Bayesian Wavelet-Based Image Deconvolution: A GEM Algorithm Exploiting a Class of Heavy-Tailed Priors

José M. Bioucas-Dias, *Member, IEEE*

***Abstract*—Image deconvolution is formulated in the wavelet domain under the Bayesian framework. The well-known sparsity of the wavelet coefficients of real-world images is modeled by heavy-tailed priors belonging to the Gaussian scale mixture (GSM) class; i.e., priors given by a linear (finite of infinite) combination of Gaussian densities. This class includes, among others, the *generalized Gaussian*, the *Jeffreys*, and the *Gaussian mixture* priors. Necessary and sufficient conditions are stated under which the prior induced by a thresholding/shrinking denoising rule is a GSM. This result is then used to show that the *prior induced* by the "nonnegative garrote" thresholding/shrinking rule, herein termed the *garrote* prior, is a GSM. To compute the *maximum a posteriori* estimate, we propose a new *generalized expectation maximization* (GEM) algorithm, where the missing variables are the scale factors of the GSM densities. The maximization step of the underlying *expectation maximization* algorithm is replaced with a *linear stationary second-order iterative method*. The result is a GEM algorithm of $O(N \log N)$ computational complexity. In a series of benchmark tests, the proposed approach outperforms or performs similarly to state-of-the art methods, demanding comparable (in some cases, much less) computational complexity.**

***Index Terms*—Bayesian, deconvolution, expectation maximization (EM), generalized expectation maximization (GEM), Gaussian scale mixtures (GSM), heavy-tailed priors, wavelet.**

I. Introduction

IMAGE deconvolution is a longstanding linear inverse problem with applications in remote sensing, medical imaging, astronomy, seismology, and, more generally, in image restoration [3].

The challenge in many linear inverse problems is that they are ill posed, i.e., either the linear operator does not admit inverse or it is nearly singular yielding highly noise sensitive solutions. To cope with the ill-posed nature of these problems, a large number of techniques has been developed, most of them under the regularization or the Bayesian frameworks [4]–[8].

The heart of the regularization and Bayesian approaches is the *a priori* knowledge expressed by the prior/regularization term, which attaches higher scores to images or structures believed to be more likely. However, tailoring a prior for real-world images is a nontrivial and subjective matter, to which many directions have been proposed. Relevant classes of priors are, in chronological order, the Gaussian (quadratic energy) implicit in the Wiener filter [3], the *compound Gauss Markov random field* [9] (weak membrane [10] in the regularization setup), the *Markov random field* with nonquadratic potencials [11], [12], and heavy-tailed densities on the wavelet domain [13]–[24].

The weak membrane [10], in the regularization setup, and the compound Gauss Markov random field [9], in the Bayesian setup, were conceived to model piecewise-smooth images. Algorithms [10], [25]–[27], and [28] are but a few examples using piecewise-smooth priors. By signaling boundaries between smooth regions with discrete random variables, the so-called *line field*, these priors improve the modeling accuracy near the edges in comparison with the classical quadratic ones. Piecewise-smooth priors were not, however, designed to model texture, this being a major limitation, as many real-world images are partially or totally textured.

Wavelets have been increasingly used in the last years in statistical and data analysis applications [29]–[31]. Underlying this trend is the parsimonious representation provided by the wavelet transform of a large class of real-world images: elements of this class are essentially described by a few large wavelet coefficients. This fact has fostered Bayesian and regularization approaches, where the prior favors a few large wavelet coefficients and many nearly zero ones (the so-called heavy-tailed density priors) [13]–[15], [17]–[21], [23], [24]. Some examples of heavy-tailed densities adopted in image restoration are the equivalent garrote [32], [33], the generalized Gaussian [34], the Jeffreys noninformative prior [35], [36], and the Gaussian mixture (GM) [20], [37].

Wavelet descriptors are among the best in representing real-world images. However, very often, restoration criteria resulting from using wavelets are very hard to implement, at least from the computational point of view. The reasons are usually two-fold: 1) heavy-tailed priors lead frequently to huge nonconvex optimization problems and 2) in formulating linear space-invariant inverse problems in the wavelet domain, one is frequently faced with linear operations resulting from the composition of Toeplitz operators with wavelet transforms. This composed operator is not diagonal and introduces unbearable computational complexity in the wavelet-based deconvolution schemes. Recent works [24], [38], [43] have circumvented this difficulty by recognizing that each of these operations *per se* can be computed efficiently with fast algorithms.

Manuscript received June 29, 2004; revised April 11, 2005. This work was supported by the Fundação para a Ciência e Tecnologia under the projects PDCTE/CPS/49967/2003 and POSC/EEA-CPS/61271/2004. References [1] and [2] are short versions of this work. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Vicent Caselles.

The author is with the Department of Electrical and Computer Engineering, Instituto of Telecommunications, Instituto Superior Técnico, 1049-001 Lisboa, Portugal (e-mail: bioucas@lx.it.pt).

Digital Object Identifier 10.1109/TIP.2005.863972

A. Proposed Approach

We formulate image deconvolution in the wavelet domain following a Bayesian approach. The observed images are assumed to be degraded by space-invariant blur and additive Gaussian noise. The wavelet coefficients are assumed to be independent with density given by a Gaussian scale mixture (GSM) [39]–[41]. This set of densities contains many heavy-tailed priors adopted in image restoration of real-world images, namely the generalized Gaussian class [34], the Jeffreys noninformative prior [36], and the GM [20], [37]. We show that the prior induced by the *garrote* thresholding rule [32] (see also [33]) is also a GSM. Furthermore, we state necessary and sufficient conditions under which the prior induced by a thresholding/shrinking denoising rule is a GSM.

To compute the MAP estimate, we propose an expectation maximization (EM) algorithm, where the missing variables are the scale factors of the prior GMs. The maximization step of the EM algorithm includes a huge nondiagonal linear system with unbearable computational complexity. To avoid this difficulty, we approximate the linear system solution by a few iterations of a *linear stationary second-order iterative method*. The resulting scheme is a *generalized expectation maximization* (GEM) [42] algorithm, achieving convergence in a few tens of iterations. The fast Fourier transform (FFT) and the discrete wavelet transform (DWT) are the heaviest computations of each GEM step. Thus, the overall algorithm complexity is $O(N \log N)$. In a set of experiments, the proposed algorithm either equals or outperforms state-of-the-art methods [14], [19], [23], [24], [43].

