Efficient unsupervised variational Bayesian image reconstruction using a sparse gradient prior

Yuling Zheng, Aurélia Fraysse, Thomas Rodet

Version of Record: https://www.sciencedirect.com/science/article/pii/S0925231219308033
Manuscript_36187a4891db7c531e507bbaa1336b6c

Efficient Unsupervised Variational Bayesian Image Reconstruction Using a Sparse Gradient Prior

Yuling Zheng[*], Aurélia Fraysse[†], Thomas Rodet[‡]

March 27, 2019

Abstract

In this paper we present an efficient unsupervised Bayesian approach and a prior distribution adapted to piecewise regular images. This approach is based on a hierarchical prior distribution promoting sparsity on image gradients. It is fully automatic since hyperparameters are estimated jointly with the image of interest. The estimation of all unknowns is performed efficiently thanks to a fast variational Bayesian approximation method. We highlight the good performance of the proposed approach through comparisons with state of the art approaches on an application to a diffraction tomographic problem.

1 Introduction

Efficient reconstruction methods for inverse problems is a topic of great interest in image processing society. Lots of such problems are ill-posed and may not have an unique solution [12]. To circumvent this issue, existing

[*]IBM Research China. Building 28, Zhongguancun Software Park, 8 Dongbeiwang Western Road, Haidian District, Beijing, 11 100193, CN (email: ylzh@cn.ibm.com)

[†]Laboratory of signals and systems (CNRS-Supélec-University of Paris-Sud). 3, Rue Joliot-Curie, 91190 Gif-sur-Yvette, France. (email: aurelia.fraysse@l2s.centralesupelec.fr)

[‡]Laboratory of Systems and Applications of Information and Energy Technologies (École Normale Supérieure de Cachan). 61, Avenue du Président Wilson, 94230 Cachan, France. (email: trodet@satie.ens-cachan.fr)

estimation approaches are generally based on the introduction of additional prior information on the unknown object in order to promote solutions with some specific properties, e.g. regularity or sparsity. However, the quality of the acquired estimation depends closely on hyperparameters (also known as regularization parameters) which control the compromise between fidelity to observed data and fidelity to this prior information. Several deterministic methods [24, 25] exist in order to tune these parameters in general. Nevertheless, such a way of choosing hyperparameters is generally computationally demanding. Another possibility is to consider statistical strategies given by the Bayesian framework where hyperparameters are jointly estimated with the unknown object by assigning hyperpriors to them [33, 26, 31], which leads to fully automatic approaches, known also as unsupervised approaches in the literature.

In this context, estimations of unknown object and hyperparameters are commonly obtained from a joint posterior distribution by computing a *maximum a posteriori* (MAP) estimate or a posterior mean (PM) estimate. Nevertheless, in general, both estimators are hardly tractable as the computation of the MAP estimate involves usually a non convex optimization problem and the computation of the PM an intractable integral. Numerical tools such as Markov Chain Monte Carlo (MCMC) techniques [37] are then usually used. However, MCMC are known to be computationally expensive in large dimensional problems. Some recent techniques based on improved versions of MCMC have been proposed recently. These Monte Carlo methods have a better convergence in term of computation time, see [36], but their application remains limited in large dimensional cases. Recently, neural network based methods have drawn great attention in image processing. They are especially efficient for tasks such as image classification, object detection, etc. However, for image reconstruction problems, the applications are generally limited to denoising [50] or debluring [28, 45] due to the difficulty in accumulating enough training data. In this case, even recent Bayesian learning methods of [42, 16] remain limited in their applications. In this paper, to have more efficient approaches, we resort to analytical approximation ones given by the variational Bayesian approximation methods which provide a tractable approximation, generally separable, to the true posterior distribution. The optimal approximation is determined by minimizing a measure of dissimilarity, namely the Kullback-Leibler divergence, between the approximate distribution and the original one. This problem can be mathematically formulated as a functional optimization one in the space of separable prob-

ability density functions. There exists an analytical solution underlying the classical variational Bayesian approximation approach [44]. Nevertheless, as stated in [15], this method suffers from a slow convergence. However, in prior works [15, 53], more efficient iterative variational Bayesian approximation methods have been proposed based on a transposition of classical optimization methods in Hilbert spaces [34, 9] into the space of probability density functions. Here in order to get an efficient Bayesian approach, we apply the improved memory gradient subspace-based variational Bayesian approximation method of [53].

Concerning the introduction of appropriate prior information, it plays an important role in the quality of reconstructions. In this work, we are mainly interested in piecewise smooth/constant images. For such images, the total variation (TV) prior has been extensively used [38, 47, 41, 3, 35, 39, 53, 32, 49, 7] due to its ability to preserve edges while reducing noise. Some other recent methods such as [51, 52] are based either on the regularity of the wavelength spectrum [51] or on tensor and automatic factorizations [52]. However we are here more interested in a model which promotes a clear reconstruction of contours. In this context, a sparser representation of the edges can be preferable. In works such as [30, 40], the spike and slab prior, also known as Bernoulli-Gaussian prior are used to enhance sparsity in the image domain. However, our prior developed in this paper allows a sparse representation of edges and at the mean time, can be used in variational Bayesian framework to develop unsupervised approaches for general linear inverse problems. This kind of prior is interesting for instance in non-destructive testing (NDT) applications where objects being inspected are generally composed of a set of homogeneous materials. In such applications, it is often needed to detect and characterize defects such as cracks present in the objects. As a result, it is very important to know precisely the contours of the objects which are in general sharper than the contours of natural images. To this end, a prior distribution which can well describe the piecewise smooth/constant property of the objects is required.

Another widely used class of model for piecewise smooth/constant images is the compound Markov random field [27] which models piecewise smooth images by introducing hidden variables to image boundaries between smooth regions [18], or modeling images as a composition of restricted number of different homogeneous material [14, 2, 20] or enforcing sparsity of image gradients [23, 48]. In [23], Giovannelli proposed a non-Gaussian compound Markov field with an analytic partition function adapted to deconvolution

problems. This field is based on the principle of the *half-quadratic* scheme proposed by Geman and Yang [17] and its Bayesian interpretation of hidden variables in terms of Gaussian location mixture given in [5]. In this work, we consider a hierarchical prior model based on an analog construction but allowing more flexibility on the contours. We also show how this model can be applied to any large dimensional linear inverse problem. In our prior model, the conditional distribution of the unknown image pixels given hidden variables is multivariate Gaussian and the hidden variables follow a separable Laplace distribution which corresponds to sparsity information. Moreover we introduce an alternate representation of this model which enables an easier tuning of the hyperparameters. In fact a different parametrization allows us to identify a shape and a scaling parameter. Hence we can fix the required sparsity thanks to the shape parameter while the scaling one, which controls the compromise between information coming from the data and information coming from the prior, is automatically estimated. Thus in this paper we present a different model for images that allows more flexibility than state of the art ones for piecewise smooth/constant images that arise in NDT for instance. We also determine how this model can be used for unsupervised inversion in the case of ill-posed linear inverse problems and finally how it can be implemented in an accelerated version of variational bayesian algorithm. We also provide implementation results in deconvolution and denoising cases in order to test the proposed prior.

The rest of this paper is organized as follows: in Section II, we give our Bayesian modeling. Next, the development of unsupervised Bayesian reconstruction approach using a fast variational Bayesian approximation method is given in Section III whereas results of numerical experiments are given in Section IV. Finally, a conclusion is drawn in Section V.

2 Bayesian modeling

2.1 Direct model

We consider here a classical linear observation model:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^M$ and $\mathbf{x} \in \mathbb{R}^N$ denote respectively data and unknown image to be estimated, arranged in column lexicographic ordering. The observation

operator $\mathbf{A} \in \mathbb{R}^{M\times N}$ is assumed to be known and $\mathbf{n}$ is an additive white noise, assumed to be i.i.d. Gaussian, $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \gamma_n^{-1}\mathbf{I})$, with γ_n as the precision parameter i.e. the inverse of the noise variance. The direct model (1) and the hypothesis of i.i.d. Gaussian noise allow an easy derivation of the likelihood function:

$$p(\mathbf{y}|\mathbf{x}, \gamma_n) \propto \gamma_n^{M/2} \exp\left[-\frac{\gamma_n \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2}{2}\right]. \tag{2}$$

2.2 A hierarchical image prior

In this work, we are interested in a image model, i.e. a prior distribution of the unknown image $\mathbf{x}$, satisfying the two following properties. Firstly, it is able to describe the piecewise smoothness property of images of interest. Secondly, we should have some knowledge about its partition function in order to develop unsupervised approaches which allow us to sidestep the difficulty of tuning hyperparameters. In this section, we define a hierarchical prior model based on Geman and Yang method, [17], which generalizes the work of Giovannelli [23].

For the conditional distribution of $\mathbf{x}$ given an hidden variable $\mathbf{b}$, we consider an extended Gaussian function:

$$\begin{aligned} p(\mathbf{x}|\mathbf{b}, \gamma_d) = & Z_{\mathbf{x}|\mathbf{b}}^{-1} \exp\Big[-\frac{\gamma_d}{2}\big(\|\mathbf{D}_h\mathbf{x} - \mathbf{b}_h\|^2 \\ & + \|\mathbf{D}_v\mathbf{x} - \mathbf{b}_v\|^2\big)\Big]. \end{aligned} \tag{3}$$

Here γ_d is the precision parameter of the conditional distribution of $\mathbf{x}$ given $\mathbf{b}$, the matrix $\mathbf{D} = [\mathbf{D}_h^T, \mathbf{D}_v^T]^T$ with $\mathbf{D}_h$ and $\mathbf{D}_v$ as horizontal and vertical first-order finite difference matrices and the hidden variable $\mathbf{b} = [\mathbf{b}_h^T, \mathbf{b}_v^T]^T$ where $\mathbf{b}_h$, $\mathbf{b}_v$ represent mean values of differences between adjacent pixels in horizontal and vertical directions. Moreover, $Z_{\mathbf{x}|\mathbf{b}}$ is the partition function of $p(\mathbf{x}|\mathbf{b}, \gamma_d)$. It is defined as

$$Z_{\mathbf{x}|\mathbf{b}} = \int \exp\Big[-\frac{\gamma_d}{2}\left(\|\mathbf{D}_h\mathbf{x} - \mathbf{b}_h\|^2 + \|\mathbf{D}_v\mathbf{x} - \mathbf{b}_v\|^2\right)\Big]\, \mathrm{d}\mathbf{x}. \tag{4}$$

In [23] $\mathbf{D}$ is a simple operator such that $\mathbf{D}\mathbf{x}$ has the same dimension as $\mathbf{x}$. In this case, the partition function has been shown to be independent of the auxiliary variable $\mathbf{b}$, and to be a function of the hyperparameter γ_d. In our case, we consider more flexibility in the horizontal and vertical components

which can be chosen independently. In this case the size of $\mathbf{Dx}$ is twice the size of $\mathbf{x}$. However we can still compute the partition function by transposing the integral into the Fourier domain, as done in [23] (see Appendix A for more details). In this case, the partition function is shown to be a product of the term $c{\gamma_d}^{-\frac{N}{2}}$ which only involves the hyperparameter and another term depending on both the hyperparameter γ_d and hidden variables $\mathbf{b}$, more precisely the Fourier transform of $\mathbf{b}$. This partition function is of a complicated form but is upper bounded by:

$$Z_{\mathbf{x}|\mathbf{b}} \leq c\gamma_d^{-N/2}. \tag{5}$$

This upper bound depends only on the hyperparameter γ_d. In the following it is used as an approximation of the partition function to reduce the computation burden.

In order to construct a prior distribution introducing piecewise smoothness information, as in [23], a Laplace prior is introduced for the hidden variables $\mathbf{b}$. We have therefore

$$p(\mathbf{b}|\gamma_b) \propto {\gamma_b}^{\xi} \exp\left[-\frac{\gamma_b}{2}\|\mathbf{b}\|_1\right], \tag{6}$$

where $\|\mathbf{b}\|_1 = \sum_i |b_i|$ is the L^1 norm, and ${\gamma_b}^{\xi}$ with $0 \leq \xi \leq 2N$ is the normalization constant. We introduce this constant ξ instead of $2N$ (size of $\mathbf{b}$) considering that in practice we may have some prior knowledge about the support of $\mathbf{b}$.

As a result, the prior distribution of $\mathbf{x}$ can be obtained by integrating out the hidden variable,

$$\begin{aligned} &p(\mathbf{x}|\gamma_d, \gamma_b) \\ &\widetilde{\propto} \int \gamma_d^{N/2} \exp\left[-\frac{\gamma_d}{2}\left(\|\mathbf{D}_h\mathbf{x} - \mathbf{b}_h\|^2 + \|\mathbf{D}_v\mathbf{x} - \mathbf{b}_v\|^2\right)\right] \\ &\qquad \times {\gamma_b}^{\xi} \exp\left[-\frac{\gamma_b}{2}\|\mathbf{b}\|_1\right] \mathrm{d}\mathbf{b}, \end{aligned} \tag{7}$$

where $\widetilde{\propto}$ means "is approximately proportional to".

Our Bayesian modeling can thus be summarized by the hierarchical model of Figure 1.

