12.00	15.43	19.03	21.06
	SSIM	0.7834	0.7950	0.6887	0.7256	0.8626
	Time	926	242	1120	1230	9423
$x^+ = 20$ Initial SNR=12.23 dB	SNR	9.24	10.25	18.27	19.66	21.90
	SSIM	0.7303	0.7405	0.7363	0.7419	0.8553
	Time	1228	203	1106	920	12459
$x^+ = 30$ Initial SNR=13.66 dB	SNR	8.20	10.10	19.74	19.85	22.76
	SSIM	0.7363	0.7320	0.7586	0.7555	0.8709
	Time	1271	568	1601	938	22888

Conclusion

In this chapter, we have proposed a variational Bayesian approach for solving signal recovery problems in the presence of non-Gaussian noise. Our approach has two main advantages. First, the regularization parameter is tuned automatically during the recovery process. Second, the designed method is applicable to a wide range of prior distributions and data fidelity terms. As the posterior density of the unknown parameters is analytically intractable, the estimation problem is derived in a variational Bayesian framework where the goal is to provide a good approximation to the posterior distribution in order to compute posterior mean estimates. Moreover, a majorization technique is employed to circumvent the difficulties raised by the intricate forms of the non-Gaussian likelihood and of the prior density. Simulations carried out on various images corrupted with mixed Poisson-Gaussian noise showcase the good performance of our approach compared with methods using the discrepancy principle for estimating the regularization parameter. Moreover, we propose variants of our method leading to a significant reduction of the computational cost while maintaining a satisfactory restoration quality.

- Chapter 6 -

Conclusion

The signal processing community has embraced the challenge to shed light on the interesting overlap existing between deterministic optimization approaches and Bayesian framework, being for a long time divergent. Interactions between these two classes of approaches have contributed to the development of algorithms that make use of stochastic simulation, approximation and optimization in order to provide novel efficient methods for signal restoration. Following this promising direction, this PhD thesis has contributed to the development of Bayesian algorithms whose efficiency is improved using deterministic optimization tools.

§ 1 Contributions

In the first part of this thesis, we have investigated the problem of sampling from intractable target distribution in high dimensional spaces. In the second part, we have proposed an image recovery algorithm using a VBA framework. Validation of the proposed approaches has been conducted through applications of signal and image recovery.

Majorize-Minimize adapted Metropolis Hastings: We have been interested on Langevin Metropolis Hastings algorithms defined for differentiable laws. Inspired from gradient descent optimization tools, such methods incorporate directional information about the target distribution in the proposal density in order to guide the chain towards the target space where most samples should be concentrated. To this end, the directional component is chosen as one iteration of a preconditioned gradient descent algorithm. Noise is then injected to the update in such a way that the trajectory of the chain converges to the full posterior distribution rather to a posterior mode. Hence, a new sample of such proposal is more likely localized in a region with high probability values and is then more likely accepted, which may speed up the convergence of the chain to the stationary distribution. However, as the Markov chain is very sensitive to the dependencies between the signal coefficients, sampling in large scale problems remains a challenging task because of the high cost at each iteration or because of

poor mixing properties of the chain when the preconditioning matrix is not well chosen. These issues are similar to problems encountered in deterministic preconditioned gradient descent optimization techniques. This dissertation led us to develop a sampling algorithm which exploits natural connections between the deterministic inverse problem and the Bayesian statistical inverse problem to accelerate statistical sampling methods:

- Inspired from Majorize-Minimize approaches frequently used in the deterministic framework, we have proposed a preconditioned version of the Langevin Metropolis Hastings algorithm that uses adaptive preconditioning matrices derived from a quadratic tangent majorant function of the negative logarithm of the posterior distribution.
- We have proposed different variants of tangent majorant functions involving full, constant and diagonal curvature matrices which allows the scalability of the proposed algorithms to large size problems.
- We have demonstrated the geometric ergodicity of the proposed sampling algorithm for the class of super-exponential distributions.
- The proposed algorithm has been validated on a sparse signal deconvolution problem with a Cauchy prior. This experiment allowed us to study the impact of the preconditioning matrix on the performance of the sampling process. Results have shown that, similarly to preconditioning in optimization, using an exact full matrix is often computationally expensive in high-dimensional settings which deteriorates the efficiency of the algorithm. The diagonal matrix has given the best tradeoff between convergence speed and mixing properties due to the minimal cost per iteration it induces.

Gibbs sampler with auxiliary variables: When the parameter space is high dimensional, the performance of stochastic sampling algorithms is very sensitive to dependencies between parameters. For instance, this problem emerges when one aims to sample from a high dimensional Gaussian distribution whose covariance matrix does not present simple structure i.e., it is neither sparse, nor circulant, nor Toeplitz etc. In this context, we often resort to sampling algorithms based on Perturbation-Optimization rule that requires to solve at each iteration a cost function using an iterative algorithm [Orieux et al., 2012] which makes the sampling process prohibitive especially when used within a Gibbs sampler. Another challenge is the design of MH proposals that make use of information about the local geometry of the target density

in order to speed up the convergence and improve mixing in the parameter space without being too computationally expensive. These two issues are mainly related to the presence of two heterogeneous sources of dependencies coming either from the prior or the likelihood in the sense that the related covariances matrices can not be diagonalized in the same domain. To overcome these difficulties, we have proposed 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 will only intervene through the auxiliary variables. Note that this strategy is highly related to half-quadratic approaches often used in optimization problems [Allain et al., 2006].

- We have described a strategy for adding auxiliary variables in a Gaussian distribution and we have extended it to scale mixture of Gaussian models.
- In the new augmented space, the Gibbs sampler needs to draw samples from the conditional distribution of the auxiliary variables at each iteration. We have therefore given some strategies to perform this task directly depending on the properties of the covariance matrix related to the auxiliary variables.
- We have studied the efficiency of the proposed approach on a problem of multichannel image recovery from its blurred and noisy version, the noise being assumed Gaussian with known variance. The problem has been addressed in the wavelet domain, where a multivariable GMEP prior has been adopted to model the wavelet coefficients at the same spatial position through all the channels allowing to exploit the spectral-intercorrelation. A separation strategy has been proposed to estimate the hyper-parameters involved in the GMEP regularization from the degraded observation of the image to be reconstructed. By adding auxiliary variables in the Gaussian data fidelity term, the observation matrix is no longer related directly to the image. Thanks to the separability property of the prior law (wavelets coefficients belonging to different spatial positions, orientation or scales are supposed to be independent), vectors of wavelet coefficients belonging to different wavelet subbands have been sampled independently in a parallel manner.
- We have shown the good performance of the proposed data augmentation approach when dealing with the problem of sampling from high-dimensional Gaussian distribution. Auxiliary variables can be added either to the prior model or the data fidelity term or to both of them depending on the properties of the related

covariance matrices. The experimental results illustrate, through an example of an image recovery problem under mixed Gaussian noise, the effectiveness of our proposed methods compared to Perturbation-Optimization approaches combined with a Gibbs sampling algorithm.

VBA approach for image recovery: The goal of VBA is to infer the posterior distribution of a set of parameters given observed data by seeking for a separable approximating distribution which is as close as possible to the true posterior distribution in terms of Kullback-Leibler divergence. In many instances, expressions of these approximate distributions are analytically intractable, especially when it is not possible to directly calculate expectation of the log joint likelihood with respect to the factorized approximate distribution. This is the case for example of non-Gaussian and signal dependent noise models. By resorting to majorization strategies based on half-quadratic tools, we have constructed a lower bound on the Kullback-Leibler divergence that we want to minimize. Approximate distribution has been derived by making this lower-bound as close as possible to the Kullback-Leibler divergence using a coordinate-ascent iterative algorithm. In particular, the approximate distribution of the target image is a Gaussian one, whose covariance matrix has been approximated either with a diagonal one or with a Monte Carlo estimate using a Perturbation-Optimization simulation algorithm. We have illustrated the usefulness of our algorithm via experiments in the context of image restoration when the noise follows a mixed Poisson-Gaussian model.

- Results have shown that, in most studied situations, the diagonal approximation of the covariance matrix gives satisfactory results. The Monte Carlo estimate ensures a better approximation of the covariance matrix but at the expense of a higher computational time. One can decrease the cost of each iteration by using a higher number of cores to perform the Perturbation-Optimization simulation tasks in a parallel manner
- We have compared the restoration results of our VBA algorithm to those obtained by optimization algorithms computing the MAP estimate where the regularization parameter is either estimated using a discrepancy principle or fixed manually according to the ground truth. Results have shown that for most tested scenarii, discrepancy based approaches perform relatively poorly compared with the other methods, especially in the case of low count images. The VBA method and the MAP approach with the best regularization parameter achieve competitive results while the proposed algorithm has the advantage to automatically tuning the regular-

ization parameter from observations.