This paper is organized as follows. Section II formulates the restoration problem in the wavelet domain under the Bayesian framework. Section III studies the GSM class of heavy-tailed priors, focusing on the generalized Gaussian densities, the Jeffreys noninformative prior, the GM, and the equivalent garrote prior. Necessary and sufficient conditions are stated under which the prior induced by a thresholding/shrinking denoising rule is a GSM. Section IV presents an EM algorithm to compute the maximum *a posteriori* (MAP) image estimate and a GEM version aimed at the fast computation of the MAP image estimate. Section VI addresses convergence and numerical aspects of the proposed GEM algorithm. Finally, Section VII presents a series of experimental results illustrating the effectiveness of the proposed methodology.

II. Problem Formulation

Let $\mathbf{x}$ and $\mathbf{y}$ be the vectors containing the original and the observed degraded image gray levels, respectively, arranged in column lexicographic ordering. We assume, without loss of generality, that images are square of size N (number of pixels).

The observation model herein considered is

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

where $\mathbf{H}$ is a square block-Toeplitz matrix accounting for space-invariant blur and $\mathbf{n}$ is a sample of zero-mean white Gaussian noise vector with density $p_N(\mathbf{n}) = \mathcal{N}(\mathbf{n}|\mathbf{0}, \sigma^2\mathbf{I})$ [$\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{C})$ denotes a Gaussian multivariate density of mean $\mathbf{m}$ and covariance $\mathbf{C}$ evaluated at $\mathbf{z}$, and $\mathbf{I}$ is the identity matrix.

Let a given wavelet forward and inverse transforms be represented by the $M \times N (M \geq N)$ and $N \times M$ matrices $\mathbf{W}$ and $\mathbf{P}$, respectively, and $\boldsymbol{\theta} = \mathbf{W}\mathbf{x}$ the wavelet coefficients of $\mathbf{x}$. We assume that the system has the perfect reconstruction property (i.e., $\mathbf{P}\mathbf{W} = \mathbf{I}$) and that $\mathbf{P}^T = \mathbf{W}$. Rivaz in [16] termed couples $\mathbf{W}$ and $\mathbf{P}$ exhibiting the latter property *balanced wavelet transforms*. Herein, we adopt this definition. Balanced transforms use the conjugate time-reverse of the analysis filter for the reconstruction filters. The orthogonal discrete wavelet transform (DWT) is an example of a balanced transform. The Q-shift dual tree complex wavelets (DT-CWT) [44] are nearly balanced in the sense that $\mathrm{tr}\{(\mathbf{P}^T - \mathbf{W})(\mathbf{P}^T - \mathbf{W})^T\}/N$ is negligible (see [16, Ch. 2.5]).

Introducing $\mathbf{x} = \mathbf{P}\boldsymbol{\theta}$ in (1), we have

$$\mathbf{y} = \mathbf{H}\mathbf{P}\boldsymbol{\theta} + \mathbf{n}. \tag{2}$$

The density of the observed vector $\mathbf{y}$ given $\boldsymbol{\theta}$ is then $p_N(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{H}\mathbf{P}\boldsymbol{\theta}, \sigma^2\mathbf{I})$. Given a prior $p_\Theta(\boldsymbol{\theta})$, the MAP estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{3}$$

where

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log p_N(\mathbf{y}|\boldsymbol{\theta}) + \log p_\Theta(\boldsymbol{\theta}) \\ &= \frac{-\|\mathbf{y} - \mathbf{H}\mathbf{P}\boldsymbol{\theta}\|^2}{2\sigma^2} + \log p_\Theta(\boldsymbol{\theta}). \end{aligned} \tag{4}$$

As in many recent works, we assume that the wavelet coefficients are mutually independent and identically distributed, i.e.,

$$p_\Theta(\boldsymbol{\theta}) = \prod_{i=1}^{N} p_\theta(\theta_i).$$

The independence assumption is motivated by the high degree of decorrelation exhibited by wavelet coefficients of real-world images. Although decorrelation does not imply independence, the latter has led to very good results.

If $\mathbf{H} = \mathbf{I}$, i.e., there is no blur, the image restoration at hand falls into a denoising problem (see, e.g., [32]–[34], [36], [45], and [46]). In this case, and by using the DWT, maximization (3) reduces to N decoupled coefficient-wise maximizations, which can be efficiently solved exploiting the orthogonality of $\mathbf{W}$ and using fast implementations of the DWT (see, e.g., [33], [34]).

If $\mathbf{H} \neq \mathbf{I}$, i.e., there exists blur, the maximization (3) cannot be decoupled, leading to a hard computational problem, mainly due to the $N \times M$ matrix $\mathbf{H}\mathbf{P}$. In fact, even in the cases where the prior term $\log p_\Theta(\boldsymbol{\theta})$ is of the form $\boldsymbol{\theta}^T\mathbf{D}\boldsymbol{\theta}$, i.e., quadratic on $\boldsymbol{\theta}$, the solution for the linear system $(\sigma^2\mathbf{D} + \mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P})^{-1}\boldsymbol{\theta} = \mathbf{P}^T\mathbf{H}^T\mathbf{y}$ one is led to is not an easy task, often involving iterative procedures.

To compute the MAP estimate (3), we adopt an EM approach. The first step in designing an EM algorithm is the choice of the so-called missing data [42]. In the present approach, the missing data is a set of random variables playing the role of scale coefficients in the GSM decomposition of the wavelet prior. The next section addresses aspects of this decomposition for commonly used heavy-tailed priors.

III. GSM: A Unifying Framework for Commonly Used Heavy-Tailed Priors

A random variable θ is said to be a GSM if its density can be decomposed into a linear combination (finite or infinite) of zero-mean Gaussian densities; i.e.,

$$p_\theta(\theta) = \int_0^\infty p_{\theta|z}(\theta|z)\, p_z(z)\, dz = E_z[p_{\theta|z}(\theta|z)] \tag{5}$$

where $p_{\theta|z}(\theta|z) \equiv \mathcal{N}(\theta|0, z)$. According to (5), the random variable θ is interpretable as

$$\theta = \sqrt{z}u \tag{6}$$

where u and z are random variables with densities $\mathcal{N}(u|0,1)$ and $p_z(z)$, respectively. For a given z, the term $\sqrt{z}$ plays the rule of a scale factor multiplying the Gaussian random variable u. Thus, the designation GSM.

A symmetric density p_θ satisfying the condition $p_\theta(0) < \infty$ is a GSM if and only if $p_\theta(\sqrt{\theta})$ is *completely monotone*[1] [41]. Many heavy-tailed priors used in wavelet-based image denoising/restoration admit the GSM decomposition (5). Some examples are listed in Table I. The GM is itself a GSM; the garrote density is addressed below; details about the other densities can be found in [47] (generalized Gaussian), [41] [Laplace (i.e., generalized Gaussian with $\nu = 1$) and Hardy], and [35] (Jeffreys).