From (7) we can see that this distribution depends on two hyperparameters γ_d and γ_b. As those parameters define together the shape and the scale of the prior distribution, and this in an implicit manner, their automatic determination is hardly tractable. To overcome this difficulty, we propose here

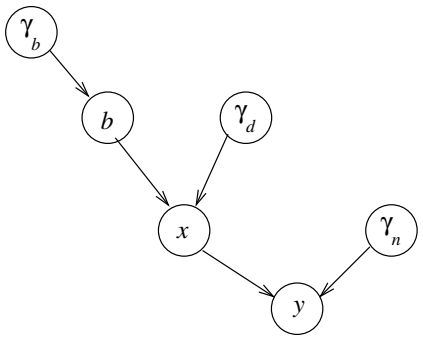


Figure 1: The proposed hierarchical model: the observation $\mathbf{y}$ given $\mathbf{x}$ follows a Gaussian distribution, the image $\mathbf{x}$ given the hidden variable $\mathbf{b}$ follows a Gaussian distribution and $\mathbf{b}$ follows a Laplace distribution, γ_n, γ_b and γ_d are hyperparameters.

to carry out a reparameterization which introduces a shape parameter and a scale one. Then the shape parameter can be fixed in order to impose a prior behaviour and the scale parameter is automatically estimated.

From (7), we can easily obtain that for all $a \in \mathbb{R}^+$

$$p(\mathbf{x}|\gamma_d', \gamma_b') = a^N p(a\mathbf{x}|\gamma_d, \gamma_b) \text{ with } \gamma_d' = a^2\gamma_d, \gamma_b' = a\gamma_b. \tag{8}$$

The above equation indicates that when we change the value of parameters γ_d and γ_b while keeping the value of the ratio $\frac{\gamma_b}{\sqrt{\gamma_d}}$, the form of the distribution does not change, i.e. only scale and amplitude change. We can then identify the ratio of γ_b to $\sqrt{\gamma_d}$ as the shape parameter of $p(\mathbf{x}|\gamma_d, \gamma_b)$ and take either γ_d or γ_b as the scale parameter. In the following, we note this form parameter by ν, that is

$$\nu = \frac{\gamma_b}{\sqrt{\gamma_d}}. \tag{9}$$

We can see in Figure 2 that when ν is small, the distribution seems heavy-tailed, more specially Cauchy-like, and when ν increases the shape of distribution becomes less heavy-tailed (Gaussian-like). In fact, as we can see in Eq. (7), in the scalar case we can see that when x tends to infinity we have

$$p(x|\gamma_d, \gamma_b) \sim \exp(-\frac{x\nu\sqrt{\gamma_d}}{2}).$$

thus the tail of the distribution depends on ν.

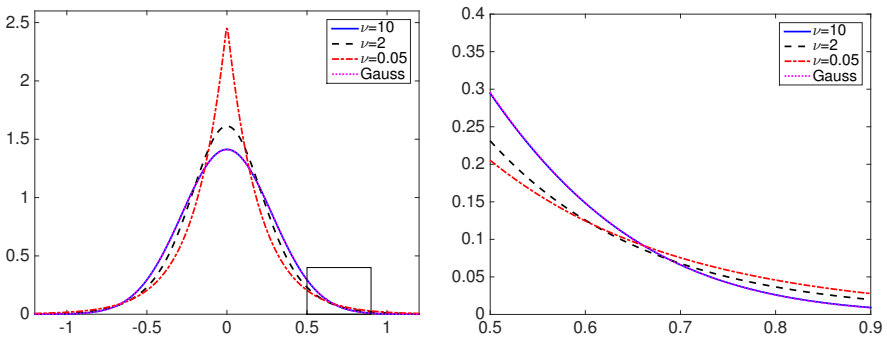


Figure 2: (a) Shape of the prior distribution of x using approximate partition function, (b) Zoom of the part within the rectangle box in (a)

Here γ_d is taken as the scale parameter. Then replacing γ_b by $\nu\sqrt{\gamma_d}$, a joint distribution involving the auxiliary variables can be given as follows:

$$\begin{aligned} & p(\mathbf{y}, \mathbf{x}, \mathbf{b} | \gamma_n, \gamma_d, \nu) \\ & = p(\mathbf{y}|\mathbf{x}, \gamma_n) p(\mathbf{x}|\mathbf{b}, \gamma_d) p(\mathbf{b}|\gamma_d, \nu) \\ & \approx C \gamma_n^{M/2} \exp\left[-\frac{\gamma_n}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2\right] \\ & \quad \times \gamma_d^{N/2} \exp\left[-\frac{\gamma_d}{2}\left(\|\mathbf{D}_h\mathbf{x} - \mathbf{b}_h\|^2 + \|\mathbf{D}_v\mathbf{x} - \mathbf{b}_v\|^2\right)\right] \\ & \quad \times (\nu\sqrt{\gamma_d})^{\xi} \exp\left[-\frac{\nu\sqrt{\gamma_d}}{2}\|\mathbf{b}\|_1\right]. \end{aligned} \tag{10}$$

The hierarchical model corresponding to the new parametrization is given by Figure 3 below. This model allows a better control on the sparsity of the gradients than TV model or the model of [23].

2.3 Hyperpriors

Hyperparameters γ_n and γ_d play an important role in the performance of algorithms. In practice, choosing correct hyperparameters is far from a trivial task. Therefore, we prefer to automatically determine their values. This is done by introducing hyperpriors for the hyperparameters. In order to obtain numerically implementable approaches, conjugate hyperpriors are employed.

For γ_n, we use a Gamma distribution,

$$p(\gamma_n) = \mathcal{G}(\gamma_n|\tilde{a},\tilde{b}) = \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\gamma_n^{\tilde{a}-1}\exp\left(-\tilde{b}\gamma_n\right) \tag{11}$$

As we do not have any prior information about γ_n, we consider $\tilde{a} \approx 0$, $\tilde{b} \approx 0$, which approximates the non-informative Jeffreys' prior. Other informative hyperpriors are also possible. Interested readers can refer to [13] for more discussions.

However, for γ_d, a Gamma distribution is no longer a conjugate prior due to the term $\sqrt{\gamma_d}$ present in (10). To overcome this difficulty we firstly carry out another reparameterization where we define a new hyperparameter $\mathcal{K}_d$ as $\mathcal{K}_d = \sqrt{\gamma_d}$. Then we define a non common probability distribution which can be taken as a conjugate prior for $\mathcal{K}_d$. In the following, this distribution is denoted by Ψ and parameterized by α, β and μ. Its probability density function is given by

$$\begin{aligned}\Psi(z|\alpha,\beta,\mu) = \frac{1}{Z_\Psi}z^\alpha \exp\left[-\beta(z+\mu)^2\right], \\ \text{for } z>0 \text{ and } \alpha,\beta,\mu>0,\end{aligned} \tag{12}$$

where the partition function Z_Ψ is computed as follows

$$\begin{aligned}Z_\Psi &= \int_0^{+\infty} z^\alpha \exp\left[-\beta(z+\mu)^2\right]\mathrm{d}z \\ &= \sum_{i=0}^{\alpha}\binom{\alpha}{i}(-\mu)^i\int_\mu^{+\infty} z'^{\alpha-i}\exp(-\beta z'^2)\mathrm{d}z' \\ &= \sum_{i=0}^{\alpha}\binom{\alpha}{i}(-\mu)^i\frac{1}{2\beta^{\frac{\alpha-i+1}{2}}}\Gamma\left(\frac{\alpha-i+1}{2},\beta\mu^2\right),\end{aligned} \tag{13}$$

here $\Gamma(r,u) = \int_u^\infty t^{r-1}e^{-t}\mathrm{d}t$ is the upper incomplete Gamma function. Therefore Z_Ψ exists and belongs to $(0,+\infty)$. In this paper, we adopt the distribution Ψ as the prior of $\mathcal{K}_d$,

$$p(\mathcal{K}_d) = \Psi(\mathcal{K}_d|\tilde{\alpha},\tilde{\beta},\tilde{\mu}). \tag{14}$$

In practice, in order to get an approximate non informative Jeffreys' prior, we choose $\tilde{\alpha} = -1$, $\tilde{\beta} \approx 0$ and $\tilde{\mu} \approx 0$.

Replacing γ_d by $\mathcal{K}_d^2$ and using the prior distribution defined above, we can obtain a joint distribution as follows,

$$\begin{aligned}
&p(\mathbf{y},\mathbf{x},\mathbf{b},\gamma_n,\mathcal{K}_d|\nu) \\
&\quad = p(\mathbf{y}|\mathbf{x},\gamma_n)p(\mathbf{x}|\mathbf{b},\mathcal{K}_d)p(\mathbf{b}|\mathcal{K}_d,\nu)p(\gamma_n)p(\mathcal{K}_d) \\
&\quad \approx C\gamma_n^{M/2}\exp\left[-\frac{\gamma_n}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2\right] \\
&\qquad \times \mathcal{K}_d^N \exp\left[-\frac{\mathcal{K}_d^2}{2}\left(\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2+\|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2\right)\right] \\
&\qquad \times (\nu\mathcal{K}_d)^{\xi}\exp\left[-\frac{\nu\mathcal{K}_d}{2}\|\mathbf{b}\|_1\right]\gamma_n^{-1}\mathcal{K}_d^{-1}. \qquad (15)
\end{aligned}$$

The posterior distribution $p(\mathbf{x},\mathbf{b},\gamma_n,\mathcal{K}_d|\mathbf{y},\nu)$ is not known explicitly since its partition function is not calculable. In order to proceed the Bayesian inference which is based on the posterior distribution, we resort to variational Bayesian methods which aims at getting the best separable analytical approximation of the true posterior distribution.

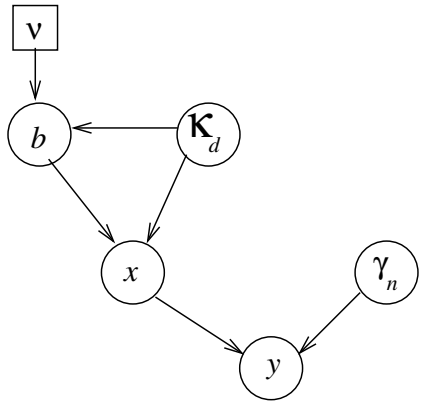


Figure 3: The hierarchical model after reparameterization. Hyperparameter ν is fixed, other ones are estimated

3 Proposed unsupervised approach using variational Bayesian approximation

In this section we present variational bayesian methods used in this paper, see [44, 15, 53] for more details. We denote by $\Theta = \{\mathbf{x}, \mathbf{b}, \gamma_n, \mathcal{K}_d\}$ the vector enclosing all the parameters to be estimated. Variational Bayesian approximation methods give a separable approximate distribution q_Θ of the true posterior by minimizing the Kullback-Leibler ($\mathcal{KL}$) divergence between them. Nevertheless, in practice, the $\mathcal{KL}$ divergence is generally intractable since it depends on the unknown true posterior distribution. However, as stated in [8], minimizing $\mathcal{KL}$ divergence is equivalent to maximizing the negative free energy $\mathcal{F}(q_\Theta)$ which depends on the joint distribution $p(\mathbf{y}, \Theta)$. The negative free energy is defined as follows:

$$\mathcal{F}(q_\Theta) = \int q_\Theta(\Theta) \ln \frac{p(\mathbf{y}, \Theta)}{q_\Theta(\Theta)} \mathrm{d}\Theta. \tag{16}$$

Since the joint distribution is known, the negative free energy is then used as an alternative to the $\mathcal{KL}$ divergence. As a result, the problem of finding the best separable approximation can be mathematically described as follows:

$$q_\Theta^{opt} = \arg\max_{q_\Theta} \mathcal{F}(q_\Theta), \;\; s.t. \;\; q_\Theta(\Theta) = \prod_i q_i(\Theta_i). \tag{17}$$

This optimization problem has the following analytic solution (see [44] for details on the variational calculus)

$$q_i(\Theta_i) \propto \exp\left(\langle \ln p(\mathbf{y}, \Theta) \rangle_{\prod_{j \neq i} q_j(\Theta_j)} \right). \tag{18}$$

Classical variational Bayesian approach (VBA) is based on this analytic solution. However, in (18), each distribution q_i depends on the other distributions q_j with j different from i. In practice, this dependence implies the employment of iterative methods such as the Gauss-Seidel one, which are not very efficient. Therefore, classical VBA is generally not very efficient for large problems.

In [15], a more efficient gradient-type variational Bayesian method was proposed and its convergence studied. It is based on a transposition of the gradient descent algorithm in Hilbert spaces into the space of probability density functions. Based on this work, we proposed in a recent work [53]

an improvement of this algorithm by a transposition of the memory gradient subspace optimization method in Hilbert spaces into the space of probability density functions. This Memory Gradient subspace based Variational Bayesian Algorithm (MG-VBA) uses the following updating formula:

$$q_i^{k+1}(\Theta_i) = K^k(\mathbf{s}) q_i^k(\Theta_i) \left[\frac{\exp\left(\langle \ln p(\mathbf{y}, \Theta) \rangle_{\prod_{j \neq i} q_j^k(\Theta_j)} \right)}{q_i^k(\Theta_i)} \right]^{s_1} \times \left[\frac{q_i^k(\Theta_i)}{q_i^{k-1}(\Theta_i)} \right]^{s_2} \tag{19}$$

where $\mathbf{s} = (s_1, s_2)$ with $s_1 > 0$ and $s_2 > 0$ are the algorithm step sizes and $K^k(\mathbf{s})$ is a normalization constant. In [53], an approximate optimal step size was proposed thanks to the second order Taylor development of the objective criterion. This step size is also used in this work.

