Wavelet-based hyperparameter estimation for solving inverse problems

Caroline Chaux, Laure Blanc-Féraud

Wavelet-based hyperparameter estimation for solving inverse problems*

Caroline Chaux† and Laure Blanc-Féraud‡

February 10, 2012

Abstract

We consider the problem of image restoration/reconstruction where the acquisition system is modeled by a linear operator with additive Gaussian noise. A variational approach is adopted for image inversion in order to compute a restored/reconstructed image, consisting in minimizing a convex criterion composed of two parts: a data fidelity term (e.g. quadratic) and a regularization term (e.g. ℓ^1-norm) expressed in the wavelet domain. The purpose of this paper is to estimate the regularization hyperparameters (one per subband) based on a Maximum Likelihood (ML) estimator, only knowing the observed data. A difficult task in such estimation is to compute the expectation according to the a posteriori probability as there is no analytical form. This expectation must be approximated numerically by sampling the distribution. However, sampling the a posteriori distribution is a difficult task because of pixel interactions introduced by the linear operator (image acquisition) in the same time as the wavelet transform (regularization). Moreover, the possible different nature (ℓ^2, ℓ^1-norm ...) of the fidelity and regularization terms does not allow to easily process them simultaneously. We show that both operators can be separated using an auxiliary (hidden) variable and splitting the a posteriori probability in two parts which are sampled alternately using MCMC (Gibbs sampling and Metropolis-Hastings). We show the equivalence between both formulations of the a posteriori distribution. Then a gradient method is used to estimate the hyperparameters. Simulation results demonstrate the good performance and behavior of the proposed approach.

1 Introduction

In various areas such as spatial or biomedical imagery, the acquisition system is modeled by a linear operator which links the observed data to the image of interest. The observation model considered in this work and widely used in image recovery is the following:

$$g = Au + n,$$

where g (vector of dimension k) denotes the observed image, degraded by a linear operator A which is assumed to be known and an additive white Gaussian noise $n \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$ of known variance σ^2. The linear operator A can model a blur due to diffraction or defocus or any other stationary blurring process. It can also be a linear operator as Radon transform in tomography reconstruction [1], or model irregular/regular sampling. The noise is assumed to be additive white Gaussian noise which is uselly the case in high level intensity acquisition. Reconstructing

*This work was supported by the Agence Nationale de la Recherche under grant ANR-09-EMER-004-03.

†Caroline Chaux is with Université Paris-Est, LIGM and UMR CNRS 8049, Marne-La-Vallée Cedex 2, France e-mail: caroline.chaux@univ-paris-est.fr.

‡Laure Blanc-Féraud is with MORPHEME I3S/INRIA/IBV, Sophia-Antipolis, France e-mail: Laure.Blanc_Feraud@inria.fr

u from the observed data g is an ill-posed inverse problem which must be regularized [2]. A variational approach can be adopted considering the minimization of a convex criterion composed of two parts: the data fidelity term which is a mean square error due to Gaussian hypothesis on the noise, and a regularizing term. Standard regularizing term is TV semi-norm [3] or ℓ^1-norm on wavelet-type transform [4–9]. We adopt here a regularization in the wavelet transform domain (e.g. a ℓ^1-norm defined on the coefficients). The criterion $J(u)$ to be minimized is then defined by:

$$J(u) = \frac{\|g - Au\|^2}{2\sigma^2} + \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}} \phi_{\mathbf{m}}(F_{\mathbf{m}} u) \tag{1}$$

where $F_{\mathbf{m}}u$ denotes a subband and F represents the global wavelet transform operator. In this work, we consider the case of an orthogonal basis [10]. Functions $\phi_{\mathbf{m}}$ are ℓ^1-norms on wavelet coefficient subbands and ℓ^2-norm on the approximation subband (low resolution image). The resolution level of the decomposition determines the finite number M of subband $\mathbf{m} = 1, ..., M$.