- When applied to image recovery with pure Poisson noise, the proposed VBA approach has yielded the best restoration results for moderate count images compared to the state of the art-methods but at the price of a larger computational cost.

§ 2 Perspectives

As a future work, this PhD opens several perspectives both from a methodological and application viewpoints. In this section, we give different ongoing works and promising ideas that could complete or improve the proposed methods in this thesis.

2.1 Short-term extensions

Comparisons of VBA and MCMC approaches: In this thesis, we have evaluated the proposed VBA and MCMC algorithms through signal and image recovery problems. An ongoing work is to compare the performance of VBA and MCMC approaches, for all the experiments throughout this thesis. On the one hand, it is well known that approximating distribution within the VBA framework is tractable when the target distribution is Gaussian conditioned to some hidden variables. Hence, VBA can be efficiently applied to the restoration of the sparse signal with $\mathcal{ST}$ prior presented in Chapter 3 by expressing $\mathcal{ST}$ as a scale mixture of Gaussian with an inverse Gamma mixing distribution [Gharsalli et al., 2012]. Similarly, multichannel image restoration problem presented in Chapter 4 can also be performed using VBA approach since the GMEP distribution can be written as a scale mixture of Gaussian where the mixing law is related to alpha stable distributions [Gómez-S-M. et al., 2008]. On the other hand, Poisson-Gaussian image restoration with total variation prior could be accomplished using proximal type MCMC sampling algorithm proposed in [Pereyra, 2016].

Other applications of the proposed methods: The proposed algorithms in this thesis have great generality and can be applied to a wide class of problems. The advantage of all these methods relies on their ability to jointly estimate the target signal of interest together with the unknown regularization and acquisition parameters. For instance, signal/image blind deconvolution can be further investigated using MCMC and VBA algorithms proposed in Chapters 5 and 3 respectively by assigning a prior probability to the blur kernel to be estimated [Babacan et al., 2012] and simulating from (respectively approximating) the resulting

conditional distribution. Moreover, the Gibbs sampling algorithms with auxiliary variables proposed in Chapter 4, can also be applied to super-resolution problems involving tight/sparse decimation matrices [Orieux et al., 2012], image processing using redundant tight frames [Pustelnik et al., 2016], image segmentation [Ayasso and Mohammad-Djafari, 2010], image reconstruction etc.

Advanced VBA approaches: The objective of VBA approaches is to find a tractable probability density function achieving the minimal Kullback-Leibler divergence according to the true density which reduces to solving a convex infinite-dimensional optimization problem. The proposed VBA algorithm in Chapter 5 computes the solution of this problem using a coordinate ascent iterative algorithm which updates at each iteration one component of the separable distribution while holding the remaining ones fixed. However, other optimization algorithms have been recently proposed by extending gradient-type iterative algorithms into the space of probability densities involved in the VBA methodology [Fraysse and Rodet, 2014; Zheng et al., 2015]. In future work, we would like to improve the algorithm proposed in Chapter 5 by using these advanced iterative algorithms which may speed up their convergence and bring significant computational savings when tackling large dimensional problems. Furthermore, the majorization technique that we have proposed to address the problem of the intricate form of the prior law could be further applied to VBA methods based on Bethe approaches [Yedidia et al., 2005].

Preconditioned MCMC algorithms: The proposed preconditioning strategy in Chapter 3 can be generalized to other sampling algorithms that are based on first order derivative information for differentiable laws. For instance, Hamiltonian Monte Carlo (HMC) are alternatives to Langevin type algorithms that propose samples based on physical interpretation of the target distribution. In addition to the variables of interest, they introduce independent auxiliary variables $\mathbf{p} \in \mathbb{R}^Q$, that follow a zero-mean Gaussian distribution with covariance matrix $\mathbf{M}$. In HMC algorithms, the minus logarithm of the joint posterior distribution $\mathcal{J}(\mathbf{x}, \mathbf{p})$ is interpreted as the total energy of the system with position variable $\mathbf{x}$ and momentum variables $\mathbf{p}$ [Neal, 2011]. This sampling algorithm is then derived using the analogy with the kinetic energy conservation in physics and the leapfrog discretization method [Leimkuhler and Reich, 2004; Neal, 2011]. Trajectories incorporating information from the target distribution can be simulated by choosing a position dependent mass matrix i.e, $\mathbf{M} = \mathbf{M}(\mathbf{x})$ that takes into account the geometry of the target density [Betancourt, 2013]. Standard choices include the Hessian matrix proposed in [Zhang and Sutton, 2011] and

the Fisher matrix proposed in [Girolami and Calderhead, 2011]. Alternatively, we propose to use the curvature matrix constructed by Majorize-Minimize strategy as the mass matrix in HMC algorithms.

2.2 Future works and open problems

For tighter proposals and cheaper metrics calculations: The computational cost and the accuracy of the proposed algorithm in Chapter 3 depend on the curvature matrix of the Gaussian proposal density. In this context, there are two possible directions for future works. From the theoretical viewpoint, an interesting perspective is to see how to build new quadratic tangent majorant functions using scale matrices that provide tighter local approximations of the target law curvature than the diagonal matrix, while being simple to manipulate. Some cheap forms may be investigated such as, block diagonal, sparse, toeplitz matrices etc. In the second direction, rather than fixing the same proposal in the whole algorithm, methods for automatic selection of the curvature matrix form based on the local curvature of the target law in each iteration, need to be developed.

Heavy tailed MH proposals: MH algorithms with Gaussian proposals based on local moves such as the Gaussian Random Walk and Langevin-type algorithms often exhibit poor performance on certain types of target distributions. For instance, their geometric ergodicity has been only demonstrated for super-exponential laws [Jarner and Hansen, 2000; Mengersen and Tweedie, 1996; Roberts and Stramer, 2002; Schreck et al., 2016]. Hence, when the target distribution are heavy-tailed, these algorithms show some difficulties to reach convergence at finite time, and/or to explore efficiently the tails of the target distribution. In fact, in such challenging frameworks, a Gaussian proposal is not a judicious choice due to its short tails. To overcome these problems, one may resort instead to proposals that have heavier tails than Gaussian. This approach has been already investigated in [Jarner and Roberts, 2007]. In future work, one appealing idea would be the selection of heavy tailed proposals based on non-quadratic Majorize-Minimize approaches.

More sophisticated prior models: The choice of more sophisticated prior distributions could greatly improve the restoration quality in practice. For example, it would be preferable to design models that enforce not only the spectral correlation of multispectral images but also other important structural properties such as spatial image correlation, dependencies between multiscale subbands with same orientation etc. One goal could be to extend the proposed algorithms in Chapter 4 to manage the different correlations presented in the same prior distribution,

i.e., dissociate the inter-spectral, inter-scale and the spatial correlation by including additional auxiliary variables.

Extension to non-Gaussian models: The Gibbs sampling algorithms proposed in Chapter 4 address sampling problems from high-dimensional distribution by adding auxiliary variables provided that either the data fidelity term or the prior is Gaussian conditioned to some hidden variables. A challenging task could be to enlarge the field of applications of the proposed algorithms to deal with problems involving non-Gaussian models.

- Appendix A -

PCGS algorithm in the case of a scale mixture of Gaussian noise

- **Parent Gibbs sampler:** Each iteration t of the Gibbs sampling algorithm is composed of 4 sampling steps:

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

- **Marginalization:** Rather than sampling only a variable at each sampling step of the Gibbs iteration, some other variables may be sampled along with instead of being conditioned upon without affecting the convergence eof the Gibbs algorithm to the desired distribution [Van Dyk and Park, 2008]. For instance, we can sample $\boldsymbol{\sigma}$ jointly with $\mathbf{v}$ in the third step, for example by first sampling from $\mathcal{P}_{\boldsymbol{\sigma}|\mathbf{x},\boldsymbol{\Theta},\mathbf{z}}$ and then from $\mathcal{P}_{\mathbf{v}|\mathbf{x},\boldsymbol{\sigma},\boldsymbol{\Theta},\mathbf{z}}$. Similarly, we can sample $\boldsymbol{\Theta}$ jointly with $\mathbf{v}$ in the fourth step, by first sampling from $\mathcal{P}_{\boldsymbol{\Theta}|\mathbf{x},\boldsymbol{\sigma},\mathbf{z}}$ and then from $\mathcal{P}_{\mathbf{v}|\mathbf{x},\boldsymbol{\sigma},\boldsymbol{\Theta},\mathbf{z}}$. In each step, we sample the variables of interest conditioning to the most recently sampled value of $\mathbf{v}$ before the current step (e.g., in Step 2, $\mathbf{x}$ is sampled conditioning to the value of $\mathbf{v}$ sampled in Step 1). Note that, at each iteration of the Gibbs sampler, the output for each variable is the most recently sampled, that is $\mathbf{x}$ is sampled in Step 2, $\boldsymbol{\sigma}$ in Step 3 and finally $\boldsymbol{\Theta}$ and $\mathbf{v}$ in Step 4.