We present in the following the application of variational Bayesian approximations for our problem. Considering the approximation set, we consider here a separability given as follows

$$\begin{aligned} q_\Theta(\Theta) &= q_{\mathbf{x}}(\mathbf{x}) q_{\mathbf{b}}(\mathbf{b}) q_{\gamma_n}(\gamma_n) q_{\mathcal{K}_d}(\mathcal{K}_d) \\ &= \prod_i q_i(x_i) q_{\mathbf{b}}(\mathbf{b}) q_{\gamma_n}(\gamma_n) q_{\mathcal{K}_d}(\mathcal{K}_d). \end{aligned} \tag{20}$$

The optimization of q_Θ can then be performed following the alternate iterative scheme:

$$q_{\mathbf{x}}^{k+1} = \arg\max_{q_{\mathbf{x}}} \mathcal{F}\left(q_{\mathbf{x}}, q_{\mathbf{b}}^k, q_{\gamma_n}^k, q_{\mathcal{K}_d}^k\right), \tag{21}$$

$$q_{\mathbf{b}}^{k+1} = \arg\max_{q_{\mathbf{b}}} \mathcal{F}\left(q_{\mathbf{x}}^{k+1}, q_{\mathbf{b}}, q_{\gamma_n}^k, q_{\mathcal{K}_d}^k\right), \tag{22}$$

$$q_{\gamma_n}^{k+1} = \arg\max_{q_{\gamma_n}} \mathcal{F}(q_{\mathbf{x}}^{k+1}, q_{\mathbf{b}}^{k+1}, q_{\gamma_n}, q_{\mathcal{K}_d}^k), \tag{23}$$

$$q_{\mathcal{K}_d}^{k+1} = \arg\max_{q_{\mathcal{K}_d}} \mathcal{F}(q_{\mathbf{x}}^{k+1}, q_{\mathbf{b}}^{k+1}, q_{\gamma_n}^{k+1}, q_{\mathcal{K}_d}). \tag{24}$$

As introduced above, conjugate priors have been chosen for $\mathbf{x}$ and hyperparameters γ_n, $\mathcal{K}_d$. Therefore either classical VBA or MG-VBA yields Gaussian distributions for $(q_i)_{i=1,\ldots,N}$ with means and variances collected in vectors $\mathbf{m}_k$ and $\boldsymbol{\sigma}_k^2$, a Gamma distribution for q_{γ_n} and a Ψ distribution for

$q_{\mathcal{K}_d}$. As a result, we have

$$q_{\mathbf{x}}^k(\mathbf{x}) = \prod_i \mathcal{N}(x_i|(\mathbf{m}_k)_i, (\boldsymbol{\sigma}_k^2)_i), \tag{25}$$

$$q_{\gamma_n}^k(\gamma_n) = \mathcal{G}(\gamma_n|a^k, b^k), \tag{26}$$

$$q_{\mathcal{K}_d}^k(\mathcal{K}_d) = \Psi(\mathcal{K}_d|\alpha^k, \beta^k, \mu^k). \tag{27}$$

Therefore, the optimization of distributions $(q_i)_{i=1,\ldots,N}$, q_{γ_n} and $q_{\mathcal{K}_d}$ is performed by updating their parameters.

However, in our problem, the optimization of $q_{\mathbf{b}}$ is more complicated since we employ a Laplace distribution, see (6), as prior distribution for $\mathbf{b}$, which is not conjugate with $p(\mathbf{x}|\mathbf{b}, \gamma_b)$ (see (3)). As a result, the free form approximation obtained by variational Bayesian approaches does not always belong to a same family of distributions. Consequently, the optimization of $q_{\mathbf{b}}$ cannot be simply performed by optimizing some parameters. To tackle this issue, we propose here to employ the restricted VBA of [44] where rather than determine a totally free form distributional approximation for $q_{\mathbf{b}}$ by using either the classical VBA based on (18) or the MG-VBA based on (19) we choose its distribution form in advance. Here, in order to sidestep the difficulty posed by the Laplace distribution, we suppose that $q_{\mathbf{b}}^k$ is a Dirac delta function,

$$q_{\mathbf{b}}^k(\mathbf{b}) = \delta(\mathbf{b} - \hat{\mathbf{b}}^k). \tag{28}$$

Therefore, instead of solving the free form optimization problem given by (22), we maximize the negative free energy under the condition that $q_{\mathbf{b}}$ is a Dirac delta function. This problem is mathematically described by

$$q_{\mathbf{b}}^{k+1} = \underset{q_{\mathbf{b}} \text{ is Dirac}}{\arg\max}\, \mathcal{F}\left(q_{\mathbf{x}}^{k+1}, q_{\mathbf{b}}, q_{\gamma_n}^k, q_{\mathcal{K}_d}^k\right). \tag{29}$$

As a result, the optimization of the distribution $q_{\mathbf{b}}$ can also be performed by optimizing its parameter $\hat{\mathbf{b}}$. In fact, in the case where the parameter distribution is restricted to a Dirac delta function (i.e. a point estimate), this optimization step can be identified as the M-step of an EM algorithm [1, 19, 4].

Since the conditional posterior $p(\mathbf{b}, \gamma_n, \mathcal{K}_d|\mathbf{x}, \mathbf{y}, \nu)$ is separable, it can be approximated efficiently thanks to the classical VBA. In fact, the MG-VBA is only adopted to approximate the posterior distribution of $\mathbf{x}$ in order to get a real gain in convergence rate.

3.1 Optimization of $q_{\mathbf{x}}$

For the optimization of $q_{\mathbf{x}}$, the MG-VBA is adopted. Due to the conjugate prior, $q_{\mathbf{x}}$ is a Gaussian distribution given by (25). Therefore, updating the distributions $q_{\mathbf{x}}$ is equivalent to updating its mean and variance vectors $\mathbf{m}_k$ and $\boldsymbol{\sigma}_k^2$.

According to (19) and the separability assumption (20), we can obtain

$$q_{\mathbf{x}}^{\mathbf{s}}(\mathbf{x}) = K^k(\mathbf{s}) q_{\mathbf{x}}^k(\mathbf{x}) \prod_i \left(\frac{q_i^r(x_i)}{q_i^k(x_i)}\right)^{s_1} \left(\frac{q_i^k(x_i)}{q_i^{k-1}(x_i)}\right)^{s_2}, \tag{30}$$

where the auxiliary functions q_i^r are given by

$$\begin{aligned}
&q_i^r(x_i) \\
&= \exp\left[\langle \ln p(\mathbf{y}, \Theta)\rangle_{\prod_{j\neq i} q_j^k q_{\mathbf{b}}^k q_{\gamma_n}^k q_{\mathcal{K}_d}^k}\right] \\
&\propto \exp\Bigg[-\int \Bigg(\frac{\gamma_n}{2}\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2 \\
&\qquad + \frac{\mathcal{K}_d^2}{2}\left(\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2 + \|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2\right)\Bigg) \\
&\qquad \times \prod_{j\neq i} q_j^k(x_j) q_{\mathbf{b}}^k(\mathbf{b}) q_{\gamma_n}^k(\gamma_n) q_{\mathcal{K}_d}^k(\mathcal{K}_d) \mathrm{d}x_j \mathrm{d}\mathbf{b} \mathrm{d}\gamma_n \mathrm{d}\mathcal{K}_d\Bigg] \\
&\propto \exp\Bigg[-\frac{\langle\gamma_n\rangle^k}{2}\Bigg(x_i^2 \mathrm{diag}\left(\mathbf{A}^T\mathbf{A}\right)_i - 2x_i\left(\mathbf{A}^T\mathbf{y}\right)_i \\
&\qquad + 2x_i\left(\mathbf{A}^T\mathbf{A}\mathbf{m}_k\right)_i - 2x_i \mathrm{diag}\left(\mathbf{A}^T\mathbf{A}\right)_i (\mathbf{m}_k)_i\Bigg) \\
&\qquad - \frac{\langle\mathcal{K}_d^2\rangle^k}{2}\Bigg(x_i^2 \mathrm{diag}\left(\mathbf{D}_h^T\mathbf{D}_h\right)_i - 2x_i\left(\mathbf{D}_h^T\hat{\mathbf{b}}_h^k\right)_i \\
&\qquad\quad + 2x_i\left(\mathbf{D}_h^T\mathbf{D}_h\mathbf{m}_k\right)_i - 2x_i \mathrm{diag}\left(\mathbf{D}_h^T\mathbf{D}_h\right)_i (\mathbf{m}_k)_i \\
&\qquad\quad + x_i^2 \mathrm{diag}\left(\mathbf{D}_v^T\mathbf{D}_v\right)_i - 2x_i\left(\mathbf{D}_v^T\hat{\mathbf{b}}_v^k\right)_i \\
&\qquad\quad + 2x_i\left(\mathbf{D}_v^T\mathbf{D}_v\mathbf{m}_k\right)_i - 2x_i \mathrm{diag}\left(\mathbf{D}_v^T\mathbf{D}_v\right)_i (\mathbf{m}_k)_i\Bigg)\Bigg]
\end{aligned} \tag{31}$$

where $\langle\gamma_n\rangle^k = \mathbb{E}_{q_{\gamma_n}^k}(\gamma_n)$, $\langle\mathcal{K}_d^2\rangle^k = \mathbb{E}_{q_{\mathcal{K}_d}^k}(\mathcal{K}_d^2)$ and $\mathrm{diag}(\mathbf{M})$ is a vector containing the diagonal elements of matrix $\mathbf{M}$.

We can see from (31) that for $i = 1, \ldots, N$, q_i^r corresponds, up to a normalization constant, to the density of a Gaussian distribution with variance $(\boldsymbol{\sigma}_r^2)_i$ and mean $(\mathbf{m}_r)_i$ expressed explicitly by the two following expressions:

$$\begin{aligned}(\boldsymbol{\sigma}_r^2)_i =& \Big[\langle\gamma_n\rangle^k \mathrm{diag}\big(\mathbf{A}^T\mathbf{A}\big)_i \\ &+\langle\mathcal{K}_d^2\rangle^k\Big(\mathrm{diag}\big(\mathbf{D}_h^T\mathbf{D}_h\big)_i+\mathrm{diag}\big(\mathbf{D}_v^T\mathbf{D}_v\big)_i\Big)\Big]^{-1} \qquad (32)\\ (\mathbf{m}_r)_i =&(\boldsymbol{\sigma}_r^2)_i\Big[\langle\gamma_n\rangle^k\left(\mathbf{A}^T\mathbf{y}-\mathbf{A}^T\mathbf{A}\mathbf{m}_k+\mathrm{diag}\big(\mathbf{A}^T\mathbf{A}\big)\circ\mathbf{m}_k\right) \\ &+\langle\mathcal{K}_d^2\rangle^k\Big(\mathbf{D}_h^T\hat{\mathbf{b}}_h^k-\mathbf{D}_h^T\mathbf{D}_h\mathbf{m}_k+\mathrm{diag}\big(\mathbf{D}_h^T\mathbf{D}_h\big)\circ\mathbf{m}_k\Big) \\ &+\langle\mathcal{K}_d^2\rangle^k\Big(\mathbf{D}_v^T\hat{\mathbf{b}}_v^k-\mathbf{D}_v^T\mathbf{D}_v\mathbf{m}_k+\mathrm{diag}\big(\mathbf{D}_v^T\mathbf{D}_v\big)\circ\mathbf{m}_k\Big)\Big]_i \qquad (33)\end{aligned}$$

where $\circ$ denotes the pointwise product.

Based on the above results for q_i^r, using (30), we can derive the expression of $q_{\mathbf{x}}^{\mathbf{s}}(\mathbf{x}) = \prod_i q_i^{\mathbf{s}}(x_i)$ where each component $q_i^{\mathbf{s}}(x_i)$ is a Gaussian distribution with mean $(\mathbf{m}_{\mathbf{s}})_i$ and variance $(\boldsymbol{\sigma}_{\mathbf{s}}^2)_i$ satisfying:

$$\boldsymbol{\sigma}_{\mathbf{s}}^2 = \left[\frac{1}{\boldsymbol{\sigma}_k^2}+s_1\left(\frac{1}{\boldsymbol{\sigma}_r^2}-\frac{1}{\boldsymbol{\sigma}_k^2}\right)+s_2\left(\frac{1}{\boldsymbol{\sigma}_k^2}-\frac{1}{\boldsymbol{\sigma}_{k-1}^2}\right)\right]^{-1}, \qquad (34)$$

$$\mathbf{m}_{\mathbf{s}} = \boldsymbol{\sigma}_{\mathbf{s}}^2\left[\frac{\mathbf{m}_k}{\boldsymbol{\sigma}_k^2}+s_1\left(\frac{\mathbf{m}_r}{\boldsymbol{\sigma}_r^2}-\frac{\mathbf{m}_k}{\boldsymbol{\sigma}_k^2}\right)+s_2\left(\frac{\mathbf{m}_k}{\boldsymbol{\sigma}_k^2}-\frac{\mathbf{m}_{k-1}}{\boldsymbol{\sigma}_{k-1}^2}\right)\right]. \qquad (35)$$

In above equations, we omit all the indication of vector component $(\cdot)_i$ to lighten notations.

For the step-size $\mathbf{s}$, as introduced above, a sub-optimal one defined in [53] is adopted. Therefore, $\boldsymbol{\sigma}_{k+1}^2 = \boldsymbol{\sigma}_{\mathbf{s}^{\mathrm{subopt}}}^2$ and $\mathbf{m}_{k+1} = \mathbf{m}_{\mathbf{s}^{\mathrm{subopt}}}$.