In such variational approach, the knowledge of $\boldsymbol{\lambda} = (\lambda_{\mathbf{m}=1,...,M})_{\mathbf{m}}$ which are called hyperparameters, is of crucial importance for the quality of the result and the question is to know how to fix these parameters from observations. In some applications and when there are only few hyperparameters, they can be empirically fixed by the user by trial and error. However, when the convergence of the $J(u)$ minimization algorithm is slow or when the number of hyperparameters to compute is more than one or two, this solution becomes difficult to apply. Our objective is to propose a method to estimate automatically hyperparameters. Estimate $\boldsymbol{\lambda}$ only knowing a degraded observation g is the difficult problem of hyperparameter estimation in incomplete data [11]. Among the most popular methods, we can cite Maximum Likelihood (ML) approaches [11, 12], EM algorithms (*Expectation Maximization*) [6, 13] or cross-validation methods [14]. More recently, in [8], authors proposed to use Stein principle in order to determine regularization hyperparameters adaptively. An other approach consists of considering the parameters as random variables and thus to associate probability density function and to proceed the estimation using MCMC sampling methods [15, 16].

In this work, we propose to consider $\boldsymbol{\lambda} = (\lambda_{\mathbf{m}})_{\mathbf{m}}$ as deterministic parameters to be estimated, because we do not have any a priori on these parameters and we estimate it in the Maximum Likelihood sense. On the one hand, this allows to exploit the good asymptotic properties of the ML estimator and on the other hand, $\boldsymbol{\lambda}$ being deterministic, to avoid to introduce new hyperparameters which would have been necessary if probabilistic laws have been fixed for $\boldsymbol{\lambda}$. ML estimation of $\boldsymbol{\lambda}$ is a difficult task as the gradient of the likelihood involves expectation according to the a priori and a posteriori probabilities and there are no analytical forms. These expectations must be approximated numerically by sampling these distributions. The main difficulty is to be able to sample the a posteriori distribution because of large pixel interactions introduced by the linear operator A inside the fidelity term in the same time as the regularization realized in the wavelet transform domain [1]. The proposed method allows to sample the a posteriori distribution by splitting it into two parts introducing an auxiliary (hidden) variable. We show that it is equivalent to sample these two distributions where the linear operator A and the wavelet transform $F_{\mathbf{m}}$ are now splitted. Then we can use MCMC (Gibbs sampling and Metropolis-Hastings) in the Fourier domain for he term involving operator A, in the wavelet domain for the term involving operator $F_{\mathbf{m}}$. Finally a gradient method is proceeded to compute the ML hyperparameter estimate.

This paper is organized as follows. Section 2, is the main section that describes the entire process leading to parameter estimation. We show how to use an auxiliary variable to sample a distribution modelising an image, which is, to our knowledge, impossible without the auxiliary variable. Then, the algorithm based on a gradient method is given in Section 3 and the choice of the step size, essential for the good behavior of the algorithm, is discussed. The efficiency of

the proposed approach is shown in Section 4, performing tests on synthetic data. Finally, the paper is concluded in Section 5.

2 Hyperparameter estimation

2.1 Maximum likelihood estimation

In order to proceed to $\boldsymbol{\lambda} = (\lambda_{\mathbf{m}})_{\mathbf{m}=1,...,M}$ parameter estimation in energy (1), this criterion is interpreted in a stochastic setting. It corresponds to the energy of the *a posteriori* probability in a Bayesian framework where the first term corresponds to the energy of the conditional probability $p(g|u)$ given by the noise n. This probability is assumed to be Gaussian

$$p(g|u) = \frac{1}{K_\sigma} \exp\Big(\frac{-\|g - Au\|^2}{2\sigma^2}\Big) \tag{2}$$

where the normalization constant is $K_\sigma = (2\pi)^{k/2}\sigma^k$ with k the number of pixels. The second term of (1), corresponds to the energy of an *a priori* model given by

$$p_{\boldsymbol{\lambda}}(u) = \frac{1}{Z_{\boldsymbol{\lambda}}} \prod_{\mathbf{m}=1}^{M} \exp\left(-\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\right) \tag{3}$$

where the constant $Z_{\boldsymbol{\lambda}}$ is a normalization constant defined by

$$Z_{\boldsymbol{\lambda}} = \int_u \exp\left(-\sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\right)du \tag{4}$$

The variable u is a discrete field of k real random variables which model a 2d image of k pixels. Thus u takes values in $\mathbb{R}^k$, which is the integration domain of (4). F being a projection onto a basis, $\mathrm{Ker}(F) = \{0\}$ and the integral (4) converges assuming that $\exists\, \mathbf{m} \in \{1, ..M\}, \lambda_{\mathbf{m}} \geq \epsilon > 0$, which can be always assumed.