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

The last Gibbs sampler may be inefficient since $\mathbf{v}$ is drawn many times at each iteration. However, we can not remove arbitrarily redundant samplers of $\mathbf{v}$. For instance, samples of $\mathbf{v}$ in Step 1 and 4 cannot be dropped from the respective sampling distribution since $\mathbf{v}^{(t+\frac{1}{3})}$ is conditioned upon in Step 2 and $\mathbf{v}^{(t+1)}$ belongs to the output of the Gibbs iteration. We can remove instead unused intermediate values namely $\mathbf{v}^{(t+\frac{2}{3})}$ since it is never conditioned upon and does not belong to the output of the Gibbs iteration. Such procedure is called trimming. Note that it remains possible to permute some steps of this Gibbs sampler without altering the convergence of the algorithm to the desired distribution. Hence, it is reasonable to use a good sampling order such that trimming can be performed to a maximum extent.

- **Permutation:** In the following permuted Gibbs iteration, the redundant samples of $\mathbf{v}$ in Step 1 and 2 give intermediate variables. Thus, they can be removed from the algorithm.

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

- **Trimming:** This step means removing the redundant variables from the last Gibbs algorithm. The resulted Gibbs sampler has the same stationary distribution as the parent Gibbs sampler.

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

- Appendix B -

Proof of Proposition 2.1

Let $i \in \{1, \dots, N\}$. Let us define $g_i : \mathbb{R} \to \mathbb{R}$ such that

$$(\forall v \in \mathbb{R}) \quad g_i(v) = \frac{v^2}{2} - \frac{\phi_i(v; z_i)}{\mu_i(z_i)}. \tag{B.1}$$

According to Assumption 2.1, g_i is convex, proper and lower semi-continuous (lsc). Its conjugate function [Bauschke and Combettes, 2011, Chapter 13] reads:

$$\begin{aligned}
(\forall w \in \mathbb{R}) \quad g_i^*(w) &= \sup_{v \in \mathbb{R}} \left(vw - g_i(v) \right) && \text{(B.2)} \\
&= \sup_{v \in \mathbb{R}} \left(vw + \frac{\phi_i(v; z_i)}{\mu_i(z_i)} - \frac{v^2}{2} \right) && \text{(B.3)} \\
&= \sup_{v \in \mathbb{R}} \left(-\frac{1}{2}(v - w)^2 + \frac{\phi_i(v; z_i)}{\mu_i(z_i)} \right) + \frac{w^2}{2}. && \text{(B.4)}
\end{aligned}$$

According to Definition (5.8),

$$(\forall w \in \mathbb{R}) \quad g_i^*(w) = \varsigma_i(w; z_i) + \frac{w^2}{2}. \tag{B.5}$$

The conjugate of g_i^* is

$$\begin{aligned}
(\forall v \in \mathbb{R}) \quad g_i^{**}(v) &= \sup_{w \in \mathbb{R}} \left(vw - g_i^*(w) \right) && \text{(B.6)} \\
&= \sup_{w \in \mathbb{R}} \left(vw - \frac{w^2}{2} - \varsigma_i(w; z_i) \right) \\
&= \sup_{w \in \mathbb{R}} \left(-\frac{1}{2}(v - w)^2 - \varsigma_i(w; z_i) \right) + \frac{v^2}{2} \\
&= - \inf_{w \in \mathbb{R}} \left(\frac{1}{2}(v - w)^2 + \varsigma_i(w; z_i) \right) + \frac{v^2}{2}. && \text{(B.7)}
\end{aligned}$$

Since g_i is convex, proper and lsc [Bauschke and Combettes, 2011, Theorem 13.32], $g_i = g_i^{**}$ so that

$$(\forall v \in \mathbb{R}) \quad - \frac{\phi_i(v; z_i)}{\mu_i(z_i)} = - \inf_{w \in \mathbb{R}} \left(\frac{1}{2}(v - w)^2 + \varsigma_i(w; z_i) \right) \tag{B.8}$$

which is equivalent to

$$(\forall v \in \mathbb{R}) \quad \phi_i(v; z_i) = \mu_i(z_i) \inf_{w \in \mathbb{R}} \left(\frac{1}{2}(v-w)^2 + \varsigma_i(w; z_i) \right), \tag{B.9}$$

so that (5.9) holds.

For every $v \in \mathbb{R}$, let

$$\widehat{w}_i(v) = g_i'(v). \tag{B.10}$$

The function g_i being convex, proper and lsc, according to [Bauschke and Combettes, 2011, Corollary 16.24], the above relation can be reexpressed by making use of the subdifferential ∂g_i^* of the convex function g_i^* (see [Bauschke and Combettes, 2011, Chapter 16] for more details). More precisely, (B.10) is equivalent to

$$v \in \partial g_i^*\big(\widehat{w}_i(v)\big). \tag{B.11}$$

According to Fermat's rule [Bauschke and Combettes, 2011, Theorem 16.2], (B.11) is a necessary and sufficient condition for $\hat{w}_i(v)$ to be a minimizer of the convex function $w \mapsto g_i^*(w) - vw$.

This minimizer is unique since $\widehat{w}_i(v)$ is uniquely defined by (B.10). We have therefore established that

$$\widehat{w}_i(v) = \arg\max w \in \mathbb{R}\big(vw - g_i^*(w)\big). \tag{B.12}$$

The definition of g_i in (B.1) shows that (B.10) also reads

$$\widehat{w}_i(v) = v - \frac{1}{\mu_i(z_i)} \phi_i'(v; z_i). \tag{B.13}$$

According to (B.6), it is straightforward that $\widehat{w}_i(v)$ also reaches the infimum in (B.9). Hence the result by using (B.13) and (5.10).

List of Figures

List of Tables

List of Algorithms

Bibliography

Aharon, M., Elad, M., and Bruckstein, A. (2006). An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322. 21

Ahmed, N., Natarajan, T., and Rao, K. R. (1974). Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93. 21

Allain, M., Idier, J., and Goussard, Y. (2006). On global and local convergence of half-quadratic algorithms. *IEEE Trans. Image Process.*, 15(5):1130–1142. xi, 62, 149

Almeida, M. S. and Figueiredo, M. A. (2013). Parameter estimation for blind and non-blind deblurring using residual whiteness measures. *IEEE Trans. Image Process.*, 22(7):2751–2763. 32, 118

Altmann, Y., Ren, X., McCarthy, A., Buller, G. S., and McLaughlin, S. (2016). Lidar waveform-based analysis of depth images constructed using sparse single-photon data. *IEEE Trans. Image Process.*, 25(5):1935–1946. 11, 117

Altmann, Y., Wallace, A., and McLaughlin, S. (2015). Spectral unmixing of multispectral Lidar signals. *IEEE Trans. Signal Process.*, 63(20):5525–5533. 119

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102. 85

Anthoine, S., Aujol, J. F., Boursier, Y., and Mélot, C. (2012). Some proximal methods for Poisson intensity CBCT and PET. *Inverse Probl. Imaging*, 6(4):565–598. 11

Antoniadis, A., Leporini, D., and Pesquet, J.-C. (2002). Wavelet thresholding for some classes of non–Gaussian noise. *Statistica neerlandica*, 56(4):434–453. 63

Archer, G. and Titterington, D. (1995). On some bayesian/regularization methods for image restoration. *IEEE Trans. Image Process.*, 4(7):989–995. 32

Atchadé, Y. F. (2006). An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability*, 8(2):235–254. 55, 58, 64, 67, 68, 72, 105

Aubert, G. and Aujol, J.-F. (2008). A variational approach to removing multiplicative noise. *SIAM J. Appl. Math.*, 68(4):925–946. 10, 117

Ayasso, H. and Mohammad-Djafari, A. (2010). Joint NDT image restoration and segmentation using Gauss–Markov–Potts prior models and variational Bayesian computation. *IEEE Trans. Image Process.*, 19(9):2265–2277. 152