A. GSM and Thresholding/Shrinking Functions

Consider the following question: in a denoising problem (i.e., $\mathbf{H} = \mathbf{I}$) using the DWT, given an antisymmetric nondecreasing thresholding/shrinking function T, is there any prior $p_\theta(\theta) \propto e^{-\phi(\theta)}$ such that the solution of the MAP estimate (3) is given by $\widehat{\theta} = T(\omega)$, where θ and ω are homonymous components of the DWTs of $\mathbf{x}$ and of $\mathbf{y}$, respectively?

To address this question, let us compute the MAP estimate (3) when $\mathbf{H} = \mathbf{I}$. Using the fact that $\mathbf{PW} = \mathbf{I}$ and denoting the ith component of $\mathbf{Wy}$ by ω_i, we may write the log-posterior $L(\theta)$ [see (4)] as

$$L(\boldsymbol{\theta}) = \frac{-\|\mathbf{y} - \mathbf{P}\boldsymbol{\theta}\|^2}{2\sigma^2} - \sum_{i=1}^N \phi(\theta_i) = \sum_{i=1}^N \left(\frac{-(\omega_i - \theta_i)^2}{2\sigma^2} - \phi(\theta_i) \right). \tag{7}$$

Therefore, the MAP estimate is obtained by maximizing $-(\omega_i - \theta_i)^2 - 2\sigma^2\phi(\theta_i)$, for $i = 1, \ldots, N$. Assume that $\phi(\theta)$ is differentiable in $\mathbb{R}$ except, perhaps, at $\theta = 0$. Then, the maxima of the log-posterior $L(\theta)$ is either $\widehat{\theta} = 0$ or a solution of

$$\omega = \widehat{\theta} + \sigma^2\phi'(\widehat{\theta}). \tag{8}$$

If T^{-1} exists for $|\theta| > 0$, then $\phi(\theta)$ is symmetric and given by

$$\phi(\theta) = c^{te} - \frac{\theta^2}{2\sigma^2} + \frac{1}{\sigma^2}\int_a^\theta T^{-1}(u)\, du, \quad \theta > 0 \tag{9}$$

[1]A $C^\infty(0,\infty)$ function f is said to be completely monotonic on $(0,\infty)$ provided that $(-1)^l f^{(l)}(x) \geq 0$, for all $x \in (0,\infty)$ and for all natural number l.

TABLE I
Gaussian Scaled Mixture (GSM) Densities

Generalized Gaussian	$\propto e^{-\lambda\|\theta\|^\nu}$, $0 < \nu \leq 2$, $\lambda > 0$
Laplace	$\propto e^{-\lambda\|\theta\|}$, $\lambda > 0$
Hardy	$\propto e^{-\lambda\sqrt{\alpha^2+\theta^2}}$, $\lambda > 0$, $\alpha \geq 0$
Jeffreys	$\propto 1/\|\theta\|$
Gaussian Mixture	$\propto \sum_{i=1}^n \omega_i \mathcal{N}[\theta\|0, \sigma_i^2]$, $\omega_i \geq 0$, $\sum_{i=1}^n \omega_i = 1$
Garrote	$\propto \exp\left\{ \frac{\theta^2}{4\sigma^2} - \frac{\|\theta\|\sqrt{\theta^2+4\lambda^2}}{4\sigma^2} - \frac{\lambda^2}{\sigma^2}\operatorname{arcsinh}\left(\frac{\|\theta\|}{2\lambda}\right)\right\}$, $\lambda > 0$

where $a > 0$ is any positive constant. We term $p_\theta(\theta) \propto e^{-\phi(\theta)}$ the prior *induced* by T.

As an application example of expression (9), consider the so-called "nonnegative garrote" [32] thresholding function given by

$$T(\omega) = \frac{(\omega^2 - \lambda^2)_+}{\omega} \tag{10}$$

where $(x)_+$ stands for "the positive part of", i.e., $(x)_+ = x$, if $x > 0$, and $(x)_+ = 0$, if $x \leq 0$. The inverse of T is

$$T^{-1}(\theta) = \frac{\theta + \sqrt{\theta^2 + 4\lambda^2}}{2}, \quad \theta > 0. \tag{11}$$

Introducing (11) into (9), we obtain the prior induced by the nonnegative garrote thresholding function, herein termed garrote prior. Its expression is shown in the last line of Table I. It should be noted that this density is an empirical Bayesian prior, in the sense that it depends on the noise variance σ^2 and, therefore, on the observed data. We have considered this pior since, with $\lambda = \sqrt{3}\sigma$, it leads to very good results in denoising applications, as shown in [33] following an empirical Bayes approach.

A straightforward way to confirm whether a given density $p_\theta(\sqrt{\theta})$ is a GSM, is to check if $(-1)^l p_\theta^{(l)}(\sqrt{\theta}) \geq 0$, for all $\theta \in (0,\infty)$[2] and for all natural l. However, if the prior is induced by a thresholding/shrinking function, the following lemma provides an easier way to test the GSM nature of the prior.

Lemma 1: Let $T : \mathbb{R} \to \mathbb{R}$ be an antisymmetric nondecreasing thresholding/shrinking function such that T^{-1} exists and is C^∞ for $|\theta| > 0$. Define $f_T(\theta) \equiv T^{-1}(\sqrt{\theta}) - \sqrt{\theta}$ for $\theta > 0$. Then the induced density $p_\theta(\theta) \propto e^{-\phi(\theta)}$, with $\phi(\theta)$ given by (9), is a GSM iff $(-1)^l f_T^{(l)}(\theta) \geq 0$ for $\theta > 0$ and for $l = 0, 1, \ldots$

Proof: See Appendix. ■

We now use Lemma 1 to show that the induced garrote prior is a GSM.

Proposition 1: The garrote prior induced by the thresholding function (10) is a GSM.

Proof: The function $f_T(\theta)$ introduced in Lemma 1 is for the garrote thresholding function [see (10) and (11)] given by $f_T(\theta) = (1/2)(\sqrt{\theta + 4\lambda^2} - \sqrt{\theta})$ for $\theta > 0$. A straightforward calculus based on the derivatives of $\sqrt{\theta}$ leads to the conclusion that $(-1)^l f_T^{(l)}(\theta) \geq 0$ for $\theta > 0$ and for $l = 0, 1, \ldots$ ■

[2]The notation $f^{(l)}(\sqrt{\theta})$ is to be understood as $(d^l q(\theta))/(d\theta^l)$, with $q(\theta) \equiv f(\sqrt{\theta})$.