The maximum likelihood estimation of $\boldsymbol{\lambda}$ consists in maximizing $p_{\boldsymbol{\lambda}}(g)$ w.r.t $\boldsymbol{\lambda}$ where $p_\lambda(g)$ is given by inference:

$$p_{\boldsymbol{\lambda}}(g) = \int_u p_{\boldsymbol{\lambda}}(g, u)du = \int_u p(g|u)p_{\boldsymbol{\lambda}}(u)du = \frac{Z_{\sigma,\boldsymbol{\lambda}}(g)}{K_\sigma Z_{\boldsymbol{\lambda}}} \tag{5}$$

where $Z_{\sigma,\boldsymbol{\lambda}}(g) = \int_u \exp\Big(-\frac{\|g-Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)du$.

Then the maximization of $p_{\boldsymbol{\lambda}}(g)$ with respect to $\boldsymbol{\lambda}$ can be computed by a gradient method [11]. Let us derive these steps. We compute each derivative of (5) with respect to each $\lambda_{\mathbf{m}}$. Note that, as it is not possible to demonstrate the convexity of $p_{\boldsymbol{\lambda}}(g)$ with respect to $\boldsymbol{\lambda}$, a gradient algorithm will probably give a local minimum, depending on the initial values of $\boldsymbol{\lambda}$. For a given application in image processing, these parameters are in a given range. Then our purpose is to propose an algorithm which makes more accurate these values for each new observation.

The derivative of (5) with respect to each $\lambda_{\mathbf{m}}$ is (applying the Lebesgue dominate convergence theorem in order to invert derivative and integral operators):

$$\frac{\partial \log p_{\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} = \frac{\partial \log Z_{\sigma,\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} - \frac{\partial \log(Z_{\boldsymbol{\lambda}})}{\partial \lambda_{\mathbf{m}}}$$

and

$$\frac{\partial \log Z_{\sigma,\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} = \frac{-1}{Z_{\sigma,\boldsymbol{\lambda}}(g)} \int_u \phi_{\mathbf{m}}(F_{\mathbf{m}}u) \exp\Big(\frac{-\|g-Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)du$$
$$= -\mathrm{E}_{\sigma,\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] \tag{6}$$
$$\frac{\partial \log(Z_{\boldsymbol{\lambda}})}{\partial \lambda_{\mathbf{m}}} = \frac{-1}{Z_{\boldsymbol{\lambda}}} \int_u \phi_{\mathbf{m}}(F_{\mathbf{m}}u) \exp\big(-\sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\big)du$$
$$= -\mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] \tag{7}$$

The expectation in (6) is defined according to the *a posteriori* law

$$p_{\boldsymbol{\lambda}}(u|g) = \frac{1}{Z_{\sigma,\boldsymbol{\lambda}}(g)} \exp\Big(\frac{-\|g-Au\|^2}{2\sigma^2} - \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big) \tag{8}$$

whereas the one in (7) is defined according to the *a priori* law $p_{\boldsymbol{\lambda}}(u)$ defined in (3).

Following [11], our aim is to apply a gradient type algorithm to estimate $\boldsymbol{\lambda}$ in the ML sense. Then we must be able to compute the gradient

$$\frac{\partial \log p_{\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} = \mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \mathrm{E}_{\sigma,\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] \tag{9}$$

2.2 Difficulties

In order to apply a gradient method, one has to numerically compute the two expectations in (9). The difficulty comes from the second expectation of (9), defined by (6) which explicitely involves the *a posteriori* probability $p_{\boldsymbol{\lambda}}(u|g)$ given in (8). Actually, this expectation is not analytically computable and we have to estimate it by an empirical mean, computed on samples generated along with this law. However, sampling by Gibbs or Metropolis algorithms is not possible due to the linear operator A which creates pixel interaction on a large neighborhood, whatever it models a blur [12] or a Radon transform [1]. Due to this large interaction between pixels, sampling algorithm will not converges in reasonable computing time, see [1] for details. A way to tackle this problem, in the case of a convolution operator, is to make the calculations in the Fourier domain (or cosine transform) in order to diagonalize the A operator. The problem is that F cannot be diagonalized in the Fourier domain too because of its composition with functions $\phi_{\mathbf{m}}$ which can be non necessarily quadratic. The difficulty comes from the simultaneous presence of A and F with different norms, that cannot be diagonalized in the same space. Consequently, these twooperators have to be separated in the probability energy.