Azzari, L. and Foi, A. (2014). Gaussian-Cauchy mixture modeling for robust signal-dependent noise estimation. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2014)*, pages 5357–5361, Florance, Italy. 10

Babacan, S. D., Molina, R., Do, M. N., and Katsaggelos, A. K. (2012). Bayesian blind deconvolution with general sparse image priors. In *Proc. European Conference on Computer Vision (ECCV 2012)*, pages 341–355, Firenze, Italy. Springer. 151

Babacan, S. D., Molina, R., and Katsaggelos, A. K. (2011). Variational Bayesian super resolution. *IEEE Trans. Image Process.*, 20(4):984–999. 50, 52, 127

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106. 9, 77

Bajić, B., Lindblad, J., and Sladoje, N. (2016). Blind restoration of images degraded with mixed Poisson-Gaussian noise with application in transmission electron microscopy. In *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI 2016)*, pages 123–127, Prague, Czech Republic. 118

Balakrishnan, N. and Nevzorov, V. B. (2004). *A primer on statistical distributions*. John Wiley & Sons. 15

Bardsley, J. M. (2012). MCMC-based image reconstruction with uncertainty quantification. *SIAM Journal on Scientific Computing*, 34(3):A1316–A1332. 77

Bardsley, J. M. and Goldes, J. (2009). Regularization parameter selection methods for ill-posed Poisson maximum likelihood estimation. *Inverse Probl.*, 25(9):095005. 32, 33, 131

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311. 100, 101

Barndorff-Nielsen, O. E. and Kluppelberg, C. (2000). *Complex stochastic systems*. CRC Press. 39

Bartlett, M. S. (1934). The vector representation of a sample. In *Proc. Mathematical Proceedings of the Cambridge Philosophical Society*, volume 30, pages 327–340. Cambridge Univ Press. 19

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York. 159, 160

Bect, J., Blanc-Féraud, L., Aubert, G., and Chambolle, A. (2004). A l1-unified variational framework for image restoration. In *Proc. European Conference on Computer Vision (ECCV 2004)*, pages 1–13, Prague, Czech Republic. 80, 81

Benvenuto, F., La Camera, A., Theys, C., Ferrari, A., Lantéri, H., and Bertero, M. (2008). The study of an iterative method for the reconstruction of images corrupted by Poisson and Gaussian noise. *Inverse Probl.*, 24(3):035016. 9, 118, 128, 129

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147. 29, 30

Bernardo, J. M. and Smith, A. F. (2001). Bayesian theory. 12

Bertero, M., Boccacci, P., Talenti, G., Zanella, R., and Zanni, L. (2010). A discrepancy principle for Poisson data. *Inverse Probl.*, 26(10):105004. 32, 33, 119, 131, 143

Betancourt, M. (2013). A general metric for Riemannian manifold Hamiltonian Monte Carlo. In *Geometric science of information*, pages 327–334. Springer. 152

Bonettini, S. and Prato, M. (2015). New convergence results for the scaled gradient projection method. *Inverse Probl.*, 31(9):095008. 11

Bonettini, S. and Ruggiero, V. (2011). An alternating extragradient method for total variation-based image restoration from Poisson data. *Inverse Probl.*, 27(9):095001. 11

Boubchir, L., Al-Maadeed, S., and Bouridane, A. (2014). Undecimated wavelet-based Bayesian denoising in mixed Poisson-Gaussian noise with application on medical and biological images. In *Proc. Int. Conf. Image Process. Theory, Tools and Applicat. (IPTA 2014)*, pages 1–5, Paris, France. 10

Boulanger, J., Condat, L., Piolot, T., Sengmanivong, L., and Pustelnik, N. (2015). Nonsmooth convex optimization for structured illumination microscopy image reconstruction. Research report, GIPSA-lab. 10

Box, G. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons. 16

Breyer, L. A., Piccioni, M., and Scarlatti, S. (2004). Optimal scaling of MALA for nonlinear regression. *The Annals of Applied Probability*, 14(3):1479–1505. 57

Brooks, S. P. (1996). *Convergence diagnostics for Markov Chain Monte Carlo*. PhD thesis, University of Cambridge. 41

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455. 40

Brooks, S. P. and Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8(4):319–335. 39, 40

Buades, A., Coll, B., and Morel, J.-M. (2005). A non-local algorithm for image denoising. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition (CVPR 2005)*, volume 2, pages 60–65, San Diego, CA, USA. 10

Bui-Thanh, T. and Ghatas, O. (2012). A scaled stochastic Newton algorithm for Markov chain Monte Carlo simulations. Technical report. http://users.ices.utexas.edu/~tanbui/PublishedPapers/SNanalysis.pdf. 58

Cai, J.-F., Chan, R. H., and Nikolova, M. (2010). Fast two-phase image deblurring under impulse noise. *J. Math. Imaging Vision*, 36(1):46–53. 10, 118

Calatroni, L., Cao, C., De Los Reyes, J. C., Schönlieb, C.-B., and Valkonen, T. (2016). Bilevel approaches for learning of variational imaging models. to appear in RADON book series. 11

Candes, E., Demanet, L., Donoho, D., and Ying, L. (2006). Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899. 21

Candes, E. and Donoho, D. L. (2002). Recovering edges in ill-posed inverse problems: Optimality of curvelet frames. *Annals of statistics*, 30(3):784–842. 24

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174. 47

Casella, G., Robert, C. P., and Wells, M. T. (2004). Mixture models, latent variables and partitioned importance sampling. *Statistical Methodology*, 1(1):1–18. 43

Cavicchioli, R., Chaux, C., Blanc-Féraud, L., and Zanni, L. (2013). ML estimation of wavelet regularization hyperparameters in inverse problems. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2013)*, pages 1553–1557, Vancouver, Canada. 80

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problems. *Computational Statistics*, 2:73–82. 32

Chaâri, L., Vincent, T., Forbes, F., Dojat, M., and Ciuciu, P. (2013). Fast joint detection-estimation of evoked brain activity in event-related FMRI using a variational approach. *IEEE Trans. Med. Imaging*, 32(5):821–837. 119

Chakrabarti, A. and Zickler, T. E. (2012). Image restoration with signal-dependent camera noise. Research report. 10, 129

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145. 32

Champagnat, F. and Idier, J. (2004). A connection between half-quadratic criteria and EM algorithms. *IEEE Signal Process. Lett.*, 11(9):709–712. 32, 80, 120

Chang, E. S., Hung, C.-C., Liu, W., and Yina, J. (2016). A denoising algorithm for remote sensing images with impulse noise. In *Proc. IEEE Int. Symp. Geoscience and Remote Sensing (IGARSS 2016)*, pages 2905–2908, Beijing, China. IEEE. 107

Chantas, G., Galatsanos, N., Likas, A., and Saunders, M. (2008). Variational Bayesian image restoration based on a product of t-distributions image prior. *IEEE Trans. Image Process.*, 17(10):1795–1805. 69

Charbonnier, P., Blanc-Féraud, L., Aubert, G., and Barlaud, M. (1994). Two deterministic half-quadratic regularization algorithms for computed imaging. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 2, pages 168–172. IEEE. 32

Charbonnier, P., Blanc-Féraud, L., Aubert, G., and Barlaud, M. (1997). Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Process.*, 6(2):298–311. 62

Chaux, C., Pesquet, J.-C., and Pustelnik, N. (2009). Nested iterative algorithms for convex constrained image recovery problems. *SIAM J. Imaging Sci.*, 2(2):730–762. 11

Chellappa, R. and Chatterjee, S. (1985). Classification of textures using Gaussian Markov random fields. *IEEE Trans. Acoust., Speech, Signal Process.(1985)*, 33(4):959–963. 77

Chen, Z., Babacan, S. D., Molina, R., and Katsaggelos, A. K. (2014). Variational Bayesian methods for multimedia problems. *IEEE Trans. Multimedia*, 16(4):1000–1017. 50, 123

Chierchia, G., Pustelnik, N., Pesquet-Popescu, B., and Pesquet, J.-C. (2014). A nonlocal structure tensor based approach for multicomponent image recovery problems. *IEEE Trans. Image Process.*, 23(12):5531–5544. 138

Chmielewski, M. (1981). Elliptically symmetric distributions: a review and bibliography. *International Statistical Review/Revue Internationale de Statistique*, 49(1):67–74. 19

Chouzenoux, E., Idier, J., and Moussaoui, S. (2011). A majorize–minimize strategy for subspace optimization applied to image restoration. *IEEE Trans. Image Process.*, 20(6):1517–1528. 32