3.2 Optimization of $q_{\mathbf{b}}$

As mentioned above, we aim at getting a Dirac delta function for $q_{\mathbf{b}}$ which maximizes the negative free energy in order to solve the problem (29). Since the Dirac delta function is parameterized by $\hat{\mathbf{b}}$, the optimization of $q_{\mathbf{b}}$ is

performed by optimizing $\hat{\mathbf{b}}$. The resolution of (29) leads to

$$\begin{aligned}
\hat{\mathbf{b}}^{k+1} &= \arg\max_{\hat{\mathbf{b}}} \left[\langle \ln p(\mathbf{y}, \Theta) \rangle_{q_{\mathbf{x}}^{k+1} \delta(\mathbf{b}-\hat{\mathbf{b}}) q_{\gamma_n}^k q_{\mathcal{K}_d}^k} \right] \\
&= \arg\max_{\hat{\mathbf{b}}} \left[-\int \left(\frac{\mathcal{K}_d^2}{2} \left(\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2 + \|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2 \right) \right.\right. \\
&\qquad \left.\left. + \frac{\nu\mathcal{K}_d}{2}\|\mathbf{b}\|_1 \right) q_{\mathbf{x}}^{k+1}(\mathbf{x}) \delta(\mathbf{b}-\hat{\mathbf{b}}) \right. \\
&\qquad \left. \times q_{\gamma_n}^k(\gamma_n) q_{\mathcal{K}_d}^k(\mathcal{K}_d) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{b}\mathrm{d}\gamma_n\mathrm{d}\mathcal{K}_d \right] \\
&= \arg\min_{\hat{\mathbf{b}}} \left[\frac{\langle \mathcal{K}_d^2 \rangle^k}{2} \left(\|\mathbf{D}_h\mathbf{m}_{k+1} - \hat{\mathbf{b}}_h\|^2 + \|\mathbf{D}_v\mathbf{m}_{k+1} - \hat{\mathbf{b}}_v\|^2 \right) \right. \\
&\qquad \left. + \frac{\nu\langle \mathcal{K}_d \rangle^k}{2}\|\hat{\mathbf{b}}\|_1 + \text{const} \right],
\end{aligned} \tag{36}$$

which falls to a L^1 norm regularized linear regression problem [46, 11] whose solution is known as a soft-thresholding which can be expressed as

$$\hat{\mathbf{b}}_h^{k+1} = \mathrm{sgn}\left(\mathbf{D}_h\mathbf{m}_{k+1}\right) \max\left(|\mathbf{D}_h\mathbf{m}_{k+1}| - \frac{\nu\langle \mathcal{K}_d \rangle^k}{2\langle \mathcal{K}_d^2 \rangle^k}, \mathbf{0} \right), \tag{37}$$

$$\hat{\mathbf{b}}_v^{k+1} = \mathrm{sgn}\left(\mathbf{D}_v\mathbf{m}_{k+1}\right) \max\left(|\mathbf{D}_v\mathbf{m}_{k+1}| - \frac{\nu\langle \mathcal{K}_d \rangle^k}{2\langle \mathcal{K}_d^2 \rangle^k}, \mathbf{0} \right). \tag{38}$$

3.3 Optimization of q_{γ_n}

Due to the use of the conjugate prior distribution, classical VBA induces that $q_{\gamma_n}^{k+1}$ is a Gamma distribution with parameters a^{k+1} and b^{k+1}. Using the formula (18), we can obtain

$$q_{\gamma_n}^{k+1}(\gamma_n) \propto \gamma_n^{M/2-1} \exp\left[-\frac{\gamma_n}{2} \mathbb{E}_{q_{\mathbf{x}}^{k+1}} \left[\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2 \right] \right]. \tag{39}$$

The quantity $\mathbb{E}_{q_{\mathbf{x}}^{k+1}} \left[\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2 \right]$ can be calculated as

$$\begin{aligned}
\mathbb{E}_{q_{\mathbf{x}}^{k+1}} \left[\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2 \right] &= \left[\|\mathbf{y}-\mathbf{A}\mathbf{m}_{k+1}\|^2 \right] \\
&\quad + \sum_i \mathrm{diag}\left(\mathbf{A}^T\mathbf{A}\right)_i (\boldsymbol{\sigma}_{k+1}^2)_i
\end{aligned} \tag{40}$$

Parameters of the Gamma distribution are then identified as

$$a^{k+1} = M/2 = a, \tag{41}$$

$$b^{k+1} = \frac{1}{2}\mathbb{E}_{q_{\mathbf{x}}^{k+1}}\left[\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2\right]. \tag{42}$$

The mean of the distribution (39), which is used as an estimate of the hyperparameter γ_n, is then given by

$$\langle\gamma_n\rangle^{k+1} = \frac{M}{\mathbb{E}_{q_{\mathbf{x}}^{k+1}}\left[\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2\right]}. \tag{43}$$

3.4 Optimization of $q_{\mathcal{K}_d}$

As above, the optimization of $q_{\mathcal{K}_d}$ is still performed using the classical VBA. Therefore, using (18), we can obtain

$$\begin{aligned} & q_{\mathcal{K}_d}^{k+1}(\mathcal{K}_d) \\ & \propto \mathcal{K}_d^{N+\xi-1}\exp\Bigg[-\frac{\mathcal{K}_d^2}{2}\mathbb{E}_{q_{\mathbf{x}}^{k+1}q_{\mathbf{b}}^{k+1}}\left[\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2+\|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2\right] \\ & \qquad\qquad - \frac{\nu\mathcal{K}_d}{2}\mathbb{E}_{q_{\mathbf{b}}^{k+1}}\left[\|\mathbf{b}\|_1\right]\Bigg] \\ & = \Psi(\mathcal{K}_d|\alpha^{k+1}, \beta^{k+1}, \mu^{k+1}) \end{aligned} \tag{44}$$

with α^{k+1}, β^{k+1} and μ^{k+1} identified as

$$\alpha^{k+1} = N + \xi - 1 = \alpha, \tag{45}$$

$$\beta^{k+1} = \frac{1}{2}\mathbb{E}_{q_{\mathbf{x}}^{k+1}q_{\mathbf{b}}^{k+1}}\left(\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2+\|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2\right), \tag{46}$$

$$\mu^{k+1} = \frac{\nu\mathbb{E}_{q_{\mathbf{b}}^{k+1}}(\|\mathbf{b}\|_1)}{4\beta^{k+1}}. \tag{47}$$

Here,

$$\begin{aligned}
&\mathbb{E}_{q_{\mathbf{x}}^{k+1} q_{\mathbf{b}}^{k+1}} \left[\|\mathbf{D}_h \mathbf{x} - \mathbf{b}_h\|^2 + \|\mathbf{D}_v \mathbf{x} - \mathbf{b}_v\|^2 \right] \\
&\quad = \left[\|\mathbf{D}_h \mathbf{m}_{k+1} - \hat{\mathbf{b}}_h^{k+1}\|^2 + \|\mathbf{D}_v \mathbf{m}_{k+1} - \hat{\mathbf{b}}_v^{k+1}\|^2 \right] \\
&\qquad + \sum_i \operatorname{diag}\left(\mathbf{D}_h^T \mathbf{D}_h\right)_i (\boldsymbol{\sigma}_{k+1}^2)_i \\
&\qquad + \sum_i \operatorname{diag}\left(\mathbf{D}_v^T \mathbf{D}_v\right)_i (\boldsymbol{\sigma}_{k+1}^2)_i, \qquad (48)
\end{aligned}$$

$$\mathbb{E}_{q_{\mathbf{b}}^{k+1}}[\|\mathbf{b}\|_1] = \|\hat{\mathbf{b}}^{k+1}\|_1. \qquad (49)$$

As shown in (32), (33) and (37), (38), the parameters of the distributions $q_{\mathbf{x}}$ and $q_{\mathbf{b}}$ depend on the first and second-order moments of the distribution Ψ. As a result, we need to update the above moments while we update the distribution $q_{\mathcal{K}_d}$. Nevertheless, as we work with large-dimensional problem, N is large and the parameter α takes very large value, which makes the direct computation of these moments numerically intractable. To sidestep this difficulty, in our approach, we propose to adopt a Markov Chain Monte Carlo technique – the random walk Metropolis-Hastings algorithm which provides us a set of samples following the distribution $\Psi(\mathcal{K}_d|\alpha^{k+1}, \beta^{k+1}, \mu^{k+1})$. We determine then its first and second-order moments using these samples. Note that $\mathcal{K}_d$ is a one-dimensional variable. In this case the computation burden of the MCMC step is quite small.

For the random walk Metropolis-Hastings algorithm, its convergence depends on its initialization and step-sizes. Therefore, to get an algorithm relatively efficient and adaptive to all Ψ distributions, we have taken the following settings. Firstly, a random value between zero and the maximiser of the Ψ function is used as the initialization point of the random walk. This maximiser of the Ψ function has an analytical simple form:

$$\mathcal{K}_d^{\max} = -\frac{\mu}{2} + \sqrt{\frac{\mu^2}{4} + \frac{\alpha}{2\beta}}, \qquad (50)$$

which can be easily computed. Moreover, a fixed step size equal to $\frac{1}{50}\mathcal{K}_d^{\max}$ is adopted. We can show that this choice of initialization gives a good convergence of the MCMC algorithm.

Altogether, our proposed algorithm for linear inverse problem can be summed up in Algorithm 1.

Algorithm 1 Proposed unsupervised Bayesian reconstruction algorithm

1. Initialize parameters of $(q_i^0)_{i=1,\ldots,N}$, $q_{\mathbf{b}}^0$ $q_{\gamma_n}^0$ and $q_{\mathcal{K}_d}^0$
2. Update means and variances of q_i^{k+1} for $i = 1, \ldots, N$
   a. Compute parameters of intermediary functions q_i^r using (32), (33)
   b. Determine the suboptimal step-sizes $(s_1^{subopt}, s_2^{subopt})$
   c. Update means and variances of q_i^{k+1} using (34), (35)
3. Update parameters of $q_{\mathbf{b}}^{k+1}$ using (37), (38)
4. Determine parameters of $q_{\gamma_n}^{k+1}$ then compute its mean using (43)
5. Update parameters of $q_{\mathcal{K}_d}^{k+1}$ using (45–47), then determine its first and second-order moments using MCMC
6. Go back to 2 until convergence

4 Simulation results

In order to demonstrate the performance of the proposed unsupervised reconstruction approach (see Algorithm 1), we show in this section some simulation results on image reconstruction. We firstly present some results in diffraction tomography (DT) applied to Non-Destructive Testing where objects are generally piecewise smooth. In this case, the relevant prior information is well fit by our hierarchical sparse gradient prior model. Afterward, the proposed approach is tested on natural images where our prior model does not fit very well. The objective is to evaluate the robustness of our approach with respect to the piecewise smoothness hypothesis. Then, since in the involved DT problems an approximate linear forward model is used (see Section 4.1 for details), we make some discussions on the robustness of the proposed approach to model errors in Section 4.6. Finally, in order to show the versatility of the proposed approach to general linear inverse problems, applications to image denoising are also shown.

4.1 Diffraction tomography

In DT, the object of interest is subjected to an incident wave at a given frequency. Diffraction occurs when the wave encounters the object and the diffracted wave is measured by a set of sensors located around the object.

The objective of DT problem is then to reconstruct an image of the unknown object from the data collected by the sensors.

In a transverse magnetic (TM) mode, the following pair of equations can be used to model the interaction between the incident wave and the object:

$$\mathbf{y} = \mathbf{G}_{obs}\text{Diag}(\mathbf{x})\mathbf{E} \tag{51}$$

$$\mathbf{E} = \mathbf{E}_{inc} + \mathbf{G}_{cou}\text{Diag}(\mathbf{x})\mathbf{E} \tag{52}$$

where $\text{Diag}(\mathbf{x})$ is a diagonal matrix with the vector $\mathbf{x}$ containing the diagonal elements, $\mathbf{G}_{obs}$ is the Green's matrix whereas $\mathbf{G}_{cou}$ represents the coupling matrix, $\mathbf{E}_{inc}$ and $\mathbf{E}$ denote the incident field and the total field applied to the object, respectively. The above equations ((51) and (52)) show a nonlinear relationship between the observed data $\mathbf{y}$ and the unknown parameter $\mathbf{x}$ representing the permittivity of the object in our problem. The inverse problem associated to this nonlinear model is generally difficult. A classical way to bypass this issue consists of a Born approximation which is based on the assumption that the total field $\mathbf{E}$ is equal to the incident field $\mathbf{E}_{inc}$. Under such assumption, a linear approximate forward model can be obtained:

$$\mathbf{y} = \mathbf{G}_{obs}\text{Diag}(\mathbf{x})\mathbf{E}_{inc}. \tag{53}$$

This approximation is valid when the object has low contrast with respect to the background. Considering this linear model, the proposed approach (Algorithm 1) can be applied for the associated reconstruction problem.