2.3 Auxiliary variable

We propose to dissociate both terms with operators A and F by introducing an auxiliary variable w which acts as a hidden variable. This work is inspired from [7], where the following result is used

$$\frac{\|g-Au\|^2}{2\sigma^2} = \min_w \frac{1}{2\sigma^2\mu}\Big(\|u-w\|^2 + \langle Cw, w\rangle\Big) + \frac{1}{2\sigma^2}\Big(\|g\|^2 - 2\langle Au, g\rangle\Big), \tag{10}$$

where $C = B(\mathrm{I}-B)^{-1}$ et $B = \mu A^*A$ (μ such that $\mu\|A^*A\| < 1$).

We thus consider a new criterion $J(u,w)$ which is defined by

$$J(u,w) = \frac{1}{2\sigma^2\mu}\Big(\|u-w\|^2 + \langle Cw, w\rangle\Big) + \frac{1}{2\sigma^2}\Big(\|g\|^2 - 2\langle Au, g\rangle\Big) + \sum_{\mathbf{m}=1}^{M} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u) \tag{11}$$

In [7], it is shown that minimizing $J(u)$ w.r.t u is equivalent to minimize $J(u,w)$ w.r.t (u,w). In this work, we show that $p_\lambda(g) = \int_{u,w} p_\lambda(g,u,w)dwdu$, where the joint probability $p_\lambda(g,w,u)$ is a Gibbs distribution whose energy is $J(u,v)$ defined in (11). This property, adding the fact that the operators A and F are no longer applied on the same variables in energy (11) due to $\langle Au, g\rangle = \langle u, A^*g\rangle$, allows to derive a new Markov Chain to sample the distributions involved in the ML estimation of λ hyper-parameter.

Let us show that $p_\lambda(g) = \int_{u,w} p_\lambda(g,u,w)dwdu$. After computations, we can show that the energy in (11) can be written as:

$$J(u,w) = \frac{1}{2\sigma^2\mu}\Big((w-(\mathrm{I}+C)^{-1}u)^{\mathrm{T}}(\mathrm{I}+C)(w-(\mathrm{I}+C)^{-1}u)\Big) + \frac{1}{2\sigma^2}\Big(\|g-Au\|^2\Big) + \sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u) \tag{12}$$

as $(\mathrm{I}+C)^{-1} = \mathrm{I} - \mu A^*A$.

Form this equation, one can note that variables g and w are independant conditionally to u, i.e. $p(g,w|u) = p(g|u)p(w|u)$, and that

$$p(w|u) = \frac{1}{K_\mu}\exp\Big(-\frac{(w-(\mathrm{I}+C)^{-1}u)^{\mathrm{T}}(\mathrm{I}+C)(w-(\mathrm{I}+C)^{-1}u)}{2\sigma^2\mu}\Big) \tag{13}$$

is a Gaussian law $\mathcal{N}\big((\mathrm{I}+C)^{-1}u, \sigma^2\mu(\mathrm{I}+C)^{-1}\big)$, with $K_\mu = (2\pi\sigma^2\mu)^{k/2}(\det(\mathrm{I}+C))^{-1/2}$.

Then due to (12) and (13), we have

$$p_\lambda(g,u,w) = p_\lambda(g|u,w)p_\lambda(u,w) = p(g|u)p(w|u)p_\lambda(u). \tag{14}$$

As $p(w|u)$ is a (Gaussian) normalized distribution, we have

$$\int_w p(w|u)dw = 1, \tag{15}$$

then we deduce from (14) and (15) that

$$p_\lambda(g) = \int_{u,w} p_\lambda(g,u,w)dwdu = \int_u p(g|u)p_\lambda(u)du.$$

We now proceed similarly to the previous section, but now by inferring w.r.t u and w variables. In order to make the ML estimation of the parameters, we have to maximize the probability law $p_{\boldsymbol{\lambda}}(g)$ w.r.t $\boldsymbol{\lambda}$.