Chouzenoux, E., Jezierska, A., Pesquet, J.-C., and Talbot, H. (2013). A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization. *SIAM Journal on Imaging Sciences*, 6(1):563–591. 63

Chouzenoux, E., Jezierska, A., Pesquet, J.-C., and Talbot, H. (2015). A convex approach for image restoration with exact Poisson-Gaussian likelihood. *SIAM J. Imag. Sci.*, 8(4):2662–2682. 11, 63, 118, 129, 130, 131

Chouzenoux, E., Pesquet, J.-C., and Repetti, A. (2014). Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132. 64

Christensen, O. F., Roberts, G. O., and Rosenthal, J. S. (2005). Scaling limits for the transient phase of local Metropolis–Hastings algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):253–268. 58

Ciuciu, P. (2000). *Méthodes Markoviennes en estimation spectrale non paramétriques. Application en imagerie radar Doppler*. PhD thesis, Université Paris Sud-Paris XI. 80

Ciuciu, P. and Idier, J. (2002). A half-quadratic block-coordinate descent method for spectral estimation. *Signal Processing*, 82(7):941–959. 32, 80

Combettes, P. L. and Pesquet, J.-C. (2007). A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Sel. Topics Signal Process.*, 1(4):564–574. 32

Combettes, P. L. and Pesquet, J.-C. (2010). Proximal splitting methods in signal processing. In Bauschke, H. H., Burachik, R., Combettes, P. L., Elser, V., Luke, D. R., and Wolkowicz, H., editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer-Verlag, New York. 123

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer. 32

Condat, L. (2014). Semi-local total variation for regularization of inverse problems. In *Proc. European Signal Processing Conf. (EUSIPCO 2014)*, pages 1806–1810, Lisbon, Portugal. 138

Costa, F., Batatia, H., Oberlin, T., and Tourneret, J.-Y. (2016). A partially collapsed Gibbs sampler with accelerated convergence for EEG source localization. In *Proc. IEEE Stat. Signal Process. Workshop (SSP 2016)*, pages 1–5, Palma de Mallorca, Spain. IEEE. 91

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904. 40, 41

Damlen, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Series B Stat. Methodol.*, 61(2):331–344. 1, 48, 79

David, H. M. (1997). Auxiliary variable methods for Markov Chain Monte Carlo with applications. *J. Am. Stat. Assoc.*, 93:585–595. 79

Dawid, A. (1970). On the limiting normality of posterior distributions. In *Proc. Mathematical Proceedings of the Cambridge Philosophical Society*, volume 67, pages 625–633. Cambridge Univ Press. 29

Deledalle, C.-A., Vaiter, S., Fadili, J., and Peyré, G. (2014). Stein unbiased gradient estimator of the risk (SUGAR) for multiple parameter selection. *SIAM J. Imaging Sci.*, 7(4):2448–2487. 118

Delp, E. and Mitchell, O. (1979). Image compression using block truncation coding. *IEEE Trans. Commun.*, 27(9):1335–1342. 95

Delpretti, S., Luisier, F., Ramani, S., Blu, T., and Unser, M. (2008). Multiframe sure-let denoising of timelapse fluorescence microscopy images. In

Proc. IEEE Int. Symp. Biomedical Imaging (ISBI 2008), Paris, France. IEEE. 128

Demoment, G. (1989). Image reconstruction and restoration: Overview of common estimation structures and problems. *IEEE Trans. Acous., Speech Signal Process*, 37(12):2024–2036. 118

Dennis Jr, J. E. and Welsch, R. E. (1978). Techniques for nonlinear least squares and robust regression. *Communications in Statistics-Simulation and Computation*, 7(4):345–359. 63

Dong, Y. and Zeng, T. (2013). A convex variational model for restoring blurred images with multiplicative noise. *SIAM J. Imaging Sci.*, 6(3):1598–1625. 10

Doucet, A., Sénécal, S., and Matsui, T. (2005). Space alternating data augmentation: Application to finite mixture of gaussians and speaker recognition. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2005)*, volume 4, pages 708–713. IEEE. 79

Drémeau, A., Herzet, C., and Daudet, L. (2012). Boltzmann machine and mean-field approximation for structured sparse decompositions. *IEEE Trans. Signal Process.*, 60(7):3425–3438. 50

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216 – 222. 79

Dupé, F.-X., Fadili, M. J., and Starck, J.-L. (2009). A proximal iteration for deconvolving Poisson noisy images using sparse representations. *IEEE Trans. Image Process.*, 18(2):310–321. 9

Dupé, F.-X., Fadili, M. J., and Starck, J.-L. (2012). Deconvolution under Poisson noise using exact data fidelity and synthesis or analysis sparsity priors. *Stat. Methodol.*, 9(1):4–18. 11

Durmus, A. and Moulines, E. (2015). Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *arXiv preprint arXiv:1507.05021*. 57

Durmus, A., Moulines, E., and Pereyra, M. (2016). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *arXiv preprint arXiv:1612.07471*. 57

Eldar, Y. C. (2009). Generalized SURE for exponential families: Applications to regularization. *IEEE Trans. Signal Process.*, 57(2):471–481. 118

Esser, E., Zhang, X., and Chan, T. F. (2010). A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046. 32

Fadili, J. and Starck, J.-L. (2009). Curvelets and ridgelets. In *Encyclopedia of Complexity and Systems Science*, pages 1718–1738. Springer. 24

Féron, O., Orieux, F., and Giovannelli, J.-F. (2016). Gradient Scan Gibbs Sampler: an efficient algorithm for high-dimensional Gaussian distributions. *IEEE J. Sel. Topics Signal Process.*, 10(2):343–352. 77

Févotte, C., Cappé, O., and Cemgil, A.-T. (2011). Efficient Markov chain Monte Carlo inference in composite models with space alternating data augmentation. In *Proc. IEEE Statist. Signal Process. Workshop (SSP 2011)*, pages 221–224. IEEE. 79

Fink, D. (1997). A compendium of conjugate priors. *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, page 46. 27, 101

Fitzpatrick, J. M. and Sonka, M. (2000). *Medical Image Processing and Analysis*. SPIE Press. 9

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional brownian motion. *IEEE Trans. Inf. Theory*, 38(2):910–917. 102

Flegal, J. M., Haran, M., and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, pages 250–260. 40

Foi, A., Trimeche, M., Katkovnik, V., and Egiazarian, K. (2008). Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Trans. Image Process.*, 17:1737–1754. 11

Forbes, F. and Fort, G. (2007). Combining Monte Carlo and mean-field-like methods for inference in hidden Markov random fields. *IEEE Trans. Image Process.*, 16(3):824–837. 50

Fraysse, A. and Rodet, T. (2014). A measure-theoretic variational Bayesian algorithm for large dimensional problems. *SIAM Journal on Imaging Sciences*, 7(4):2591–2622. 50, 152

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science. Chapman and Hall/CRC. 2, 33, 38

Ganan, S. and McClure, D. (1985). Bayesian image analysis: An application to single photon emission tomography. *Amer. Statist. Assoc*, pages 12–18. 63

Gauss, C. F. (1995). *Theory of the Combination of Observations Least Subject to Error: Part One, Part Two, Supplement*, volume 11. Siam. 12

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472. 41

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. *Handbook of markov chain monte carlo*, pages 163–174. 40

Geman, D. and Reynolds, . (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(3):367–383. 80

Geman, D. and Yang, C. (1995). Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Process.*, 4(7):932–946. 32, 77, 80, 120

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI(6):721–741. 47, 80, 85

Geman, S. A. and McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. *MIT Industrial Liaison Program report*, 8:5–21. 69

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, 57(6):1317–1339. 35

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–483. 41

Gharsalli, L., Mohammad-Djafari, A., Fraysse, A., and Rodet, T. (2012). Variational Bayesian approximation with scale mixture prior: a comparison between three algorithms. In *Proc. Int. Workshop on Bayesian Inference and Maximun Entropy Methods in Sciences and Engineerin (MaxEnt 2012)*, volume 1353, pages 130–138, IPP, Garching near Munich. 151

Gibbs, A. L. (2000). *Convergence of Markov Chain Monte Carlo algorithms with applications to image restoration*. PhD thesis, University of Toronto. 39

Gilavert, C., Moussaoui, S., and Idier, J. (2015). Efficient Gaussian sampling for solving large-scale inverse problems using MCMC. *IEEE Trans. Signal Process.*, 63(1):70–80. 77, 111