Many heavy-tailed priors satisfy $\phi'(0+) \neq 0$, implying the existence of a threshold on ω below which $\widehat{\theta} = 0$. Therefore, this type of nonsmooth prior leads to sparse representations of the estimated images. Sparseness is a desired property in many applications such as sparse regression, variable selection, and feature selection. This is not the case in image deconvolution as we will see in Section VII.

The Hardy density is obtained from the Laplace density by replacing $|\theta|$ with $\sqrt{\theta^2+\alpha^2}$ [41]. Both are heavy-tailed; however, the latter is sparse, implying the existence of a threshold, whereas the former does not, no matter how small is $\alpha \in (0,\infty)$. With respect to optimization, the Hardy prior is preferable, as it is C^∞, widening the range of applicable optimization algorithms.

In the context of GSM densities, the considerations made in the paragraph above about Hardy and Laplace priors bring to mind the following question: Given a sparse GSM prior $p_\theta(\theta)$, is $q_\theta(\theta) = a(\alpha)p_\theta(\sqrt{\theta^2+\alpha^2})$ also a GSM? The answer is yes, since, for $\theta > 0$, the derivatives of $(-1)^l q_\theta^{(l)}(\sqrt{\theta}) = (-1)^{(l)}a(\alpha)p_\theta^{(l)}(\sqrt{\theta+\alpha^2}) \geq 0$, with $a > 0$. This fact can be exploited, for example, to eliminate difficulties in dealing with the Jeffreys prior at the origin.

IV. EM Algorithm

In this section, we develop an EM algorithm that converts the maximization (3) into a sequence of quadratic problems, each one solved iteratively and efficiently by using fast algorithms to compute $\boldsymbol{\theta} = \mathbf{W}\mathbf{x}$ (forward wavelet transform), $\mathbf{x} = \mathbf{P}\boldsymbol{\theta}$, (inverse wavelet transform), and $\mathbf{y} = \mathbf{H}\mathbf{x}$ (image convolution), thus avoiding the direct manipulation of matrix $\mathbf{HP}$.

Let $\mathbf{z} \equiv (z_1,\ldots,z_N)$ be a random vector of independent components, where $\sqrt{z_i}$ plays the role of scale in the Gaussian decomposition of $p_\theta(\theta_i)$, as referred to in (6). Let random vectors $\mathbf{z}$ and $(\mathbf{y},\mathbf{z})$ play the role of *missing data* and *complete data*, respectively, in our EM formulation. The density of the complete data, denoted by $p_{YZ\Theta}(\mathbf{y},\mathbf{z},\boldsymbol{\theta})$, is then given by

$$\begin{aligned} p_{YZ\Theta}(\mathbf{y},\mathbf{z},\boldsymbol{\theta}) &= p_{Y|Z\Theta}(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})p_{Z\Theta}(\mathbf{z},\boldsymbol{\theta}) \\ &= p_N(\mathbf{y}|\boldsymbol{\theta})p_{\Theta|Z}(\boldsymbol{\theta}|\mathbf{z})p_Z(\mathbf{z}) \end{aligned} \tag{12}$$

where the independence of $\mathbf{y}$ on $\mathbf{z}$ was used in the second line of (12), $p_{\Theta|Z}(\boldsymbol{\theta}|\mathbf{z}) \equiv \prod_{i=1}^N p_{\theta|z}(\theta_i|z_i)$, and $p_Z(\mathbf{z}) = \prod_{i=1}^N p_z(z_i)$.

The EM algorithm yields a nondecreasing log-posterior sequence [42] $\{L(\widehat{\boldsymbol{\theta}}_t),\, t=0,1,\ldots\}$, where $\{\widehat{\boldsymbol{\theta}}_t,\, t=0,1,\ldots\}$ is generated by the two-step iteration presented in Algorithm 1.

Algorithm 1 Wavelet GSM-Based EM Algorithm for Image Deconvolution.

Initialization: $\widehat{\boldsymbol{\theta}}_0$

for $t := 0$ to $StopRule$ **do**

{Expectation Step (E-Step)}

$Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_t) = E[\log[p_{YZ\Theta}(\mathbf{y},\mathbf{z},\boldsymbol{\theta})|\mathbf{y},\widehat{\boldsymbol{\theta}}_t]$

{Maximization Step (M-step)}

$\widehat{\boldsymbol{\theta}}_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_t)$

end for

The mean value in the E-step is computed with respect to $p_{Z|Y\Theta}(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$. From (12), we conclude that

$$\begin{aligned} p_{Z|Y\Theta}(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) &= \frac{p_{YZ\Theta}(\mathbf{y},\mathbf{z},\boldsymbol{\theta})}{p_{Y\Theta}(\mathbf{y},\boldsymbol{\theta})} \\ &= \frac{p_N(\mathbf{y}|\boldsymbol{\theta})p_{\Theta|Z}(\boldsymbol{\theta}|\mathbf{z})p_Z(\mathbf{z})}{p_N(\mathbf{y}|\boldsymbol{\theta})p_\Theta(\boldsymbol{\theta})} \\ &= \frac{p_{\Theta|Z}(\boldsymbol{\theta}|\mathbf{z})p_Z(\mathbf{z})}{p_\Theta(\boldsymbol{\theta})} \\ &= p_{Z|\Theta}(\mathbf{z}|\boldsymbol{\theta}). \end{aligned}$$

Since $\log p_N(\mathbf{y}|\boldsymbol{\theta}) = -\|\mathbf{y}-\mathbf{HP}\boldsymbol{\theta}\|^2/(2\sigma^2) + c_1$ does not depend on the missing data vector $\mathbf{z}$, and that $p_\theta(\theta_i) = \mathcal{N}(\theta_i|0,z_i)$, and, thus, $\log p_\theta(\theta_i) = -\theta_i^2/(2z_i) + c_2$, ($c_1$ and c_2 do not depend on $\boldsymbol{\theta}$), then

$$\begin{aligned} &Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_t) \\ &\quad = E[\log p_N(\mathbf{y}|\boldsymbol{\theta}) + \log p_{\Theta|Z}(\boldsymbol{\theta}|\mathbf{z}) + \log p_Z(\mathbf{z})|\widehat{\boldsymbol{\theta}}_t] \\ &\quad = -\frac{\|\mathbf{y}-\mathbf{HP}\boldsymbol{\theta}\|^2}{2\sigma^2} - \frac{1}{2}\boldsymbol{\theta}^T E\left[\operatorname{diag}\left\{z_1^{-1},\ldots,z_N^{-1}\right\}|\widehat{\boldsymbol{\theta}}_t\right]\boldsymbol{\theta} \\ &\qquad + E[\log p_Z(\mathbf{z})|\widehat{\boldsymbol{\theta}}_t] \\ &\quad = -\frac{\|\mathbf{y}-\mathbf{HP}\boldsymbol{\theta}\|^2}{2\sigma^2} - \frac{1}{2}\boldsymbol{\theta}^T\mathbf{D}_t\boldsymbol{\theta} + c \end{aligned} \tag{13}$$