$$p_{\boldsymbol{\lambda}}(g) = \int_{u,w} p_{\boldsymbol{\lambda}}(g,u,w)dudw = \int_{u,w} p(g|u)p(w|u)p_{\boldsymbol{\lambda}}(u)dudw$$

and using (11) we have

$$\begin{aligned} p_{\boldsymbol{\lambda}}(g) =& \frac{1}{K_\sigma K_\mu Z_{\boldsymbol{\lambda}}}\int_{u,w}\exp\Big(-\frac{1}{2\sigma^2\mu}\Big(\|u-w\|^2 + \langle Cw, w\rangle\Big) - \frac{1}{2\sigma^2}\Big(\|g\|^2 - 2\langle u, A^*g\rangle\Big) \\ & - \sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)dudw \end{aligned}$$

We now compute the partial derivatives of the log-likelihood with respect to each parameter $\lambda_{\mathbf{m}}$ invoking again the Lebesgue dominated convergence theorem:

$$\frac{\partial \log p_{\boldsymbol{\lambda}}(g)}{\partial \lambda_{\mathbf{m}}} = \mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \mathrm{E}_{\sigma,\boldsymbol{\lambda},\mu}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)]. \tag{16}$$

where

$$\mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] = \frac{1}{Z_{\boldsymbol{\lambda}}}\int_u \phi_{\mathbf{m}}(F_{\mathbf{m}}u)\exp\big(-\sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\big)du \tag{17}$$

$$\mathrm{E}_{\sigma,\boldsymbol{\lambda},\mu}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] = \frac{1}{Z_{\sigma,\boldsymbol{\lambda}}(g)}\int_u \phi_{\mathbf{m}}(F_{\mathbf{m}}u)\exp\Big(\frac{-\|g-Au\|^2}{2\sigma^2}-\sum_{\mathbf{m}=1}^{M}\lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)du \tag{18}$$

$$\tag{19}$$

The partial derivatives of the log-likelihood with respect to parameter $\lambda_{\mathbf{m}}$ in (16) involves two expectations. The first expectation, defined in (17), involves the *a priori* law $p_{\boldsymbol{\lambda}}(u)$ which can be analytically computed in some particular cases. For example, when $\phi_{\mathbf{m}}(F_{\mathbf{m}}u) = \|F_{\mathbf{m}}u\|_1$, $p_{\boldsymbol{\lambda}}(u)$ reduces to a Laplace law in the subbands and thus this expectation can be computed easily and is equal to $1/\lambda_{\mathbf{m}}$. The second expectation, defined in (18) involves the *a posteriori* law $p_{\boldsymbol{\lambda}}(u,w|g)$. Contrary to the previous case in (5) where A and F were both applied on u, now only F is applied on u, and A is applied on g and w. Then it be possible to build a sampler acting on u and w alternately, each sampling can now be diagonalized in a given space (Fourier and wavelet respectively). Then the second expectation in (16) is numerically approximated using a sequence of samples $u_l, l = l_0, ..., L$ which follows the *a posteriori* probability density $p_{\boldsymbol{\lambda}}(u,w|g)$

$$\mathrm{E}_{\sigma,\boldsymbol{\lambda},\mu}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] \simeq \left(\frac{1}{L-l_0+1}\right)\sum_{l=l_0}^{L}\phi_{\mathbf{m}}\big(F_{\mathbf{m}}u_l\big) \tag{20}$$

where u_l is the lth sample generated according to the *a posteriori* probability density $p_{\boldsymbol{\lambda}}(u,w|g)$. L is the number of samples computed and l_0 defines the number of samples needed to initialize the chain.

The precise sampling algorithm as well as the entire estimation algorithm are described in the next section.

3 Algorithm

A gradient method is used in order to compute the parameters $\lambda_{\mathbf{m}}$ [11]:

$$\lambda_{\mathbf{m}}^{(n+1)} = \lambda_{\mathbf{m}}^{(n)} + \alpha_n\Big[\mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \frac{2}{L}\sum_{l=L/2+1}^{L}\phi_{\mathbf{m}}\big(F_{\mathbf{m}}{u^{(n)}}_{\sigma,\boldsymbol{\lambda}^{(n)},\mu}\big)_l\big)\Big] \tag{21}$$

where $\mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)]$ is computed explicitly as mentioned previously and $\big({u^{(n)}}_{\sigma,\boldsymbol{\lambda}^{(n)},\mu}\big)_l$ is the lth sample generated according to the *a posteriori* probability density $p_{\boldsymbol{\lambda}^{(n)}}(u,w|g)$. α_n is here the step-size of the algorithm and can vary along with the iterations (this point is discussed in the next subsection).