Gilboa, G. and Osher, S. (2008). Nonlocal operators with applications to image processing. *Multiscale Model. Simul.*, 7(3):1005–1028. 138

Gilks, W. R., Best, N., and Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 44(4):455–472. 47

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1999). *Markov Chain Monte Carlo in practice*. Interdisciplinary Statistics. Chapman and Hall/CRC. 33, 38, 47, 80

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348. 35

Giovannelli, J.-F. (2008). Unsupervised Bayesian convex deconvolution based on a field with an explicit partition function. *IEEE Trans. Image Process.*, 17(1):16–26. 79

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(91):123–214. 46, 57, 58, 59, 79, 81, 153

Giryes, R., Elad, M., and Eldar, Y. C. (2011). The projected GSURE for automatic parameter tuning in iterative shrinkage methods. *Appl. Comput. Harmon. Anal.*, 30(3):407–422. 118

Glynn, P. W. and Iglehart, D. L. (1989). Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392. 34

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223. 33

Gómez-S-M., E., Gómez-Villegas, M., and Marín, J. (2006). Sequences of elliptical distributions and mixtures of normal distributions. *Journal of multivariate analysis*, 97(2):295–310. 19

Gómez-S-M., E., Gómez-Villegas, M., and Marín, J. (2008). Multivariate exponential power distributions as mixtures of normal distributions with bayesian applications. *Communications in StatisticsTheory and Methods*, 37(6):972–985. 19, 151

Gribonval, R. (2011). Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Trans. Signal Process.*, 59(5):2405–2410. 15

Gribonval, R. and Machart, P. (2013). Reconciling" priors" &" priors" without prejudice? In *Advances in Neural Information Processing Systems*, pages 2193–2201. Curran Associates, Inc. 15

Grimmer, J. (2010). An introduction to Bayesian inference via variational approximations. *Political Analysis*, 24(4):mpq027. 49

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton university bulletin*, 13(49-52):28. 12

Han, D. and Larson, D. R. (2000). *Frames, bases and group representations*, volume 697. American Mathematical Soc. 24

Hansen, P. C., Kilmer, M. E., and Kjeldsen, R. H. (2006). Exploiting residual information in the parameter choice for discrete ill-posed problems. *BIT Numer. Math.*, 46(1):41–59. 119

Hansen, P. C. and O'Leary, D. P. (1993). The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal on Scientific Computing*, 14(6):1487–1503. 33

Harizanov, S., Pesquet, J.-C., and Steidl, G. (2013). Epigraphical projection for solving least squares Anscombe transformed constrained optimization problems. In *Proc. Int. Conf. Scale Space and Variational Methods Comput. Vision (SSVM 2013)*, pages 125–136, Schloss Seggau, Austria. 11

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109. 2, 3, 38, 42

Healey, G. E. and Kondepudy, R. (1994). Radiometric CCD camera calibration and noise estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(3):267–276. 10

Hebert, T. and Leahy, R. (1989). A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. medical imaging*, 8(2):194–202. 69

Hobert, J. P., Jones, G. L., Presnell, B., and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika*, pages 731–743. 40

Huang, Y.-M., Ng, M. K., and Wen, Y.-W. (2009). A new total variation method for multiplicative noise removal. *SIAM J. Imaging Sci.*, 2(1):20–40. 10

Huber, P. J. (2011). *Robust statistics*. Springer. 63

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37. 59

Hurn, M. (1997). Difficulties in the use of auxiliary variables in Markov chain Monte Carlo methods. *Stat. Comput.*, 7:35–44. 79

Idier, J. (2001). Convex half-quadratic criteria and interacting auxiliary variables for image restoration. *IEEE Trans. Image Process.*, 10(7):1001–1009. 80, 120

Irony, T. Z. and Singpurwalla, N. D. (1997). Non-informative priors do not exist A dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference*, 65(1):159–177. 16

Janesick, J. R. (2007). *Photon transfer*. SPIE press San Jose. 10, 118

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of metropolis algorithms. *Stochastic Processes and their Applications*, 85(2):341 – 361. 39, 67, 68, 153

Jarner, S. F. and Roberts, G. O. (2007). Convergence of Heavy-tailed Monte Carlo Markov Chain Algorithms. *Scandinavian Journal of Statistics*, 34(4):781–815. 153

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford. 30

Jeong, T., Woo, H., and Yun, S. (2013). Frame-based Poisson image restoration using a proximal linearized alternating direction method. *Inverse Probl.*, 29(7):075007. 11

Jezierska, A., Chaux, C., Pesquet, J.-C., Talbot, H., and Engler, G. (2014). An EM approach for time-variant Poisson-Gaussian model parameter estimation. *IEEE Trans. Signal Process.*, 62(1):17–30. 11

Jezierska, A., Chouzenoux, E., Pesquet, J.-C., and Talbot, H. (2012). A primal-dual proximal splitting approach for restoring data corrupted with Poisson-Gaussian noise. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2012)*, pages 1085–1088, Kyoto, Japan. 11, 118

Johnson, A. A. (2009). *Geometric ergodicity of Gibbs samplers*. PhD thesis, university of minnesota. 39

Joksch, H., Raiffa, H., and Schlaifer, R. (1964). Applied statistical decision theory. division of research, graduate school of business administration. 27

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability surveys*, 1(299-320):5–1. 39

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547. 40

Julliand, T., Nozick, V., and Talbot, H. (2015). Image noise and digital image forensics. In *Proc. Digital-Forensics and Watermarking: 14th International Workshop (IWDW 2015)*, pages 3–17. 128

Kai-Tai, F., Kotz, S., and Kai Wang, N. (1990). *Symmetric multivariate and related distributions*. Chapman and Hall. 19

Kai-Tai, F. and Yao-Ting, Z. (1990). *Generalized multivariate analysis*. Science Press. 19

Kail, G., Tourneret, J.-Y., Hlawatsch, F., and Dobigeon, N. (2012). Blind deconvolution of sparse pulse sequences under a minimum distance constraint: A partially collapsed Gibbs sampler method. *IEEE Trans. Signal Process.*, 60(6):2727–2743. 91

Kaipio, J. and Somersalo, E. (2006). *Statistical and computational inverse problems*, volume 160. Springer Science & Business Media. 14

Kamilov, U., Bostan, E., and Unser, M. (2012). Generalized total variation denoising via augmented Lagrangian cycle spinning with Haar wavelets. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2012)*, pages 909–912. 26, 77

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795. 33, 48

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370. 16, 29

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā: The Indian Journal of Statistics, Series A*, 39(4):419–430. 19

Kendall, W. (2004). Geometric ergodicity and perfect simulation. *Electronic Communications in Probability*, 9:140–151. 39

Khelil-Cherif, N. and Benazza-Benyahia, A. (2004). Wavelet-based multivariate approach for multispectral image indexing. In *SPIE Conference on Wavelet Applications in Industrial Processing*. 19

Kittisuwan, P. (2016). Medical image denoising using simple form of MMSE estimation in Poisson–Gaussian noise model. *Int. J. Biomath.*, 9(02):1650020. 11

Kolehmainen, V., Lassas, M., Niinimäki, K., and Siltanen, S. (2012). Sparsity-promoting Bayesian inversion. *Inverse Problems*, 28(2):025005. 77

Kong, A. (1992). A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348. 33, 35

Kwitt, R., Meerwald, P., and Uhl, A. (2009). Color-image watermarking using multivariate power-exponential distribution. In *Proc. IEEE Int. Conf. on Image Process. (ICIP 2009)*, pages 4245–4248. 20

Lange, K. (1990). Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Trans. medical imaging*, 9(4):439–446. 62

Lantéri, H. and Theys, C. (2005). Restoration of astrophysical images - the case of Poisson data with additive Gaussian noise. *EURASIP J. Adv. Signal Process.*, 2005(15):2500–2513. 9, 11, 118

Laplace, P. (1820). *Théorie analytique des probabilités*. V. Courcier. 29

Lasanen, S. (2012). Non-Gaussian statistical inverse problems. *Inverse Problems & Imaging*, 6(2). 77

Le Pennec, E. and Mallat, S. (2005). Sparse geometric image representations with bandelets. *IEEE Trans. Image Process.*, 14(4):423–438. 21

Lefkimmiatis, S., Bourquard, A., and Unser, M. (2012). Hessian-based norm regularization for image restoration with biomedical applications. *IEEE Trans. Image Process.*, 21(3):983–995. 138

Leimkuhler, B. and Reich, S. (2004). *Simulating Hamiltonian dynamics*, volume 14. Cambridge University Press. 152