where c stands for terms not depending on $\boldsymbol{\theta}$ and $\mathbf{D}_t \equiv E[\operatorname{diag}\{z_1^{-1},\ldots,z_N^{-1}\}|,\widehat{\boldsymbol{\theta}}_t]$ ($\operatorname{diag}(\cdot)$ stands for diagonal matrix). Given that $p_{Z|\Theta}(\mathbf{z}|\widehat{\boldsymbol{\theta}}_t)$ is conditionally independent, then it follows that $\mathbf{D}_t = \operatorname{diag}\{d(\widehat{\theta}_{it})\}$, where $d(\theta) \equiv E[z^{-1}|\theta]$ and $\widehat{\theta}_{it}$ is the ith component of $\widehat{\boldsymbol{\theta}}_t$.

The M-step consists in maximizing (13) with respect to $\boldsymbol{\theta}$, i.e.,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{t+1} &= \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta},\widehat{\boldsymbol{\theta}}_t) \\ &= \arg\min_{\boldsymbol{\theta}} \left\{\boldsymbol{\theta}^T\left(\mathbf{D}_t + \mathbf{P}^T\mathbf{H}^T\mathbf{HP}\right)\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{P}^T\mathbf{H}^T\mathbf{y}\right\} \\ &= (\sigma^2\mathbf{D}_t + \mathbf{P}^T\mathbf{H}^T\mathbf{HP})^{-1}\mathbf{P}^T\mathbf{H}^T\mathbf{y}. \end{aligned} \tag{14}$$

In (14), it is assumed that the matrix $\mathbf{A}_t \equiv (\sigma^2\mathbf{D}_t + \mathbf{P}^T\mathbf{H}^T\mathbf{HP})$ is positive definite, i.e., $\mathbf{A}_t \succ 0$. Since $\mathbf{P}^T\mathbf{H}^T\mathbf{HP} \succeq 0$, i.e., it is nonnegative definite, then a sufficient condition for $\mathbf{A}_t \succ 0$ is that $\mathbf{D}_t \succ 0$. This inequality is always satisfied, since the diagonal elements of $\mathbf{D}_t$ are mean values of positive quantities.

Recalling that $p_{\theta|z}(\theta|z) = \mathcal{N}(\theta|0,z)$, the mean value $d(\theta) = E[z^{-1}|\theta]$ can be expanded as

$$d(\theta) = \int_0^\infty z^{-1}p_{z|\theta}(z|\theta)\,dz \tag{15}$$

$$= \frac{1}{p_\theta(\theta)}\int_0^\infty z^{-1}p_{\theta|z}(\theta|z)p_z(z)\,dz \tag{16}$$

$$= \frac{-1}{\theta p_\theta(\theta)}\int_0^\infty \frac{dp_{\theta|z}}{d\theta}p_z(z)\,dz \tag{17}$$

$$= -\frac{1}{\theta p_\theta(\theta)}\frac{dp_\theta}{d\theta} \tag{18}$$

where we have used $(dp_{\theta|z})/(d\theta) = -(\theta/z)p_{\theta|z}$ in (17) and exchanged the integral with the derivative in (18). Equation (18) is very useful, since it allows to compute $d(\theta)$ directly from p_θ without the explicit knowledge of the density p_z. Moreover, if

the prior is induced by a thresholding/shrinking function such that T^{-1} exists for $|\theta| > 0$, then $p_\theta(\theta) \propto e^{-\phi(\theta)}$ and

$$d(\theta) = -\frac{1}{\theta p_\theta(\theta)} \frac{dp_\theta}{d\theta} \tag{19}$$

$$= \frac{\phi'(\theta)}{\theta} \tag{20}$$

$$= \frac{T^{-1}(\theta) - \theta}{\theta \sigma^2} \tag{21}$$

where $\phi'(\theta) = (T^{-1}(\theta) - \theta)/\sigma^2$ [see (8)] was used in (21). This expression allows to compute $d(\theta)$ directly from a given thresholding/shrinking function without the knowledge of the induced density p_θ or the density p_z.

Table II presents $d(\theta)$ for the listed priors. The expression correspondent to the induced garrote prior was obtained using the formula (21); the remaining expressions were obtained using the formula (18). The last line of Table II means that the mean value $E[z^{-1}|\theta]$ associated to the prior $q(\theta) \propto p(\sqrt{\theta^2 + \alpha^2})$ is simply the mean value $E[z^{-1}|\theta]$ associated to $p(\theta)$ computed at $\sqrt{\theta^2 + \alpha^2}$. This is a simple and useful result, providing a tool to build nonsparse priors from sparse ones.

Expressions (18) and (21) allow the direct specification of the mean value d from a given thresholding/shrinking law T or from a given density p_θ, respectively, without computing the underlying GSM. We can push further this link and apply the iteration (14), even if the underlying prior is not a GSM. Of course, in this case, it is not guaranteed that we obtain a nondecreasing log-posterior sequence.

V. Generalized Expectation Maximization (GEM) Algorithm

M-step (14) is impracticable from the computational point of view, as it amounts to solving the linear system $\mathbf{A}_t\boldsymbol{\theta} = \mathbf{y}'$ of size N^2, where $\mathbf{A}_t = \sigma^2\mathbf{D}_t + \mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}$. We tackle this difficulty by replacing the maximization of (14) with a few steps of an iterative procedure that increments $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_t)$, with respect to $\boldsymbol{\theta}$. The resulting scheme is, thus, a GEM algorithm.