3.1 Global algorithm

The pseudo-code of the resulting algorithm is given in Algorithm 1.

Remark: In practice, μ is taken equal to $0.99/\|A^*A\|$. Furthermore, to initialize $\lambda_{\mathbf{m}}^{(0)}$ in the gradient algorithm 1, we perform a Wiener filtering on the degraded image g and estimate the parameters by a classical maximum likelihood approach on the resulting image.

Algorithm 1 General form of the gradient algorithm

Set $\mu < 1/\|A^*A\|$.
Set $u^{(-1)} = g$.
for every subbands $\mathbf{m}$, **do**
  Initialize $\lambda_{\mathbf{m}}^{(0)}$.
end for
for $n = 0, 1, \ldots$ **do**
  Fix an algorithm step-size α_n
  for $l = 1, \ldots, L$ **do**
    Generate w_l directly in the frequency domain according to $p(w|u^{(n-1)})$ (see Algo. 2).
    for every subbands $\mathbf{m}$ **do**
      Generate $\big(F_{\mathbf{m}}u^{(n)}{}_{\sigma,\boldsymbol{\lambda}^{(n)},\mu}\big)_l$ in the wavelet transform domain according to $p_{\boldsymbol{\lambda}}(u^{(n-1)}|w_l, g)$ (see Algo. 3)
    end for
    ($u^{(n)}$ is obtained applying an inverse wavelet transform on coefficients $F_{\mathbf{m}}u^{(n)}$)
  end for
  for every subbands $\mathbf{m}$ **do**
    $\lambda_{\mathbf{m}}^{(n+1)} = \lambda_{\mathbf{m}}^{(n)} + \alpha_n \Big[\mathrm{E}_{\boldsymbol{\lambda}}[\phi_{\mathbf{m}}(F_{\mathbf{m}}u)] - \frac{2}{L}\sum_{l=L/2+1}^{L} \phi_{\mathbf{m}}\big(F_{\mathbf{m}}u^{(n)}{}_{\sigma,\boldsymbol{\lambda}^{(n)},\mu}\big)_l\Big]$.
  end for
end for

3.2 Gibbs sampler and Metropolis-Hastings algorithm

We focus here on the second term of iteration (21). The sampling has to be done according to $p_{\boldsymbol{\lambda}}(u, w|g)$ for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(n)}$. For that, we proceed in two successive steps as described bellow:

(a) we first generate (Gibbs sampler) samples according to $p(w|u)$ given by (13) (Gaussian law). The variable w is directly expressed in the Fourier transform domain as the covariance matrix can be diagonalized easily. The Gibbs sampler is described by Algorithm 2.

Algorithm 2 Gibbs sampler: generation of w samples according to $p(w|u)$

Given u, compute the mean $(\mathrm{I} + C)^{-1}u = (\mathrm{I} - \mu A^*A)u$ and the variance $\sigma^2\mu(\mathrm{I} - \mu A^*A)$ of the Gaussian distribution.
Generate first w as $\mathcal{N}(0, 1)$ and then compute the Fourier transform of w.
Restore the mean of the distribution and then the variance.
w is obtain applying an inverse Fourier transform on the obtained coefficients.

(b) secondly, having a generation of w samples, we generate u samples according to $p_{\boldsymbol{\lambda}}(u|w, g)$ where:

$$p_{\boldsymbol{\lambda}}(u|w, g) \propto \exp\Big(-\frac{1}{2\sigma^2\mu}\|u - w\|^2 + \frac{1}{\sigma^2}\langle u, A^*g\rangle - \sum_{\mathbf{m}} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big) \tag{22}$$

In that case, we use a Metropolis-Hastings algorithm directly in the wavelet transform domain which is equivalent to write (22) as

$$p_{\boldsymbol{\lambda}}(u|w, g) \propto \exp\Big(-\frac{1}{2\sigma^2\mu}\|Fu - Fw\|^2 + \frac{1}{\sigma^2}\langle Fu, FA^*g\rangle - \sum_{\mathbf{m}} \lambda_{\mathbf{m}}\phi_{\mathbf{m}}(F_{\mathbf{m}}u)\Big)$$

The Metropolis-Hastings algorithm is described by Algorithm 3.

This operation compound of two steps is done I times and a mean over the $I/2$ last samples is taken into account.