Li, J., Shen, Z., Jin, R., and Zhang, X. (2015). A reweighted ℓ_2 method for image restoration with Poisson and mixed Poisson-Gaussian noise. *Inverse Probl. Imaging*, 9(3):875–894. 11, 118, 129

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. In *Proc. Berkeley Symp. on Math. Stat. and Prob*, volume 1, pages 453–468, California, USA. 29

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New-York. 33, 38

Liu, X., Tanaka, M., and Okutomi, M. (2014). Practical signal-dependent noise parameter estimation from a single noisy image. *IEEE Trans. Image Process.*, 23(10):4361–4371. 10

Luisier, F., Blu, T., and Unser, M. (2011). Image denoising in mixed Poisson-Gaussian noise. *IEEE Trans. Image Process.*, 20(3):696–708. 11, 119

Mäkitalo, M. and Foi, A. (2011). Optimal inversion of the Anscombe transformation in low-count Poisson image denoising. *IEEE Trans. Image Process.*, 20(1):99–109. 121, 163

Mäkitalo, M. and Foi, A. (2012). Poisson-Gaussian denoising using the exact unbiased inverse of the generalized Anscombe transformation. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2012)*, pages 1081–1084, Kyoto, Japan. 11, 118, 119, 129

Mäkitalo, M. and Foi, A. (2013). Optimal inversion of the generalized Anscombe transformation for Poisson-Gaussian noise. *IEEE Trans. Image Process.*, 22(1):91–103. 11, 121, 129, 163

Mäkitalo, M. and Foi, A. (2014). Noise parameter mismatch in variance stabilization, with an application to Poisson–Gaussian noise estimation. *IEEE Trans. Image Process.*, 23(12):5348–5359. 11

Mallat, S. (1999). *A wavelet tour of signal processing*. Academic press. 21, 23

Marnissi, Y., Benazza-Benyahia, A., Chouzenoux, E., and Pesquet, J.-C. (2013). Generalized multivariate exponential power prior for wavelet-based multichannel image restoration. In *Proc. IEEE Int. Conf. on Image Process. (ICIP 2013)*, pages 2402–2406. 20, 25, 96, 100

Marnissi, Y., Chouzenoux, E., Benazza-Benyahia, A., Pesquet, J.-C., and Duval, L. (2015). Reconstruction de signaux parcimonieux à l'aide d'un algorithme rapide d'échantillonnage stochastique. In *Actes du 25e colloque GRETSI*, Lyon, France. 81

Marnissi, Y., Chouzenoux, E., Pesquet, J.-C., and Benazza-Benyahia, A. (2016a). An auxiliary variable method for langevin based MCMC algorithms. In *Proc. IEEE Stat. Signal Process. Workshop (SSP 2016)*, pages 297–301, Palma de Mallorca, Spain. 120

Marnissi, Y., Zheng, Y., and Pesquet, J.-C. (2016b). Fast variational Bayesian signal recovery in the presence of Poisson-Gaussian noise. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2016)*, pages 3964–3968, Shanghai, China. 129

Martin, J., Wilcox, C. L., Burstedde, C., and Ghattas, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM J. Sci. Comput.*, 34(3):1460–1487. 46, 58

McGrory, C. A. and Titterington, D. (2009). Variational Bayesian analysis for hidden Markov models. *Australian & New Zealand Journal of Statistics*, 51(2):227–244. 50

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121. 39, 40, 43, 153

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092. 3, 42

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media. 39

Mira, A. and Tierney, L. (1997). On the use of auxiliary variables in Markov chain Monte Carlo sampling. Technical report. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7814. 79

Molina, R., Katsaggelos, A. K., and Mateos, J. (1999). Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Trans. Image Process.*, 8(2):231–246. 32

Moser, S. M. (2012). Capacity results of an optical intensity channel with input-dependent Gaussian noise. *IEEE Trans. Inf. Theory*, 58(1):207–223. 10, 117

Murphy, J. and Godsill, S. (2011). Joint Bayesian removal of impulse and background noise. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2011)*, pages 261–264, Prague, Czech Republic. 119

Murphy, K. P. (2007). Conjugate Bayesian analysis of the Gaussian distribution. *def*, 1(2σ2):16. 100

Murtagh, F., Starck, J.-L., and Bijaoui, A. (1995). Image restoration with noise suppression using a multiresolution support. *Astron. Astrophys. Suppl.*, 112:179–189. 9, 129

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162. 152

Nichols, T. E., Qi, J., Asma, E., and Leahy, R. M. (2002). Spatiotemporal reconstruction of list-mode PET data. *IEEE Trans. Med. Imag.*, 21(4):396–404. 128

Nikolova, M. and Ng, M.-K. (2005). Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM J. Sci. Comput.*, 27:937–966. 32, 80, 120

Oliveira, J. P., Bioucas-Dias, J. M., and Figueiredo, M. (2009). Adaptive total variation image deblurring: a majorization–minimization approach. *Signal Processing*, 89(9):1683–1693. 9

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609. 21

Orieux, F., Féron, O., and Giovannelli, J.-F. (2012). Sampling high-dimensional Gaussian distributions for general linear inverse problems. *IEEE Signal Process. Lett.*, 19(5):251–254. 72, 77, 127, 148, 152

Papandreou, G. and Yuille, A. L. (2010). Gaussian sampling by local perturbations. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Proc. Neural Inform. Process. Syst. Conf. (NIPS 2010)*, pages 1858–1866, Vancouver, Canada. 77, 127

Parisi, G. (1988). *Statistical Field Theory*. Addison Wesley, Redwood City, CA. 33, 48

Park, T. and van Dyk, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics*, 18(2):283–305. 91

Parker, A.and Fox, C. (2012). Sampling Gaussian distributions in Krylov spaces with conjugate gradients. *SIAM Journal on Scientific Computing*, 34(3):B312–B334. 77

Parker, R. L. (1994). *Geophysical inverse theory*. Princeton university press. 9

Parrilli, S., Poderico, M., Angelino, C. V., and Verdoliva, L. (2012). A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage. *IEEE Trans. Geosci. Remote Sens*, 50(2):606–616. 10

Patel, J. K. and Read, C. B. (1996). *Handbook of the normal distribution*, volume 150. CRC Press. 17

Pereyra, M. (2016). Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760. 151

Pereyra, M., Bioucas-Dias, J., and Figueiredo, M. A. T. (2015). Maximum-a-posteriori estimation with unknown regularisation parameters. In *Proc. Eur. Sig. Proc. Conf. (EUSIPCO 2015)*, pages 2244–2248, Nice, France. 23, 28, 32, 119

Pereyra, M., Schniter, P., Chouzenoux, E., Pesquet, J.-C., Tourneret, J.-Y., Hero, A. O., and McLaughlin, S. (2016). A survey of stochastic simulation and optimization methods in signal processing. *IEEE J. Sel. Topics Signal Process.*, 10(2):224–241. 49

Pesquet, J.-C., Benazza-Benyahia, A., and Chaux, C. (2009). A SURE approach for digital signal/image deconvolution problems. *IEEE Trans. Signal Process.*, 57(12):4616–4632. 118

Petropulu, A. P., Pesquet, J. C., and Yang, X. (2000). Power-law shot noise and its relationship to long-memory α-stable processes. *IEEE Trans. Signal Process.*, 48(7):1883–1892. 11

Pillai, N. S., Stuart, A. M., and Thierry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Probab.*, 22(6):2165–2616. 46, 57

Pizurica, A. and Philips, W. (2006). Estimating the probability of the presence of a signal of interest in multiresolution single-and multiband image denoising. *IEEE Trans. Image Process.*, 15(3):654–665. 19

Pizurica, A., Zlokolica, V., and Philips, W. (2004). Noise reduction in video sequences using wavelet-domain and temporal filtering. In *Proc. Photonics Technologies for Robotics, Automation, and Manufacturing*, pages 48–59. International Society for Optics and Photonics. 9

Portilla, J., Strela, V., Wainwright, M. J., and Simoncelli, E. P. (2003). Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. Image Process.*, 12(11):1338–1351. 25

Provencher, S. W. (1979). Inverse problems in polymer characterization: direct analysis of polydispersity with photon correlation spectroscopy. *Die Makromolekulare Chemie*, 180(1):201–209. 9

Pustelnik, N., Benazza-Benhayia, A., Zheng, Y., and Pesquet, J.-C. (2016). Wavelet-based image deconvolution and reconstruction. *Wiley Encyclopedia of Electrical and Electronics Engineering*. 24, 25, 49, 119, 152