Let $\mathbf{A}_t = \mathbf{C}_t - \mathbf{R}$ be a *splitting* [48] of $\mathbf{A}_t$, where $\mathbf{C}_t \equiv (\sigma^2\mathbf{D}_t + \mathbf{I})$ and $\mathbf{R} \equiv (\mathbf{I} - \mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P})$. Given that $\mathbf{A}_t \succ 0$, then the *second-order stationary iterative method* consisting of the following equations (see, e.g., [48]):

$$\begin{aligned} \mathbf{r}_i &= \mathbf{A}_t\boldsymbol{\xi}_i - \mathbf{y}' \quad i = 0, 1, \dots \\ \boldsymbol{\xi}_1 &= \boldsymbol{\xi}_0 - \beta_0\mathbf{C}_t^{-1}\mathbf{r}_0 \\ \boldsymbol{\xi}_{i+1} &= \alpha\boldsymbol{\xi}_i + (1-\alpha)\boldsymbol{\xi}_{i-1} - \beta\mathbf{C}_t^{-1}\mathbf{r}_i \quad i = 1, 2, \dots \end{aligned} \tag{22}$$

converges to the solution of $\mathbf{A}\boldsymbol{\theta} = \mathbf{y}'$, if and only if

$$\begin{cases} 0 < \alpha < 2 \\ 0 < \beta < 2\alpha/\lambda_N \end{cases} \tag{23}$$

where $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ are the eigenvalues of $\mathbf{C}_t^{-1}\mathbf{A}_t$ (see Theorem 5.9 of [48, Ch. 5]). The optimal convergence factor is

$$\rho_{\text{opt}} = [1 - \sqrt{\lambda_1/\lambda_N}]/[1 + \sqrt{\lambda_1/\lambda_N}]$$

TABLE II
Mean Values $d(\theta) = E[z^{-1} \mid \theta]$

Generalized Gaussian	$\lambda\nu/\lvert\theta\rvert^{2-\nu}$ $0 < \nu \leq 2$, $\lambda > 0$
Laplace	$\lambda/\lvert\theta\rvert$, $\lambda > 0$
Hardy	$\lambda/\sqrt{\theta^2 + \alpha^2}$, $\lambda > 0$, $\alpha \geq 0$
Jeffreys	$1/\lvert\theta\rvert^2$
Gaussian Mixture	$\frac{1}{p_\theta(\theta)} \sum_{i=1}^n \omega_i \sigma_i^{-2} \mathcal{N}(\theta \mid 0, \sigma_i^2)$
Garrote	$(2\lambda^2/\sigma^2)/(\theta^2 + \lvert\theta\rvert\sqrt{\theta^2 + 4\lambda^2})$, $\lambda > 0$
$q(\theta) \propto p(\sqrt{\theta^2 + \alpha^2})$	$d_p(\sqrt{\theta^2 + \alpha^2})$

and is achieved for

$$\begin{cases} \alpha = \rho_{\text{opt}}^2 + 1 \\ \beta = 2\alpha/(\lambda_1 + \lambda_N) \\ \beta_0 = \beta/\alpha. \end{cases} \tag{24}$$

Some algebra applied to the third line of (22) leads to

$$\begin{aligned} \boldsymbol{\xi}_{i+1} = (\alpha - \beta)\boldsymbol{\xi}_i + (1-\alpha)\boldsymbol{\xi}_{i-1} \\ + \beta\mathbf{C}_t^{-1}\left\{\boldsymbol{\xi}_i + \mathbf{P}^T\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{P}\boldsymbol{\xi}_i)\right\} \quad i = 1, 2, \dots \end{aligned} \tag{25}$$

The expression for $\boldsymbol{\xi}_1$ shown in the second line of (22) is an instance of (25) with $\alpha = 1$ and $\beta = \beta_0$. Note that if $\alpha = 1$ in (22), we obtain a *first-order stationary iterative method* converging for $0 < \beta < 2/\lambda_N$, with an optimal convergence factor of $\rho_{\text{opt}} = [1 - \lambda_1/\lambda_N]/[1 + \lambda_1/\lambda_N]$ achieved, as in the second-order case, for $\beta = 2/(\lambda_1 + \lambda_N)$ [48, Ch. 5]. For ill-conditioned systems (i.e., $\lambda_1 \ll \lambda_N$), the asymptotic rate of convergence $r \equiv -\log(\rho_{\text{opt}})$ of the first-order method is, approximately, $r = 2\lambda_1/\lambda_N$, whereas that of the second-order method is, approximately, $r = 2\sqrt{\lambda_1/\lambda_N}$. This makes a huge diference in the number of iterations necessary to reduce the error by an order of magnitude. For example, for $\lambda_1 = 0.01$ and $\lambda_N = 1$, usual values in deconvolution problems, the first-order method takes 50 iterations to reduce the error by an order of magnitude, whereas the the second-order method takes only five iterations!

Given that $\mathbf{C}_t = (\sigma^2\mathbf{D}_t + \mathbf{I})$ is diagonal, the product $\mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}\boldsymbol{\xi}_i$, necessary to determine the residual $\mathbf{r}_i$, is the heaviest computation in each iteration (25). We note, however, that $\mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}\boldsymbol{\xi}_i$ can be computed efficiently, since there exists fast implementations $[O(N)]$ for computing the the inverse wavelet transforms [30], and the product of a Toeplitz matrix by a vector can also be computed efficiently, by embedding $\mathbf{H}$ into a larger block-circulant matrix. Block-circulant matrices are diagonalized by the two-dimensional (2-D) discrete Fourier transform. Therefore, by using the 2-D fast Fourier transform, the complexity of the product of a Toeplitz matrix by a vector is $[O(N \log N)]$ [3].

A pertinent question is the choice of the number of iterations, say p. Whatever $p > 0$ might be, if parameters α and β satisfy (23), then $\|\boldsymbol{\xi}_p - \widehat{\boldsymbol{\theta}}_{t+1}\| \to 0$ as $p \to \infty$ [$\widehat{\boldsymbol{\theta}}_{t+1}$ is given by the (14)], i.e., the norm of the error tends to zero. This does not imply, however, that $Q(\boldsymbol{\xi}_p, \widehat{\boldsymbol{\theta}}_t) > Q(\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{\theta}}_t)$ as required

in the optimization step[3] (O-step). A very simple solution to this problem consists in checking if $Q(\boldsymbol{\xi}_p, \widehat{\boldsymbol{\theta}}_t) > Q(\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{\theta}}_t)$. If this inequality is not satisfied, iterate (25) until $Q(\boldsymbol{\xi}_{i+1}, \widehat{\boldsymbol{\theta}}_t) > Q(\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{\theta}}_t)$. Note that this procedure adds only a small computational complexity to the GEM algorithm, since the heaviest step in determining the quadratic function $Q(\boldsymbol{\xi}_i, \widehat{\boldsymbol{\theta}}_t)$ given by (13) is the computation of $\mathbf{HP}\boldsymbol{\theta}$, also needed to compute $\boldsymbol{\xi}_{i+1}$ in (25).

Algorithm 2 (WaveGSM) Wavelet GSM-Based GEM Algorithm for Image Deconvolution.