Algorithm 3 Metropolis-Hastings algorithm: generation of u samples according to $p_{\boldsymbol{\lambda}}(u|w,g)$

1: Let $k_{\mathbf{m}}$ be the subband size.
2: Fix the proposal density to $\mathcal{N}(0,\sigma_p^2)$ and generate coefficients $F_{\mathbf{m}}c$ according to this law.
3: Generate $k_{\mathbf{m}}$ uniformly distributed random values v in $[0,1]$.
4: **if** $\frac{p_{\boldsymbol{\lambda}}(c|w,g)}{p_{\boldsymbol{\lambda}}(u|w,g)} > v$ (component-wise computation) **then**
5: $\quad u = c$ (component-wise acceptation)
6: **end if**

3.3 Choice of the step-size α_n

The simplest way to define α_n is to fix it to some small value to insure the algorithm convergence. The drawback of this approach is that the convergence speed can be highly suboptimal thus leading to a very slow convergence rate.

As already mentioned, the step-size of the gradient algorithm could be chosen adaptively at each iteration. One can think in using $\alpha_n = \dfrac{\alpha}{n+1}$ where α is some initial fixed value. However, in our simulations, this stategy was not always satisfactory, the step-size becoming too small.

This led us to consider an adaptive strategy similar to the one used in [17, p. 3835]. However, at some point, the criterion to be fulfilled in order the step-size to be updated by $\alpha_{n+1} = \dfrac{1}{\frac{1}{\alpha_n}+1}$ was difficult the reach and thus the algorithm became very time consuming, the calculation of the step-size becoming very long.

In the next section, we considered the simplest strategy, fixing a small step-size. However, it can be noticed that instead of having an iteration adaptive step-size, one could think in having a subband adaptive step-size strategy. The related developments are under investigations and appear encouraging.

4 Numerical experiments on synthetic data

In a first time, in order to test our algorithm performances, we randomly generate wavelet coefficients according to the a priori law (3) with realistic (for natural images) subband fixed parameters $\lambda_{\mathbf{m}}$.

We use Symlets [18] of length 8 over $J = 2$ resolution levels and each subband $\mathbf{m}$ is represented by the triplet (j,l,c) where j is the resolution level index, and $(l,c)_{l\in\{0,1\},c\in\{0,1\}}$ represents the low/high-pass filtered subbands (the couple $(1,1)$ thus represents the diagonal coefficients). The corresponding size of the generated image is 128×128.

We applied our method on a degraded image g where A is a Gaussian blur of standard deviation equal to 0.5 and the noise has a variance σ^2 equal to 25 (the resulting Signal to Noise Ratio (SNR) is about 10.68 dB). The gradient algorithm is launched over $L = 2000$ iterations using a fixed step-size $\alpha_n = 0.0001$, and is initialized by applying a Wiener filter on g and then estimating the parameters by a maximum likelihood approach on the Wiener filtered image. Some obtained numerical results are given in Tab. 1; more precisely, we give: the theoretical value, the value estimated by maximum likelyhood on the ground truth (complete data) and for the proposed method, we give the mean value over the 500 last values ($l_0 = 1501$) generated by the gradient algorithm.

The results given above are encouraging since the estimated values given by our approach are very close from those obtained by maximum likelihood on the ground truth. Moreover, looking at Fig.1, one can note that the proposed gradient algorithm quickly converges to the theoretical value. However, it still slightly oscillating thus justifying that the final value is obtained by computing a mean over the 500 last generated values.

Subband $\mathbf{m}$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$	
Theoretical value	0.0611	0.0547	0.0551	
Maximum likelihood on ground truth	0.0609	0.0548	0.0555	
Maximum likelihood on degraded data (proposed method)	0.0606	0.0544	0.0555	
Subband $\mathbf{m}$	$(2,0,0)$	$(2,0,1)$	$(2,1,0)$	$(2,1,1)$
Theoretical value	0.0124	0.0256	0.0247	0.0255
Maximum likelihood on ground truth	0.0123	0.0255	0.0251	0.0266
Maximum likelihood on degraded data (proposed method)	0.0123	0.0256	0.0250	0.0265

Table 1: Numerical results.