Pustelnik, N., Chaux, C., and Pesquet, J.-C. (2011). Parallel proximal algorithm for image restoration using hybrid regularization. *IEEE Trans. Image Process.*, 20(9):2450–2462. 137, 142

Quang, P. B., Musso, C., and Le Gland, F. (2012). Multidimensional Laplace formulas for nonlinear Bayesian estimation. In *Proc. European Signal Processing Conference (EUSIPCO 2012)*, pages 1890–1894. IEEE. 35

Ramani, S., Blu, T., and Unser, M. (2008). Monte-Carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Trans. Image Process.*, 17(9):1540–1554. 118

Repetti, A. (2015). *Algorithmes d'optimisation en grande dimension: applications à la résolution de problèmes inverses*. PhD thesis, Université Paris-Est Marne la Vallée. 121, 163

Repetti, A., Chouzenoux, E., and Pesquet, J.-C. (2012). A penalized weighted least squares approach for restoring data corrupted with signal-dependent noise. In *Proc. Eur. Sig. Proc. Conf. (EUSIPCO 2012)*, pages 1553–1557, Bucharest, Roumania. 63, 118, 129, 131

Repetti, A., Pham, M. Q., Duval, L., Chouzenoux, E., and Pesquet, J.-C. (2015). Euclid in a Taxicab: Sparse Blind Deconvolution with Smoothed ℓ_1-ℓ_2 Regularization. *IEEE Signal Process. Lett.*, 22(5):539–543. 69

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media. 12, 14, 30, 34

Robert, C.and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media. 47, 79

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling or random walk Metropolis algorithms. *Ann. Appl. Probab.*, pages 110–120. 39, 43, 55

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Series B Stat. Methodol.*, 60(1):255–268. 46, 57

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367. 45, 55, 57

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367. 46

Roberts, G. O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357. 56, 58, 153

Roberts, G. O. and Tweedie, L. R. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363. 39, 46, 57, 67

Roberts, T. D. and Kingsbury, N. G. (2014). Bayesian denoising/deblurring of Poisson-Gaussian corrupted data using complex wavelets. In *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI 2014)*, pages x–x+2, Beijing, China. 11, 118

Rond, A., Giryes, R., and Elad, M. (2016). Poisson inverse problems by the Plug-and-Play scheme . *J. Vis. Commun. Image Represent.*, 41:96 – 108. 137, 142

Rubinstein, R. Y. and Kroese, D. P. (2013). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media. 35

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268. 21

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338. 76

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press. 77

Salmon, J., Harmany, Z., Deledalle, C.-A., and Willett, R. (2014). Poisson noise reduction with non-local PCA. *J. Math. Imaging Vis.*, 48(2):279–294. 11, 117

Schreck, A., Fort, G., Le Corff, S., and Moulines, E. (2016). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *IEEE J. Sel. Topics Signal Process.*, 10(2):366–375. 67, 153

Seeger, M. and Bouchard, G. (2012). Fast variational Bayesian inference for non-conjugate matrix factorization models. In *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS 2012)*, pages 1012–1018, La Palma, Canary Islands. 120

Setzer, S., Steidl, G., and Teuber, T. (2010). Deblurring Poissonian images by split Bregman techniques. *J. Vis. Comm. Image Repr.*,, 21(3):193–199. 11

Shepp, L. A. and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imaging*, 1(2):113–122. 10

Sherlock, C., Fearnhead, P., and Roberts, G. O. (2010). The random walk Metropolis: linking theory and practice through a case study. *Statistical Science*, pages 172–190. 41

Šmídl, V. and Quinn, A. (2005). *The Variational Bayes Method in Signal Processing*. Springer-Verlag, New York, USA. 3, 49, 50

Smith, B. J. (2005). Bayesian output analysis program (BOA) version 1.1 user's manual. *Dept. of Biostatistics, Univ. of Iowa, College of Public Health, http://www. public-health. uiowa. edu/boa*. 41

Snyder, D. L., Hammoud, A. M., and White, R. L. (1993). Image recovery from data acquired with a charge-coupled-device camera. *J. Opt. Soc. Am. A*, 10(5):1014–1023. 118, 128

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192. 17

Starck, J.-L. and Murtagh, F. (1998). Automatic noise estimation from the multiresolution support. *Publ. Astron. Soc. Pac.*, 110(744):193–199. 129

Stuart, M. A., Voss, J., and Wiberg, P. (2004). Conditional path sampling of SDEs and the Langevin MCMC method. *Commu. Math. Sci.*, 2(4):685–697. 46, 58

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540. 79

Thompson, A. M., Brown, J. C., Kay, J. W., and Titterington, D. M. (1991). A study of methods of choosing the smoothing parameter in image restoration by regularization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):326–339. 32

Tian, H., Fowler, B., and Gamal, E. A. (2001). Analysis of temporal noise in CMOS photodiode active pixel sensor. *IEEE J. Solid-State Circuits*, 36(1):92–101. 10

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728. 38, 39

Tikhonov, A. N. (1963). Regularization of incorrectly posed problems. SOVIET MATHEMATICS DOKLADY. 16

Tramel, E. W., Drémeau, A., and Krzakala, F. (2016). Approximate message passing with restricted Boltzmann machine priors. *J. Stat. Mech.: Theory and Experiment*, 2016(7):073401. 50

Vacar, C., Giovannelli, J.-F., and Berthoumieu, Y. (2011). Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP 2011)*, pages 3964–3967, Prague, Czech Republic. 46, 58, 59

Van Dyk, D. A. and Meng, X.-L. (2012). The art of data augmentation. *Journal of Computational and Graphical Statistics*. 75, 79

Van Dyk, D. A. and Park, T. (2008). Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796. 90, 91, 155

Veksler, O. (1999). *Efficient graph-based energy minimization methods in computer vision*. PhD thesis, Cornell University. 63

Velayudhan, D. and Paul, S. (2016). Two-phase approach for recovering images corrupted by Gaussian-plus-impulse noise. In *Proc. IEEE Int. Conf. Inventive Computation Technologies (ICICT 2016)*, volume 2, pages 1–7, Tamilnadu, India. IEEE. 107

Wainwright, M. J. and Simoncelli, E. P. (1999). Scale mixtures of Gaussians and the statistics of natural images. In *Proc. Neural Inform. Process. Syst. Conf. (NIPS 1999)*, pages 855–861. Citeseer. 19

Walden, A. and Hosken, J. (1986). The nature of the non-Gaussianity of primary reflection coefficients and its significance for deconvolution. *Geophysical Prospecting*, 34(7):1038–1066. 69

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612. 131

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648. 19, 85

Woods, J. W. (2013). *Subband image coding*, volume 115. Springer Science & Business Media. 17

Xiao, Y., Zeng, T., Yu, J., and Ng, M. K. (2011). Restoration of images corrupted by mixed gaussian-impulse noise via $\ell_1 - \ell_0$ minimization. *Pattern Recogn.*, 44(8):1708–1720. 10

Yan, M. (2013). Restoration of images corrupted by impulse noise and mixed gaussian impulse noise using blind inpainting. *SIAM J. Imaging Sci.*, 6(3):1227–1245. 10, 118

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory*, 51(7):2282–2312. 152

Ying, W., Jiang, Y., and Fan, C. (2012). A blind receiver with multiple antennas in impulsive noise with Gaussian mixtures through MCMC approaches. In *Proc. IEEE Int. Conf. Com. Technol. (ICCT 2012)*, pages 961–965, Chengdu, China. 119

Zanni, L., Benfenati, A., Bertero, M., and Ruggiero, V. (2015). Numerical methods for parameter estimation in Poisson data inversion. *J. Math. Imaging Vis.*, 52(3):397–413. 119, 131

Zhang, Y. and Sutton, C. A. (2011). Quasi-Newton methods for Markov chain Monte Carlo. In *Proc. Neural Information Processing Systems (NIPS 2011)*, number 24, pages 2393–2401, Granada, Spain. 46, 58, 152

Zheng, Y., Fraysse, A., and Rodet, T. (2015). Efficient variational Bayesian approximation method based on subspace optimization. *IEEE Trans. Image Process.*, 24(2):681–693. 50, 152

Zibulevsky, M. and Elad, M. (2010). ℓ_1– ℓ_2 optimization in signal and image processing. *IEEE Signal Process. Mag.*, 27(3):76–88. 62

Zozor, S. and Vignat, C. (2010). Some results on the denoising problem in the elliptically distributed context. *IEEE Trans. Signal Process.*, 58(1):134–150. 19, 20