Initialization:

$\widehat{\boldsymbol{\theta}}_0 = \mathbf{P}^T \widehat{\mathbf{x}}_0$ {$\widehat{\mathbf{x}}_0$ is the Wiener estimate}

for {$t := 0$ to *StopRule*}

  {E-Step}

  $\mathbf{D}_t := \text{diag}\,\{E[(z_1^{-1}, \ldots, z_N^{-1})|\widehat{\boldsymbol{\theta}}_t]\}$

  {O-step (Increases $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_t)$)}

  $\boldsymbol{\xi}_0 := \widehat{\boldsymbol{\theta}}_t, n := 0$

  repeat

    for $i := np$ to $(n+1)p - 1$ **do**

      $\mathbf{C}_t^{-1}\mathbf{r}_i = \boldsymbol{\xi}_i + \mathbf{C}_t^{-1}(\mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}\boldsymbol{\xi}_i - \boldsymbol{\xi}_i - \mathbf{y}')$

      $\boldsymbol{\xi}_{i+1} = \alpha_i\boldsymbol{\xi}_i + (1-\alpha_i)\boldsymbol{\xi}_{i-1} - \beta_i\mathbf{C}_t^{-1}\mathbf{r}_i$

      {$\alpha_0 = 1$ and $\alpha_i = \alpha$, $\beta_i = \beta$, for $i \neq 0$}

    end for

    $n := n + 1$

  until $Q(\boldsymbol{\xi}_{np}, \widehat{\boldsymbol{\theta}}_t) - Q(\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{\theta}}_t) > 0$

  $\widehat{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\xi}_{np}$

end for

Algorithm 2, named WaveGSM, shows the pseudo-code for the proposed GEM iterative procedure. It is initialized with $\widehat{\boldsymbol{\theta}}_0 = \mathbf{P}^T\widehat{\mathbf{x}}_0$, where $\widehat{\mathbf{x}}_0$ is the Wiener estimate assuming that $\mathbf{x}$ is zero-mean Gaussian with covariance matrix $\sigma_x^2\mathbf{I}$, i.e.,

$$\widehat{\mathbf{x}}_0 = \left(\frac{\sigma^2}{\sigma_x^2}\mathbf{I} + \mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}.$$

Variance σ_x^2 is estimated based on the sample variance of $\mathbf{y}$ divided by the energy of the blur filter. If the condition $Q(\boldsymbol{\xi}_{np}, \widehat{\boldsymbol{\theta}}_t) - Q(\widehat{\boldsymbol{\theta}}_t, \widehat{\boldsymbol{\theta}}_t) \geq 0$ is not satisfied, another set of p iterations of expression (25) is computed. In practice, as seen in next section, a few iterations are enough to increase Q. This and other aspects concerning convergence and parametrization of the proposed algorithm are addressed in the next section.

A. Unknown Noise Variance and Prior Parameters

Until now, we have assumed that the noise variance σ^2 is known. If this is not the case, σ^2 is estimated by computing

$$\widehat{\sigma}_t^2 = \|\mathbf{y} - \mathbf{HP}\widehat{\boldsymbol{\theta}}_{t-1}\|^2/N^2 \tag{26}$$

after step 14 of Algorithm 2. If the prior does not depend on σ, this step does not modify the nondecreasing nature of $L(\widehat{\boldsymbol{\theta}}_t)$, since it is applied before the GEM steps and only to the log-likelihood term.

An alternative to (26) is computing σ beforehand and keep it constant. In the next section, we use the MAD (i.e., median absolute deviation of the finest level wavelet coefficients divided by 0.6745) [49] estimate, for this purpose.

If the prior density has unknown parameters, the EM setup supplies a tool to infer these parameters. The idea is to maximize Q not only with respect to $\boldsymbol{\theta}$, but also with respect to the unknown parameters. Of course, if the term $E[\log \mathbf{z}|\widehat{\boldsymbol{\theta}}_t]$ discarded in the definition of Q [see expression (13)] depends on any unknown prior parameter, it should be included in Q. Since the joint maximization of Q with respect to $\boldsymbol{\theta}$ and to the prior parameters may be a hard problem, we can maximize Q first with respect to $\boldsymbol{\theta}$ and then with respect to the prior parameters. This is a cyclic maximizer with just one iteration that preserves the nondecreasing nature of $L(\widehat{\boldsymbol{\theta}}_t)$. In Section VII, we use this technique to infer parameters of a GM prior.

B. Translation-Invariant Restoration

Translation invariant (TI) wavelet-based methods outperform orthogonal DWT based ones, as the former significantly reduce the *blocky* artifacts associated to the dyadic shifts inherent to the orthogonal DWT basis functions [45]. Herein, we compare two ways of implementing TI, both based on WaveGSM algorithm

1) **WaveGSM-TI** Run WaveGSM a given number of times N_s, each time t applying $\mathbf{W}_t$, a shifted version of the DWT. The final estimate is obtained by averaging all estimates. Algorithm 3 shows the pseudo-code for WaveGSM-TI. The notation $\widehat{\boldsymbol{\theta}}_t = \text{WaveGSM}(\widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{W}_t)$ stands for the output of WaveGSM initialized with $\widehat{\boldsymbol{\theta}}_{t-1}$ and applying the DWT $\mathbf{W}_t$.
2) **WaveGSM-TIR** Run exactly WaveGSM with the TI-DWT. In the present setup, replacing the orthogonal DWT with the TI-DWT does not alter the GEM nature of the developed algorithm, as the optimization step still increments the objective function $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_t)$.

Algorithm 3 (WaveGSM-TI) Translation invariant version of WaveGSM by averaging N_s estimates.

Initialization: $\widehat{\boldsymbol{\theta}}_0 = \mathbf{P}^T\widehat{\mathbf{x}}_0$ {$\widehat{\mathbf{x}}_0$ is the Wiener estimate}

for $t := 1$ to N_s **do**

  $\widehat{\boldsymbol{\theta}}_t := \text{WaveGSM}(\widehat{\boldsymbol{\theta}}_{t-1}, \mathbf{W}_t)$

  $\widehat{\boldsymbol{\theta}}_t := ([\widehat{\boldsymbol{\theta}}_t + (t-1)\widehat{\boldsymbol{\theta}}_{t-1}]/t$

end for

VI. Convergence and Numerical Analysis of the GEM Algorithm

Assuming that $Q(\boldsymbol{\theta}, \boldsymbol{\theta}') = E[\log[p_{YZ\Theta}(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta})|\boldsymbol{\theta}']$ exists for any couple $(\boldsymbol{\theta}, \boldsymbol{\theta}')$, the proposed GEM algorithm generates a nondecreasing sequence of the log-posterior $L(\boldsymbol{\theta})$. In this section, we address the following convergence aspects: 1) Does the sequence $\{L(\widehat{\boldsymbol{\theta}}_t),\ t = 0, 1, \ldots\}$ converge to stationary points of L? If so, what type are they (saddle points, local maxima, global maxima); 2) Does the sequence $\{\widehat{\boldsymbol{\theta}}_t,\ t = 0, 1, \ldots\}$ converge? In addressing these questions, we follow closely [50].