4.2 Simulation configuration

The proposed approach is tested on synthetic data and compared with the baseline Gaussian prior based MAP estimate [12] (abbreviated as Gauss-MAP), with a Total Variation (TV) regularized MAP estimate (abbreviated as TV-MAP) computed using a primal-dual splitting algorithm [10] and with a TV prior based unsupervised variational Bayesian approach [3] (abbreviated as TV-VBA) which gives a posterior mean estimate. The previous TV refers to the more commonly adopted isotropic total variation [41, 3, 35, 39, 53]. To be more comprehensive, we include also comparisons with an anisotropic total variation [6, 43, 49] based approach. As TV-MAP, the anisotropic total variation regularized MAP estimate is computed using a primal-dual splitting algorithm [10]. In the following, the anisotropic based approach is abbreviated as AnisoTV-MAP. In Gauss-MAP, TV-MAP

and AnisoTV-MAP, the regularization parameter is manually tuned to give the best result in the sense of maximizing the Peak Signal to Noise Ratio (PSNR) value. As a result, we prefer to abbreviate these approaches as Best-Gauss-MAP, Best-TV-MAP and Best-AnisoTV-MAP in the following. Note that the manual tuning of regularization parameter requires knowledge of the ground truth of the image to be reconstructed whereas TV-VBA and the proposed approach do not require such information.

For our experiments of Section 4.3 and Section 4.4, data are simulated using the forward model (53) for the experimental configuration where 36 receivers locate around the object and a source emits incident waves of frequency 3×10^8 Hz at 36 positions uniformly distributed around the object of size 81×81.

During the implementation of Best-Gauss-MAP, Best-TV-MAP and Best-AnisoTV-MAP, $\mathbf{m}_0 = \mathbf{A}^T\mathbf{y}$ is used as the initial value of the unknown image. Concerning TV-VBA, it is implemented with the same initializations as presented in [3]. For the proposed approach, initializations are chosen similarly. Hence, at first the proposed approach uses the following initial values: $\mathbf{m}_0$ as the mean and 100 as the variance of unknown image pixels. We consider also different initializations in order to determine how it bias the results. The hidden variable $\mathbf{b}_h$ is initialized by a sparse vector $\mathbf{b}_h^0$ constructed as follows: keeping the five percent largest elements of $\mathbf{D}_h\mathbf{m}_0$ and setting all other elements to zero. The initial value of $\mathbf{b}_v$ is obtained in the same way from $\mathbf{D}_v\mathbf{m}_0$. Concerning hyperparameters, the initial value γ_n^0 is estimated from $\mathbf{m}_0$ by using its updating equation (39) and $\sqrt{\gamma_n^0}$ is used as the initial value of $\mathcal{K}_d$. Moreover, in the proposed approach, the shape parameter of our hierarchical prior distribution ν is set to 0.8 to have a heavy-tailed distribution in all our experiments presented here. Regarding ξ, it depends on the number of non-zero entries in $\mathbf{b}_h$ and $\mathbf{b}_v$. And the non-zero entries correspond to image edges. In all our experiments, we make an assumption that the proportion of edges in all image pixels is approximately 10%. As a result, we have set $\xi = 0.1N$. In our experiments, this value leads to good results even for images with different proportions of edges. We also test how this parameter influences the reconstruction. Regarding the convergence of the proposed algorithm, it is inspected through inspecting the convergence of the mean square error of the reconstructed images. Moreover, experiments shown in this paper were run using Matlab R2017b on a laptop computer with Intel Core i7 CPU (3 GHz) and 16 GB RAM.

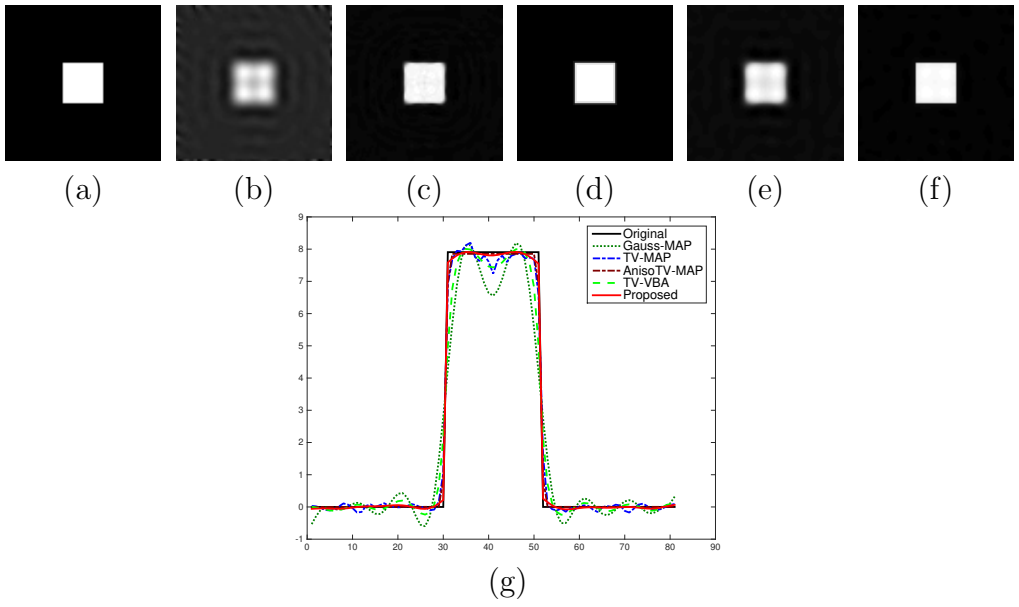


Figure 4: (a) Original image, reconstructed images with (b) Best-Gauss-MAP [12], (c) Best-TV-MAP [10], (d) Best-AnisoTV-MAP, (e)TV-VBA [3], (f) proposed approach, (g) profile of middle rows of the above images.

4.3 Validation on a simple example

We test firstly our approach by applying it to synthetic data generated from a simple image of size 81×81 composed of a black background and one square object in the center, as shown in Figure 4 (a). Moreover, data is directly generated from the approximate linear model (53). As a result, the proposed approach using the approximate linear forward model does not have any model error. A Gaussian i.i.d. background noise is added to the data which leads to Signal to Noise Ratio (SNR) equal to 30 dB. We show in Figure 4 (b)-(e) reconstructions obtained by Best-Gauss-MAP, Best-TV-MAP, Best-AnisoTV-MAP, TV-VBA and the proposed approach, respectively. As expected, Best-Gauss-MAP leads to an image with blur edges (see Figure 4 (b)). By using the TV prior, Best-TV-MAP, Best-AnisoTV-MAP and TV-VBA give much better reconstructions (see Figure 4 (c), (d) and (e)) where the noise in the background is reduced and the edges of the square object are sharper. The comparison of Figure 4 (f) with Figure 4 (b)-(e) shows

that the proposed approach gives the best reconstruction: the edges of the square object are very sharp and the background is very clear and similar to the original image shown in Figure 4 (a). To make a clearer comparison, we provide in Figure 4 (g) the profile of the middle row of the six images shown in Figure 4. We can see that the proposed approach gives sharp edges which are the closest to original edges shown with a solid line. Moreover, the reconstruction of the proposed approach presents less oscillations in flat zones. Concerning PSNR values of the reconstructions, the proposed approach gives 41.00 dB which is much higher than that of Best-TV-MAP (30.40 dB), Best-AnisoTV-MAP (34.3 dB), TV-VBA (24.88 dB) and Best-Gauss-MAP (21.79 dB).

4.4 Evaluation with different spatial frequencies and noise levels

The above results show that the proposed approach performs very well with the simple image presented in Figure 4 (a). In the following, we evaluate the proposed approach using data generated from a more complicated image which contains components of different spatial frequencies, as shown in Figure 5 (a): there are six objects of different widths and of different intensities. As in the previous case, synthetic data are generated from the object using directly the approximate linear model. Moreover, we add 6 different levels of Gaussian noise (SNR equal to 15, 20, 25, 30, 35, 40dB) to the data.

We show in Figure 5 (b)-(f) reconstructed images obtained by the five approaches for comparison in the case where SNR is equal to 30 dB. As in the previous case, our proposed approach gives better result than the other approaches: edges are much clearer and the background noise is well reduced. The profile shown in Figure 5 (g) illustrates furthermore that the proposed approach leads to a reconstruction with sharper edges. In this case, the Best-AnisoTV-MAP obtains the closest image with sharp edges to the proposed approach. However, even with quite sharp edges, the edge positions are much less correctly estimated for rightmost high spatial frequency bands. Moreover, we can see that the quality of reconstruction is poorer when the width of the object is smaller, e.g. the leftmost band which is the widest is relatively well reconstructed: the object value is quite close to the true value and positions of edges are very precise. In the opposite, for the rightmost band which has the smallest width, the object value is underestimated and the

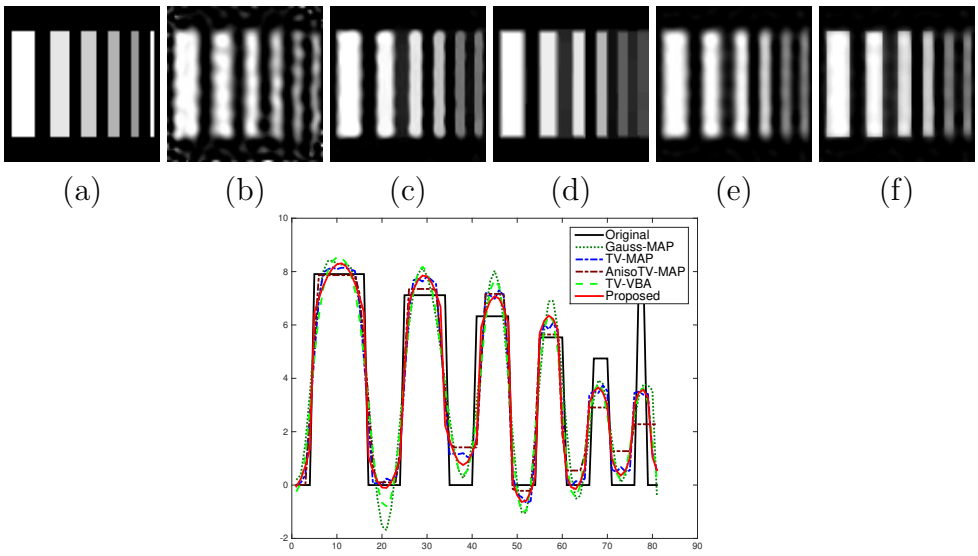


Figure 5: (a) Original image, reconstructed images with (b) Best-Gauss-MAP [12], (c) Best-TV-MAP [10], (d) Best-AnisoTV-MAP, (e) TV-VBA [3], (f) proposed approach, (g) profile of middle rows of the above images.

edges are quite smoothed. This result can be explained by the fact that narrower bands contain higher spatial frequency components, the reconstruction of which is much more difficult.

We also compare the reconstruction at different noise level. We summarise PSNR, SSIM (Structural SIMilarity), MAE (Mean Average Error) values of the reconstructions as well as CPU time used by five approaches for comparison in Table 1. For Best-Gauss-MAP, Best-TV-MAP and Best-AnisoTV-MAP, only results at SNR levels of 15 dB and 30 dB are shown in Table 1. Since Best-Gauss-MAP, Best-TV-MAP and Best-AnisoTV-MAP are supervised approaches, the regularization parameter needs to be manually tuned each time. As a result, it is computationally demanding to test all the datasets.

Moreover, the results for TV-VBA and the proposed approach reported in Table 1 are average values over 30 simulations with 30 different noise realizations at each SNR level. Results of 30 noise realizations are summarized as boxplots in Figure 6. We can see from Table 1 that for all the tested

Table 1: Performance Comparison of Best-Gauss-MAP [12], Best-TV-MAP [10], Best-AnisoTV-MAP [10], TV-VBA [3] and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds). In Table, * represents Best-.

		Reconstruction results				
Data	Metrics	*Gauss-MAP	*TV-MAP	*AnisoTV-MAP	TV-VBA	Proposed
15dB	PSNR	11.82	14.37	13.67	13.98	**14.40**
	SSIM	0.29	**0.41**	0.34	0.35	0.38
	MAE	1.51	**0.92**	1.19	1.14	1.06
	Time (s.)	30	79.6	400	6.4	7.8
20dB	PSNR	-	-	-	14.48	**14.97**
	SSIM	-	-	-	0.38	**0.40**
	MAE	-	-	-	1.06	**0.94**
	Time (s.)	-	-		5.1	8.0
25dB	PSNR	-	-	-	14.74	**15.55**
	SSIM	-	-	-	0.40	**0.44**
	MAE	-	-	-	1.02	**0.85**
	Time (s.)	-	-	-	4.0	5.3
30dB	PSNR	13.85	15.59	14.87	14.99	**16.23**
	SSIM	0.41	0.51	**0.55**	0.42	0.50
	MAE	1.18	0.76	0.81	0.96	**0.71**
	Time (s.)	25	93.5	209	5.6	8.5
35dB	PSNR	-	-	-	15.22	**16.58**
	SSIM	-	-	-	0.43	**0.54**
	MAE	-	-	-	0.93	**0.63**
	Time (s.)	-	-	-	6.8	9.1
40dB	PSNR	-	-	-	15.41	**16.81**
	SSIM	-	-	-	0.45	**0.55**
	MAE	-	-	-	0.89	**0.60**
	Time (s.)	-	-	-	8.6	4.6

data, it is the proposed approach that leads to the highest PSNR value. Averagely, PSNR values of the proposed approach is 0.95 dB greater than those obtained by TV-VBA. At the SNR level of 15 dB, the PSNR of the proposed approach is 2.58 dB higher than that of Best-Gauss-MAP, 0.03 dB higher, therefore nearly equivalent than that of Best-TV-MAP and 0.73 dB higher than that of Best-AnisoTV-MAP. At the SNR level of 30 dB, the proposed approach gains 2.38 dB with respect to Best-Gauss-MAP, 0.64 dB with respect to Best-TV-MAP and 1.36 dB with respect to Best-AnisoTV-MAP. Similar comparison result can be obtained with the metrics SSIM and

MAE except that at SNR level of 30 dB, the Best-AnisoTV-MAP obtains higher SSIM whereas lower PSNR. It is because Best-AnisoTV-MAP tends to gives very flat zone, as a result, similar visual pattern to our synthetic image. However, the approach estimates less accurately the pixel values of flat zones. In addition to the quality metrics, CPU time is also provided in Table 1. We can see that both the proposed approach and the TV-VBA takes less than 10 seconds for all the test cases whereas the other three approaches takes much longer time. We need to mention that the computation time of Best-Gauss-MAP, Best-TV-MAP and Best-AnisoTV-MAP shown in Table 1 is the time for the best regularization parameter. Considering the search of regularization parameter, the computation time will be much more longer in practice.