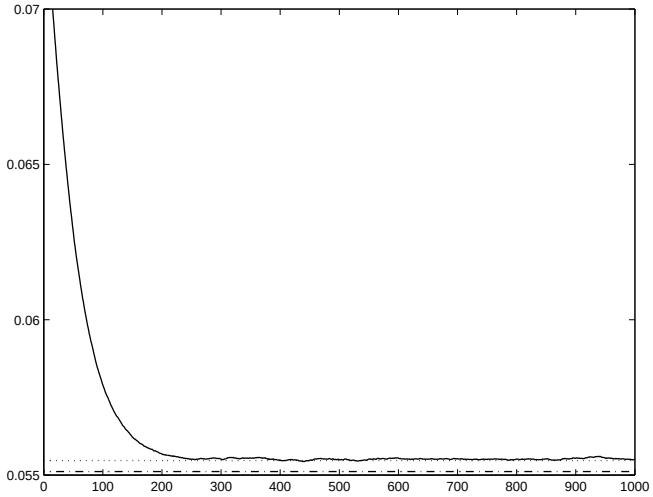

Figure 1: $\lambda_{\mathbf{m}}$ behavior (over the first 1000 iterations) associated to the subband ($\mathbf{m} = (1,1,1)$). Theoretical value (dash-dot line), maximum likelihood value (dot line) and iteratively estimated values along with the gradient algorithm iteration (continuous line).

5 Conclusion

A regularization parameter estimation approach has been proposed within a variational framework. The considered parameters to be estimated are considered as deterministic and a maximum likelihood based estimator is used. However, it could have been possible to include the samplings of $\boldsymbol{\lambda}$ and σ by associating them adapted laws and hyper hyperarameters (fully Bayesian framework). Moreover, under certain conditions, this approach can be extended to more general regularization functions.

Tests on real natural data as well as applications to data restoration are under investigation.

References

[1] D. Geman and C. Yang. Nonlinear Image recovery with Half-Quadratic Regularization. *IEEE Transactions on Image Processing*, 23(7):932–946, July 1995.

[2] G. a Demoment. Image reconstruction and restoration: overview of common estimation strucutres and problems. *IEEE Transactions on Acoustic Speech and Signal Processing*, 37(12):2024–2036, Dec. 1989.

[3] L.I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation noise removal algorithm. *Physica D*, 60:259–268, 1992. Elsevier Science Publisher.

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

[5] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574, December 2007.

[6] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. on Image Proc.*, 12(8):906–916, Aug. 2003.

[7] J. Bect, L. Blanc-Féraud, G. Aubert, and A. Chambolle. A l^1-unified variational framework for image restoration. In T. Pajdla and J. Matas, editors, *Proc. European Conference on Computer Vision (ECCV)*, volume LNCS 3024, pages 1–13, Prague, Czech Republic, May 2004. Springer.

[8] C. Vonesch, S. Ramani, and M. Unser. Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constrain. In *Proc. of the 2008 IEEE International Conf. on Image Proc. (ICIP'08)*, pages 665–668, San Diego CA, USA, October 12-15 2008.

[9] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs. A variational formulation for frame based inverse problems. *Inverse Problems*, 23:1495–1518, June 2007.

[10] S. Mallat. *A wavelet tour of signal processing*. Academic Press, San Diego, USA, 1998.

[11] L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82:625–645, 1989. Springer-Verlag.

[12] A. Jalobeanu, L. Blanc-Féraud, and J. Zerubia. Hyperparameter estimation for satellite image restoration using a MCMC maximum likelihood method. *Pattern Recognition*, 35(2):341–352, 2002.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 (series B):1–38, 1977.

[14] M. Jansen and A. Bultheel. Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise. *IEEE Trans. on Image Proc.*, 8(7):947–953, July 1999.

[15] M. M. Ichir and A. Mohammad-Djafari. Hidden markov models for wavelet-based blind source separation. *IEEE Trans. on Image Proc.*, 15(7):1887–1899, July 2006.

[16] D. Leporini and J.-C. Pesquet. Bayesian wavelet denoising: Besov priors and non-Gaussian noises. *Signal processing*, 81(1):55–67, 2001.

[17] J. Gauthier, L. Duval, and J. C. Pesquet. Optimization of synthesis oversampled complex filter banks. *IEEE Trans. on Signal Proc.*, 57(10):3827–3843, Oct. 2009.

[18] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF, SIAM Lecture Series, Philadelphia, PA, USA, 1992.