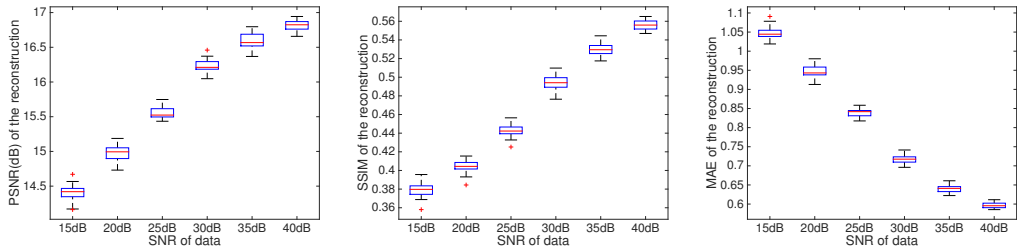


Figure 6: Results of 30 noise realizations of each noise level are summarized as boxplots.

Since the Best-Gauss-MAP is computationally demanding and has been shown to have much poorer performance than the proposed approach, we will not show its results in the following experiments.

Discussion on the hyperparameter ξ and initializations The proposed approach automatically estimate most hyperparameters but ξ is a parameter introduced to incorporate prior information on the proportion of edges within the target image. In all of our experiments, ξ has been set to $0.1N$ and this choice has led to good results. In this part, we make a discussion on the influence of ξ on the quality of the reconstruction results through experiments. We show in Figure 7 the statistics of the results through boxplots when ξ takes values between $[0, 0.2N]$. We can see that when data is not very noisy (SNR$= 30, 35, 40$dB), the results are very stable with respect to the change of the parameter ξ. When data is more noisy

(SNR= $15, 20, 25$dB), results have larger variations compared to less noisy cases.

We made moreover an exploration on the influence of the initialization of the target image to reconstruction results of the proposed approach. We show in Figure 8 boxplots obtained with two common choices of initializations: In the first case, we consider $\mathbf{m}_0 = \mathbf{A}^T\mathbf{y}$ which is used as the default initialization for all experiments shown in this paper and $\mathbf{m}_0 = 0$ is the second classical possible initialization studied. We can see that these two initializations lead to very similar results.

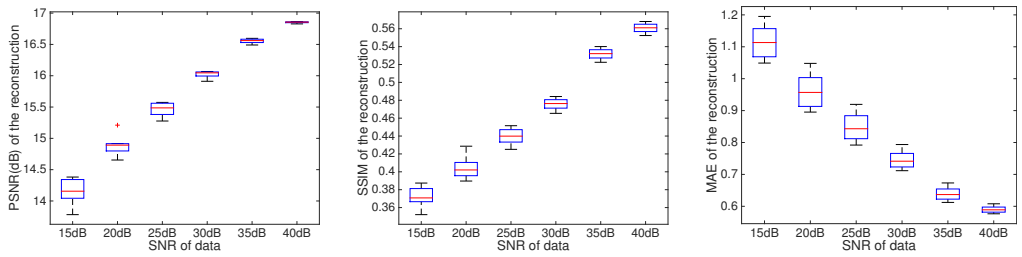


Figure 7: Boxplots of PSNR, SSIM, MAE of the reconstructions while varying ξ between $[0, 0.2N]$ (equally spaced by $0.01N$).

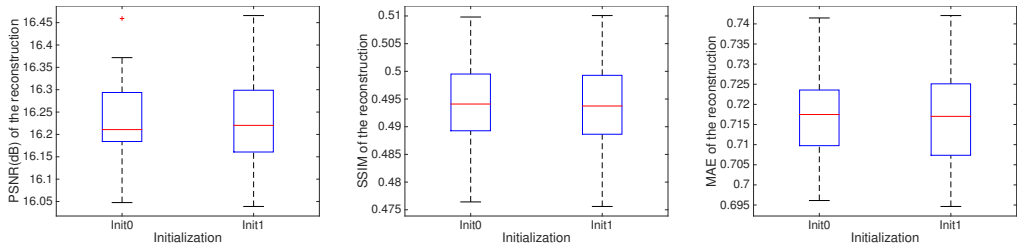
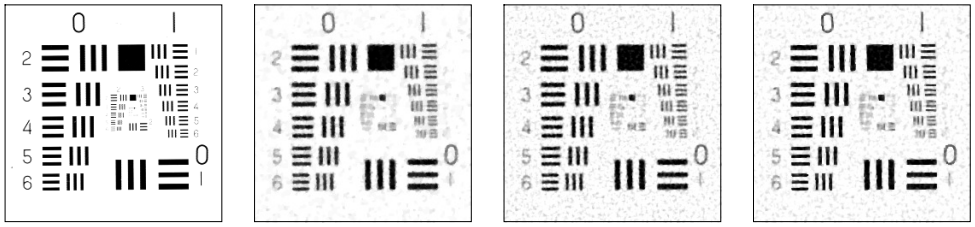


Figure 8: Boxplots of PSNR, SSIM, MAE of the reconstructions while varying the initialization of the images for data of SNR= $30dB$ (30 noise realizations). Init0: $\mathbf{m}_0 = \mathbf{A}^T\mathbf{y}$, Init1: $\mathbf{m}_0 = 0$.

4.5 Evaluation on natural images

In above simulations, images composed of only homogeneous objects have been used. In these cases, our prior promoting piecewise smoothness is well adapted to the problem. In order to study the robustness of the proposed approach with respect to variable structures, we show in this section simulation results on the reconstruction of three natural images composed of different structures, as shown in Figure 9-11 (a). The first ones are classical tests images, given by *M*ire and *L*ena. We show in Figure 11 (a) an image of Printed Circuit Board (PCB) since the quality control of PCB is a typical application of non-destructive testing in industry. To evaluate the capability of our approach on larger size images such as Figure 9-11 (a) (256×256), in this section, we simulate the data using the Radon transform with parallel-beam geometry. The data corresponds to projections collected at 270 angles uniformly distributed on $[0, \pi]$. Each projection is collected by 367 detection cells. In this case, the associated inverse problem is very high-dimensional (65536 unknowns and the forward matrix H has more than six billons entries).



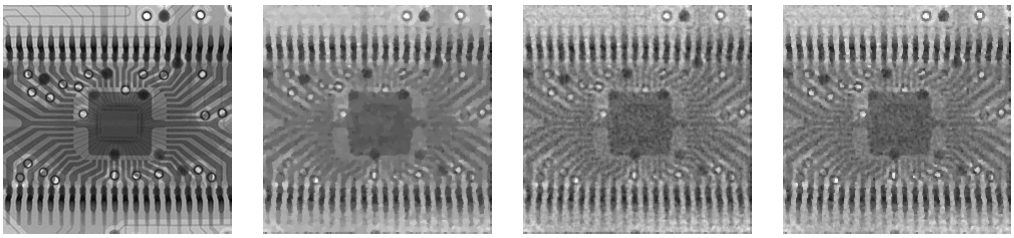
(a) (b) (c) (d)

Figure 9: (a) Original image of dimension 256×256, reconstructed images with (b) Best-TV-MAP [10], (c) TV-VBA [3] (d) proposed approach.

Reconstruction quality evaluation We show in Figure 9 (b)-(d) reconstructions obtained by Best-TV-MAP, TV-VBA and the proposed approach from data generated from the image *Mire* shown in Figure 9 (a) at SNR level of 30 dB. We can see that in this problem, the proposed approach works still better than Best-TV-MAP and TV-VBA in reconstructing regular objects (bar shape objects in Figure 9). For less regular parts like the numbers, the

(a) (b) (c) (d)

Figure 10: (a) Original image of dimension 256×256, reconstructed images with (b) Best-TV-MAP [10], (c) TV-VBA [3] (d) proposed approach.



(a) (b) (c) (d)

Figure 11: (a) Original image of dimension 256×256, reconstructed images with (b) Best-TV-MAP [10], (c) TV-VBA [3] (d) proposed approach.

proposed approach gives good reconstruction as well. In Figure 10, reconstructions of image *Lena* obtained by the compared approaches are given. Figure 10 (d) shows the reconstruction obtained by the proposed approach. We can see that this reconstruction has a quite good quality, even though a little staircasing effect at image edges. This fact suggests that the proposed approach is robust for natural images where the piecewise smoothness hypothesis is not totally satisfied. As shown in Figure 10 (b), Best-TV-MAP gives a reconstruction of little noise but with visible staircasing effect which do not appear in the reconstruction of TV-VBA (see Figure 10 (c)). These results are consistent with the analysis given in [29] which shows that the TV based posterior mean estimate does not suffer much from the staircasing whereas the Best-TV-MAP estimate generally encounters this problem. Compared to Best-TV-MAP and TV-VBA, our approach gives a compromise between the noise reduction and the staircasing effect. Similarly for

PCB image, the proposed approach gives clearer edges for the transmission lines than the TV-VBA. The Best-TV-MAP obtains an image with much less noise since in this case, the regularization parameter has been manually set to large value but the proposed approach has automatically estimated the hyperparameters.

Table 2: Performance Comparison of Best-TV-MAP [10], Best-AnisoTV-MAP [10], TV-VBA [3] and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds) for Mire. In Table, * represents Best-.

		Reconstruction results				
Data	Metrics	*TV-MAP	*AnisoTV-MAP	TV-VBA	Proposed	*Best-Proposed*
15dB	PSNR	-	-	11.09	**12.22**	-
	SSIM	-	-	0.57	**0.58**	-
	MAE	-	-	0.18	**0.15**	-
	Time (s.)	-	-	52	**39**	-
20dB	PSNR	11.97	12.34	11.28	**13.06**	*13.5*
	SSIM	0.36	0.34	0.57	**0.61**	*0.59*
	MAE	0.15	0.14	0.17	**0.12**	*0.13*
	Time (s.)	540	1100	40	**37**	*22*
25dB	PSNR	-	-	16.47	**17.83**	-
	SSIM	-	-	0.53	**0.62**	-
	MAE	-	-	0.09	**0.07**	-
	Time (s.)	-	-	50	**46**	-
30dB	PSNR	19.33	19.41	19.24	**20.68**	*23.46*
	SSIM	0.71	0.71	0.70	0.71	*0.73*
	MAE	0.06	0.06	0.07	**0.05**	*0.03*
	Time (s.)	480	1200	**30**	36	18
35dB	PSNR	-	-	21.55	**22.96**	-
	SSIM	-	-	0.71	**0.72**	-
	MAE	-	-	0.05	**0.04**	-
	Time (s.)	-	-	45	**37**	-
40dB	PSNR	-	-	24.21	**25.77**	-
	SSIM	-	-	0.73	**0.75**	-
	MAE	-	-	0.04	**0.03**	-
	Time (s.)	-	-	**24**	33	-

In order to have a numerical comparison of the proposed approach with Best-TV-MAP, Best-AnisoTV-MAP and TV-VBA, we summarise the PSNR, SSIM, MAE of reconstructed images in Table 2-4. Again, since the manual tuning of regularization parameter for Best-TV-MAP and Best-AnisoTV-

Table 3: Performance Comparison of Best-TV-MAP [10], Best-AnisoTV-MAP [10], TV-VBA [3] and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds) for Lena. In Table, * represents Best-.

		Reconstruction results				
Data	Metrics	*TV-MAP	*AnisoTV-MAP	TV-VBA	Proposed	*Best-Proposed*
15dB	PSNR	-	-	15.25	**18.32**	-
	SSIM	-	-	0.43	**0.48**	-
	MAE	-	-	0.14	**0.09**	-
	Time (s.)	-	-	**29**	32	-
20dB	PSNR	**22.69**	22.41	20.45	21.04	*22.70*
	SSIM	**0.60**	0.58	0.54	0.57	*0.57*
	MAE	**0.051**	0.053	0.07	0.06	*0.05*
	Time (s.)	350	460	**32**	36	*13*
25dB	PSNR	-	-	23.59	**23.98**	-
	SSIM	-	-	0.64	**0.67**	-
	MAE	-	-	0.044	**0.042**	-
	Time (s.)	-	-	**28**	34	-
30dB	PSNR	**25.90**	25.58	25.49	25.78	*27.28*
	SSIM	**0.72**	**0.72**	0.67	0.70	*0.77*
	MAE	**0.033**	0.034	0.037	0.035	*0.028*
	Time (s.)	320	360	**26**	31	*15*
35dB	PSNR	-	-	27.01	**27.11**	-
	SSIM	-	-	0.71	**0.74**	-
	MAE	-	-	0.032	**0.031**	-
	Time (s.)	-	-	**29**	37	-
40dB	PSNR	-	-	**28.77**	**28.77**	-
	SSIM	-	-	0.79	**0.80**	-
	MAE	-	-	**0.026**	**0.026**	-
	Time (s.)	-	-	34	**25**	-

MAP is time-consuming, we report here only the results for two noise levels. In most cases, for *Mire* image which is still quite regular, the proposed approach gives the best results in terms of PSNR, SSIM and MAE. For image *Lena*, the reconstructions of the proposed approach are comparable with TV-VBA. However, in the case where SNR = 20 dB, Best-TV-MAP and Best-AnisoTV-MAP lead to higher PSNR. This is due to the regularization parameter of these two approaches which is manually tuned to give the largest PSNR values. In very noisy cases, oversmoothed solution has generally higher PSNR. This point is further supported by the results shown in

Table 4: Performance Comparison of Best-TV-MAP [10], Best-AnisoTV-MAP [10], TV-VBA [3] and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds) for PCB. In Table, * represents Best-.

		Reconstruction results				
Data	Metrics	*TV-MAP	*AnisoTV-MAP	TV-VBA	Proposed	*Best-Proposed*
15dB	PSNR	-	-	17.12	**18.53**	-
	SSIM	-	-	0.21	**0.25**	-
	MAE	-	-	0.11	**0.09**	-
	Time (s.)	-	-	**31**	38	-
20dB	PSNR	**19.94**	19.43	18.84	19.60	*21.29*
	SSIM	**0.41**	0.37	0.33	0.36	*0.44*
	MAE	0.08	0.08	0.08	**0.07**	*0.06*
	Time (s.)	550	534	**30**	33	20
25dB	PSNR	-	-	22.91	**23.91**	-
	SSIM	-	-	0.59	**0.61**	-
	MAE	-	-	0.05	**0.04**	-
	Time (s.)	-	-	**35**	39	-
30dB	PSNR	**25.76**	25.47	25.02	25.60	*26.64*
	SSIM	**0.74**	**0.74**	0.68	0.69	*0.77*
	MAE	0.04	0.04	0.04	0.04	*0.03*
	Time (s.)	355	379	**28**	32	17
35dB	PSNR	-	-	27.32	**27.44**	-
	SSIM	-	-	0.77	0.77	-
	MAE	-	-	0.03	0.03	-
	Time (s.)	-	-	**25**	33	-
40dB	PSNR	-	-	**29.83**	29.67	-
	SSIM	-	-	0.84	0.84	-
	MAE	-	-	0.03	0.03	-
	Time (s.)	-	-	**26**	33	-

the row of Best-Proposed where instead of estimating the hyperparameters, we manually tune them to get the highest PSNR. We can see that in two noise level cases shown here, the Best-Proposed leads to the highest PSNR. Similar conclusion can be drawn from the results of PCB images.

Moreover, computation time can also be found in Table 2-4. Our proposed method is based on a fast variational Bayesian method which outperforms Best-TV-MAP and Best-AnisoTV-MAP in term of computation time and is quite competitive with TV-VBA even while estimating more parameters (d).

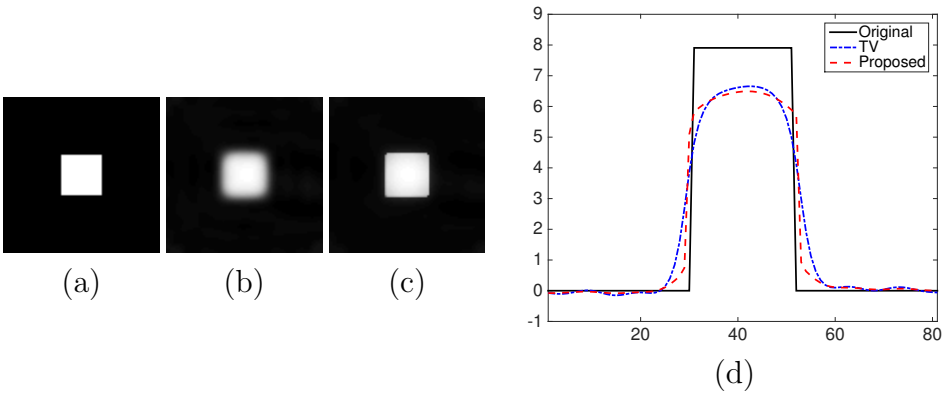


Figure 12: (a) Original image, reconstructed images with (b) TV-VBA [3] (c) proposed approach, (d) profile of middle rows of the above images.

4.6 Robustness to model errors

As stated above, the proposed approach uses a linear approximation of the non-linear forward model for NDT, which causes consequently model errors. Here, we evaluate the robustness of the proposed approach with respect to the model errors by using data generated from the non-linear forward model (see (51) and (52)). We simulate data with the simple image, Figure 12 (a) as the object and add a same level of Gaussian noise as the first experiment, i.e. SNR is equal to 30 dB. We show in Figure 12 (b) and (c) reconstructions obtained by TV-VBA and the proposed approach. Comparing to Figure 4 (d) and (e), we can see that the extra model errors lead to a decrease in reconstruction quality. However, the proposed approach manages to give a reconstruction with still a clear edge (Figure 12 (c)) which is not the case for TV-VBA (Figure 12 (b)). Profiles of the reconstructions are shown in Figure 12 (d). It shows again that the proposed approach gives a sharper edge than TV-VBA. But in this case, the value of the object is underestimated. Actually, the proposed approach uses a prior promoting sparsity information on the image gradients, as a result, sharp edges are correctly restored. Nevertheless, no prior information on the raw pixel values has been introduced. As a result, in such a badly ill-posed inverse problem with additional model errors, the proposed approach failed to estimate pixel values accurately.

4.7 Evaluation on image denoising problems

The proposed method is developed for general linear inverse problem based on observation model (1). By setting different observation operator $\mathbf{A}$, our proposed approach can be applied to solve different inverse problems. In the previous section, we have shown that our approach can be successfully applied to tomography problem. To show the versatility of the proposed approach, we show in this section some simulation results on the classical image denoising problem. In such a case, the observation operator is simply an identity matrix.

Synthetic noisy data are generated by adding Gaussian additive noise at five different SNR levels: 15dB, 20dB, 25dB, 30dB, 35dB to three images: Mire, Lena and PCB images (as shown in Figure 9 (a), Figure 10(a), Figure 11(a)). Several reconstruction results are included for comparison: TV-VBA [1], a nonlocal means (NL-means) based denoising method [21, 22] and one deep learning based denoising method (DL-denoising) [50]. For NL-means method, its performance depends on the estimation of the noise variance. However, in practice, the noise variance is unknown. To make a quite fair comparison, in this simulation, we randomly take values within $\pm 50\%$ of the true noise variance as estimations of the noise variance and report the mean reconstruction results of 30 trials. Regarding the DL-denoising, a pretrained DnCNN network is used.

The results are summarized in Table 5 - Table 7. In general, the proposed approach compare favorably with TV-VBA for the images Mire and PCB since these two images are more piecewise smooth. For the image Lena, the proposed approach gives comparable results to TV-VBA. Moreover, numerical results in Table 5 - Table 7 show that the proposed approach performs better than the tested NL-means approach. As regards DL-denoising, it gives amazing results in high noisy cases: SNR $= 15, 20$dB. This is not surprising since convolutional neural network has been shown to perform very well in various image processing tasks and the pretrained DnCNN network has been specifically trained for Gaussian denoising with hundreds to thousands of training images. However, we can remark that in less noisy cases (e.g. SNR $= 35, 40$ dB), the pretrained network performs dramatically bad. We have shown also the computation time in Table 5 - Table 7. We can see that the

[1]Since we have made a lot of comparison with three different TV approaches, in this section, we will choose to compare only with TV-VBA which is an unsupervised approach as our proposed method.

proposed approach is very fast, comparable to TV-VBA, and about 10 times faster than NL-means and about 30 times faster than DL-denoising.

Table 5: Performance Comparison of NL-means, DL-denoising, TV-VBA [3] and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds) on Mire.

		Reconstruction results			
Data	Metrics	NL-means	DL-denoising	TV-VBA	Proposed
15dB	PSNR	18.33	26.13	19.47	20.15
	SSIM	0.66	0.86	0.40	0.41
	MAE	0.09	0.03	0.08	0.07
	Time (s.)	11.9	35.1	1.1	1.1
20dB	PSNR	22.81	29.34	23.88	24.82
	SSIM	0.78	0.92	0.54	0.55
	MAE	0.05	0.02	0.05	0.04
	Time (s.)	11.9	33.5	1.2	1.3
25dB	PSNR	28.04	32.28	28.37	29.50
	SSIM	0.87	0.95	0.70	0.72
	MAE	0.023	0.014	0.028	0.026
	Time (s.)	11.5	32.8	1.3	1.4
30dB	PSNR	31.49	34.93	32.99	34.11
	SSIM	0.93	0.96	0.84	0.86
	MAE	0.014	0.011	0.017	0.015
	Time (s.)	10.9	32.9	1.5	1.5
35dB	PSNR	36.15	36.65	37.46	38.64
	SSIM	0.95	0.95	0.93	0.94
	MAE	0.01	0.01	0.01	0.009
	Time (s.)	10.2	32.8	1.6	1.7
40dB	PSNR	41.5	38.4	42.57	43.13
	SSIM	0.98	0.96	0.97	0.98
	MAE	0.005	0.008	0.006	0.005
	Time (s.)	9.8	32.9	1.7	1.8

4.8 Discussion

This section presents the results of our approach for multiple configurations. We have shown through simulations that our approach is consistently superior to the other unsupervised TV-VBA approach. In order to have a comparison with more approaches, we also considered supervised algorithms, where hyperparameters are manually tuned as optimal ones in the sense of

Table 6: Performance Comparison of NL-means, DL-denoising, TV-VBA [3] and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds) on Lena.

		Reconstruction results			
Data	Metrics	NL-means	DL-denoising	TV-VBA	Proposed
15dB	PSNR	22.87	28.74	24.84	24.90
	SSIM	0.57	0.80	0.55	0.56
	MAE	0.060	0.026	0.045	0.045
	Time (s.)	10.7	37.7	0.9	1.1
20dB	PSNR	27.37	31.24	28.95	29.06
	SSIM	0.74	0.86	0.74	0.75
	MAE	0.034	0.020	0.028	0.027
	Time (s.)	10.6	36.1	1.4	1.4
25dB	PSNR	30.82	33.54	32.86	33.00
	SSIM	0.84	0.90	0.87	0.88
	MAE	0.023	0.015	0.018	0.017
	Time (s.)	10.5	37.8	1.6	1.6
30dB	PSNR	35.40	35.55	36.85	36.94
	SSIM	0.91	0.93	0.94	0.94
	MAE	0.013	0.012	0.011	0.011
	Time (s.)	10.2	36.9	1.8	1.8
35dB	PSNR	39.42	37.44	41.07	41.11
	SSIM	0.96	0.95	0.97	0.98
	MAE	0.008	0.009	0.007	0.007
	Time (s.)	9.7	36.4	2.1	2.1
40dB	PSNR	42.74	38.63	45.58	45.59
	SSIM	0.98	0.96	0.99	0.99
	MAE	0.006	0.007	0.004	0.004
	Time (s.)	9.5	36.1	2.4	2.3

PSNR. For this, the real image needs to be known. The simulation results show that in the case of NDT-type images, our model prior is very relevant, and our approach gives the best reconstructions. Moreover, our approach is extremely robust. This is partly explained by the choice of the estimator of the posterior mean which minimizes the quadratic risk, and by the estimation of the hyperparameters. This robustness has been shown in different configurations: when the prior model is less relevant as in the case of Lena (see Figure 10, Table 3) or when an important model error has been introduced by simulating the problem with a nonlinear model and reconstructing with a linear model (see Figure 12). Compared to a learning-based denoising

Table 7: Performance Comparison of TV-VBA [3], NL-means, DL-denoising and our Proposed approach in Terms of PSNR (dB), SSIM, MAE and CPU time (in seconds) on PCB.

		Reconstruction results			
Data	Metrics	NL-means	DL-denoising	TV-VBA	Proposed
15dB	PSNR	22.91	27.82	23.96	24.87
	SSIM	0.61	0.84	0.68	0.72
	MAE	0.06	0.03	0.05	0.045
	Time (s.)	8.8	26.1	1.1	1.3
20dB	PSNR	25.36	30.88	27.77	28.9
	SSIM	0.69	0.91	0.82	0.86
	MAE	0.045	0.022	0.03	0.028
	Time (s.)	8.6	26.3	1.5	1.5
25dB	PSNR	31.21	33.20	31.60	32.72
	SSIM	0.88	0.94	0.90	0.93
	MAE	0.022	0.017	0.02	0.018
	Time (s.)	8.5	26.2	1.8	1.9
30dB	PSNR	34.43	36.51	36.10	36.90
	SSIM	0.94	0.97	0.97	0.97
	MAE	0.015	0.012	0.012	0.011
	Time (s.)	7.9	26.1	2.1	2.3
35dB	PSNR	39.41	40.87	40.92	41.2
	SSIM	0.98	0.98	0.99	0.99
	MAE	0.008	0.007	0.007	0.007
	Time (s.)	7.6	26.2	2.2	2.3
40dB	PSNR	43.05	45.32	45.53	45.68
	SSIM	0.99	0.99	0.99	0.99
	MAE	0.005	0.004	0.004	0.004
	Time (s.)	7.7	26.3	2.5	2.4

approach, the proposed solution still gives good results without learning base (see Table 5-7). Finally, Figure 7 and 8 clearly showed that our approach was unsupervised and that the parameters of the method such as the initialization of the algorithm or the fixed parameter ξ do not significantly change the quality of the reconstructed images.

5 Conclusion

In this paper, we proposed a hierarchical sparse gradient prior and its Bayesian application to linear inverse problems. The main interest of this prior is that it models well smooth images with sparse edges. Furthermore, non-informative Jeffreys' priors are employed for hyperparameters which leads to unsupervised approaches. In this case the estimation of hyperparameters is very complicated due to the non linearity of the problem and the strong correlations between these parameters. We have proposed an original estimation approach based on a different parameterization of the problem. Furthermore the memory gradient subspace based variational Bayesian approximation method is employed to allow us to obtain fast and efficient estimations of parameters. Simulation results have shown that our approach gives good performances in general and outperforms the total variation based approaches in reconstructing piecewise smooth images. We have also show that this method is very efficient when it is considered in its optimal use condition but still remains robust to model or image errors.

A The partition function $Z_{\mathbf{x}|\mathbf{b}}$

We compute the partition function $Z_{\mathbf{x}|\mathbf{b}}$ given by (4) by transposing the integral into the Fourier domain.

Let us denote the convolution kernel corresponding to matrix $\mathbf{D}_h$ and $\mathbf{D}_v$ by $\mathbf{d}_h$, $\mathbf{d}_v$, respectively. Moreover, we use $*$ to denote the convolution operator. Then we have $\mathbf{D}_h\mathbf{x} = \mathbf{d}_h * \mathbf{x}$ and $\mathbf{D}_v\mathbf{x} = \mathbf{d}_v * \mathbf{x}$. In the following, $\overset{\circ}{\mathbf{a}}$ represents the Fourier transform of the vector $\mathbf{a}$. Using the Parseval's

theorem, we can obtain

$$
\begin{aligned}
&\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2+\|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2\\
=&\|\overset{\circ}{\mathbf{d}}_h\circ\overset{\circ}{\mathbf{x}}-\overset{\circ}{\mathbf{b}}_h\|^2+\|\overset{\circ}{\mathbf{d}}_v\circ\overset{\circ}{\mathbf{x}}-\overset{\circ}{\mathbf{b}}_v\|^2\\
=&\sum_i\left[(\overset{\circ}{\mathbf{d}}_h)_i(\overset{\circ}{\mathbf{x}})_i-(\overset{\circ}{\mathbf{b}}_h)_i\right]^2+\left[(\overset{\circ}{\mathbf{d}}_v)_i(\overset{\circ}{\mathbf{x}})_i-(\overset{\circ}{\mathbf{b}}_v)_i\right]^2\\
=&\sum_i\left[(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2\right](\overset{\circ}{\mathbf{x}})_i^2+(\overset{\circ}{\mathbf{b}}_h)_i^2+(\overset{\circ}{\mathbf{b}}_v)_i^2\\
&-2\left[(\overset{\circ}{\mathbf{d}}_h)_i(\overset{\circ}{\mathbf{b}}_h)_i+(\overset{\circ}{\mathbf{d}}_v)_i(\overset{\circ}{\mathbf{b}}_v)_i\right](\overset{\circ}{\mathbf{x}})_i\\
=&\sum_i\left[(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2\right]\left[(\overset{\circ}{\mathbf{x}})_i-\frac{(\overset{\circ}{\mathbf{d}}_h)_i(\overset{\circ}{\mathbf{b}}_h)_i+(\overset{\circ}{\mathbf{d}}_v)_i(\overset{\circ}{\mathbf{b}}_v)_i}{(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2}\right]^2\\
&+\frac{\left((\overset{\circ}{\mathbf{d}}_h)_i(\overset{\circ}{\mathbf{b}}_v)_i-(\overset{\circ}{\mathbf{d}}_v)_i(\overset{\circ}{\mathbf{b}}_h)_i\right)^2}{(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2}
\end{aligned}
\tag{54}
$$

Therefore, the partition function $Z_{\mathbf{x}|\mathbf{b}}$ can be easily obtained in the Fourier transform thanks to a change of variable

$$
\begin{aligned}
Z_{\mathbf{x}|\mathbf{b}}=&\int\exp\left[-\frac{\gamma_d}{2}\left(\|\mathbf{D}_h\mathbf{x}-\mathbf{b}_h\|^2+\|\mathbf{D}_v\mathbf{x}-\mathbf{b}_v\|^2\right)\right]\mathrm{d}\mathbf{x}\\
=&(2\pi)^{N/2}\gamma_d^{-N/2}\prod_i\left[(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2\right]^{-1/2}\\
&\times\exp\left[-\frac{\gamma_d}{2}\frac{\left((\overset{\circ}{\mathbf{d}}_h)_i(\overset{\circ}{\mathbf{b}}_v)_i-(\overset{\circ}{\mathbf{d}}_v)_i(\overset{\circ}{\mathbf{b}}_h)_i\right)^2}{(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2}\right]\\
=&c{\gamma_d}^{-N/2}\exp\left[-\frac{\gamma_d}{2}\sum_i^N\frac{\left((\overset{\circ}{\mathbf{d}}_h)_i(\overset{\circ}{\mathbf{b}}_v)_i-(\overset{\circ}{\mathbf{d}}_v)_i(\overset{\circ}{\mathbf{b}}_h)_i\right)^2}{(\overset{\circ}{\mathbf{d}}_h)_i^2+(\overset{\circ}{\mathbf{d}}_v)_i^2}\right]
\end{aligned}
\tag{55}
$$

where c encloses all factors independent of γ_d and $\overset{\circ}{\mathbf{b}}_h$, $\overset{\circ}{\mathbf{b}}_v$. We can easily

obtain

$$Z_{\mathbf{x}|\mathbf{b}} \leq c{\gamma_d}^{-N/2} \tag{56}$$

since the rest of the factor is strictly smaller than one.

References

[1] H. Attias. A variational Bayesian framework for graphical models. *Adv. Neural Inf. Process. Syst.*, 12(1-2):209–215, 2000.

[2] H. Ayasso and A. Mohammad-Djafari. Joint NDT image restoration and segmentation using Gauss–Markov–Potts prior models and variational Bayesian computation. *IEEE Trans. Image Process.*, 19(9):2265–2277, 2010.

[3] S.D. Babacan, R. Molina, and A.K. Katsaggelos. Variational Bayesian super resolution. *IEEE Trans. Image Process.*, 20(4):984–999, 2011.

[4] M. J. Beal and Z. Ghahramani. The variational Bayesian em algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting*, page 453, 2002.

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

[6] T. Chan, S. Esedoglu, F. Park, and A. Yip. Recent developments in total variation image restoration. *Mathematical Models of Computer Vision*, 17(2), 2005.

[7] H. Chang, Y. Lou, Y. Duan, and S. Marchesini. Total variation–based phase retrieval for Poisson noise removal. *SIAM Journal on Imaging Sciences*, 11(1):24–55, 2018.

[8] R. A. Choudrey. *Variational Methods for Bayesian Independent Component Analysis*. PhD thesis, University of Oxford, 2002.

[9] E. Chouzenoux, A. Jezierska, J. C. Pesquet, and H. Talbot. A majorize-minimize subspace approach for l2-l0 image regularization. *SIAM J. Imaging Sci.*, 6(1):563–591, 2013.

[10] P. L. Combettes and J.-C. Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Valued and variational analysis*, 20(2):307–330, 2012.

[11] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

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

[13] M. Dumitru, L. Wang, A. Mohammad-Djafari, and N. Gac. Model selection in the sparsity context for inverse problems in bayesian framework. In *International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, pages 155–165. Springer, 2017.

[14] O. Féron, B. Duchêne, and A. Mohammad-Djafari. Gauss-Markov-Potts priors for Bayesian inversion in microwave imaging, 2010.

[15] A. Fraysse and T. Rodet. A measure-theoretic variational Bayesian algorithm for large dimensional problems. *SIAM J. Imaging Sci.*, 7(4):2591–2622, 2014.

[16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

[17] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Process.*, 4(7):932–946, July 1995.

[18] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, (6):721–741, 1984.

[19] Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. *Adv. Neural Inf. Process. Syst.*, pages 507–513, 2001.

[20] L. Gharsalli, H. Ayasso, B. Duchêne, and A. Mohammad-Djafari. Variational Bayesian inversion for microwave breast imaging. *Computer Assisted Methods in Engineering and Science*, 21(3/4):199–210, 2017.

[21] G. Gilboa and S. Osher. Nonlocal linear image regularization and supervised segmentation. *Multiscale Modeling & Simulation*, 6(2):595–630, 2007.

[22] G. Gilboa and S. Osher. Nonlocal operators with applications to image processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2008.

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

[24] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, May 1979.

[25] P. C. Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev.*, 34:561–580, 1992.

[26] D. M. Higdon, J. E. Bowsher, V. E. Johnson, T. G. Turkington, D. R. Gilland, and R. J. Jaszczak. Fully Bayesian estimation of Gibbs hyperparameters for emission computed tomography data. *IEEE Trans. Med. Imag.*, 16(5):516–526, oct 1997.

[27] F. C. Jeng and J. W. Woods. Compound Gauss-Markov random fields for image estimation. *IEEE Trans. Signal Process.*, 39(3):683–697, 1991.

[28] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[29] C. Louchet and L. Moisan. Total variation denoising using posterior expectation. In *European Signal Processing Conference (EUSIPCO)*, 2008.

[30] X. Lu, Y. Yuan, and P. Yan. Sparse coding for image denoising using spike and slab prior. *Neurocomputing*, 106:12–20, 2013.

[31] Y. Marnissi, Y. Zheng, E. Chouzenoux, and J.-C. Pesquet. A variational bayesian approach for image restorationapplication to image deblurring with poisson–gaussian noise. *IEEE Transactions on Computational Imaging*, 3(4):722–737, 2017.

[32] M. T. McCann, M. Nilchian, M. Stampanoni, and M. Unser. Fast 3d reconstruction method for differential phase contrast x-ray ct. *Opt. exp.*, 24(13):14564–14581, 2016.

[33] R. Molina. On the hierarchical Bayesian approach to image restoration: Application to astronomical images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(11):1122–1128, nov 1994.

[34] J. Nocedal and S.J. Wright. *Numerical Optimization*. Series in Operations Research. Springer Verlag, New York, 2000.

[35] J. P. Oliveira, J. M. Bioucas-Dias, and M. Figueiredo. Adaptive total variation image deblurring: a majorization–minimization approach. *Signal Processing*, 89(9):1683–1693, 2009.

[36] M. Pereyra, P. Schniter, E. Chouzenoux, J-C. Pesquet, J-Y. Tourneret, A. Hero, and S. McLaughlin. A survey of stochastic simulation and optimization methods in signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):224–241, 2016.

[37] C. P. Robert and G. Casella. *Monte-Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York, 2000.

[38] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[39] A. Sawatzky, C. Brune, F. Wubbeling, T. Kosters, K. Schafers, and M. Burger. Accurate EM-TV algorithm in PET with low SNR. In *2008 IEEE Nuclear Science Symposium Conference Record*, pages 5133–5137. IEEE, 2008.

[40] J. G. Serra, J. Mateos, R. Molina, and A. K. Katsaggelos. Parameter estimation in spike and slab variational inference for blind image deconvolution. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1495–1499. IEEE, 2017.

[41] J. h. Shen and T. F. Chan. Mathematical models for local nontexture inpaintings. *SIAM J. Appl. Math.*, 62(3):1019–1043, 2002.

[42] T. Shi and J. Zhu. Online Bayesian passive-aggressive learning. *Journal of Machine Learning Research*, 18(33):1–39, 2017.

[43] Y. Shi and Q. Chang. Efficient algorithm for isotropic and anisotropic total variation deblurring and denoising. *Journal of Applied Mathematics*, 2013, 2013.

[44] V. Šmídl and A. Quinn. *The Variational Bayes Method in Signal Processing*. Springer, 2006.

[45] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[46] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B (Methodological)*, pages 267–288, 1996.

[47] C. R. Vogel and M. E. Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. Image Process.*, 7(6):813–824, 1998.

[48] L. Wang, A. Mohammad-Djafari, N. Gac, and M. Dumitru. Bayesian 3D X-ray computed tomography with a hierarchical prior model for sparsity in Haar transform domain. *Entropy*, 20(12):977, 2018.

[49] T. Wang, K. Nakamoto, H. Zhang, and H. Liu. Reweighted anisotropic total variation minimization for limited-angle CT reconstruction. *IEEE Transactions on Nuclear Science*, 64(10):2742–2760, 2017.

[50] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[51] L. Zhang, W. Wei, Y. Zhang, C. Shen, A. Van den Hengel, and Q. Shi. Cluster sparsity field for hyperspectral imagery denoising. In *European conference on computer vision*, pages 631–647. Springer, 2016.

[52] Q. Zhao, L. Zhang, and A. Cichocki. Bayesian cp factorization of incomplete tensors with automatic rank determination. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1751–1763, 2015.

[53] Y. Zheng, A. Fraysse, and T. Rodet. Efficient variational Bayesian approximation method based on subspace optimization. *IEEE Trans. Image Process.*, 24(2):681–693, 2015.