A. Convergency of $L(\widehat{\boldsymbol{\theta}}_t)$ to Stationary Points

Let us assume that the GSM density $p_\theta(\theta) \propto e^{-\phi(\theta)}$ is bounded above and that ϕ is C^∞. Then, $d(\theta) = \phi'(\theta)/\theta$ is continuous in $\mathbb{R}$ since $\lim_{\theta\to 0} d(\theta) = \lim_{\theta\to 0} \phi'(\theta)/\theta =$

[3]From now on, we refer to O-step instead of M-step, because $Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_t)$ is not maximized with respect to $\boldsymbol{\theta}$, but only increased.

$\phi''(0) < \infty$. Herein, we term densities p_θ with these properties *nonsparse priors*. Hardy prior, GM prior, and any prior $q(\theta) \propto p(\sqrt{\theta^2+\alpha^2})$, with $p(\theta)$ being a GSM and $\alpha \neq 0$, are nonsparse priors.

If a prior is nonsparse, then $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is continuous in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Then, all limit points of any GEM sequence $\{\widehat{\boldsymbol{\theta}}_t,\ t = 0,1,\ldots\}$ generated by (25) are stationary points of L and $\{L(\widehat{\boldsymbol{\theta}}_t)\}$ converges monotonically to $L^* = L(\boldsymbol{\theta}^*)$ for some stationary point $\boldsymbol{\theta}^*$. This result is a minor modification of Theorem 2 of [50], where the condition $L(\widehat{\boldsymbol{\theta}}_{t+1}) > L(\widehat{\boldsymbol{\theta}}_t)$ for any nonstationary point $\widehat{\boldsymbol{\theta}}_t$ is assured by step 13 of the WaveGSM Algorithm.

GMs and Hardy priors are nonsparse, leading to continuous $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Therefore, the respective log-posterior sequences converge to stationary points L^*. Generalized Gaussian, Laplace, Jeffreys, and Garrote priors are sparse and then $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is not defined for $\boldsymbol{\theta}' = 0$. For these priors, it is not possible, therefore, to assure convergence of L.

Although convergence of L to stationary points can not be assured in the case of sparse priors, this can be reverted by introducing a small modification in the log-posterior, consisting in replacing the prior $p_\theta(\theta)$ with $p_\theta(\sqrt{\theta^2+\alpha^2})$, where α is a nonzero arbitrary small number. As seen before, this assures continuity of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ at $\boldsymbol{\theta}' = 0$ and, thus, convergence of L to stationary points. Furthermore, the heavy-tailed nature of the prior is kept, giving credit to the modification. Of course, the sparseness is lost in the sense that many estimated coefficients that were exactly zero are now very small, but not zero. In terms of deconvolution, usually this is not a problem.

B. Type of Stationary Points

Provided that the sequence $\{\widehat{\boldsymbol{\theta}}_t\}$ converges to a stationary point, the next natural question is what type of stationary point is $\{\widehat{\boldsymbol{\theta}}_t\}$ converging to? If all stationary points are local (global) maxima, then $\{L(\widehat{\boldsymbol{\theta}}_t)\}$ converges to a local (global) maxima. The problem is that, very often, $L(\boldsymbol{\theta})$ has saddle points that may trap the GEM sequence.

In the case of nonsparse priors and strictly convex log-posterior L, there is a unique maximum L^* and $\{L(\widehat{\boldsymbol{\theta}}_t)\}$ converges to it. Moreover, the sequence $\{\widehat{\boldsymbol{\theta}}_t\}$ converges to $\boldsymbol{\theta}^*$, such that $L^* = L(\boldsymbol{\theta}^*)$. The latter result is a direct consequence of Corollary 1 of [50], by noting that $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is a quadratic function of $\boldsymbol{\theta}$ and then $D^{10}Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is continuous with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$.

Hardy prior is strictly convex. The log-likelihood $\log p_N(\mathbf{y}|\theta)$ is also convex although not necessarily strictly. Therefore, the log posterior using the Hardy prior is strictly convex and then the GEM sequence converges to the global maximizer $\boldsymbol{\theta}^*$. Other priors are strictly convex provided that $D^2L(\boldsymbol{\theta}) < 0$. This condition may be satisfied, or not, depending on the blur matrix $\mathbf{H}$ and on the noise variance σ^2. In general terms, $D^2L(\boldsymbol{\theta})$ decreases as the blur decreases and as noise variance approaches to zero.

C. Numeric Analysis

After these considerations, one is certainly convinced that the proposed GEM iterative scheme runs into numerical and convergence troubles when using sparse priors. This happens, in fact, in (14), since $d(\theta) \to \infty$ as $\theta \to 0$. Notice, however, the GEM iteration (22) depends on $d^{-1}(\theta)$, thus, being numerically stable.

Still, there is a question on the GEM convergence when using sparse priors: If a wavelet coefficient is initialized to zero, it remains zero, irrespective of the number of iterations and the initialization of the remaining wavelet coefficients. This is implied by $d^{-1}(\theta) = 0$ when the referred wavelet coefficient is zero and, thus, so is zero the correspondent diagonal entry of $\mathbf{C}_t^{-1}$. We have found out, however, in a series of experiments, that the sequence generated by (25) leads systematically to good results, providing that the wavelet coefficients are not initialized to zero. Moreover, given a nonsparse prior obtained from a sparse prior by replacing θ with $\sqrt{\theta^2+\alpha^2}$, it was also systematically observed that $\|\widehat{\boldsymbol{\theta}}_t(\alpha) - \widehat{\boldsymbol{\theta}}_t(0)\| \to 0$ and $|L(\widehat{\boldsymbol{\theta}}_t(\alpha), \alpha) - L(\widehat{\boldsymbol{\theta}}_t(0), 0)| \to 0$ as $\alpha \to 0$, where $L(\widehat{\boldsymbol{\theta}}_t(\alpha), \alpha)$ denotes the objective function obtained by replacing in the prior θ with $\sqrt{\theta^2+\alpha^2}$. That is to say that, in a heuristic sense, $L(\widehat{\boldsymbol{\theta}}_t(\alpha), \alpha)$ and $\widehat{\boldsymbol{\theta}}_t(\alpha), \alpha)$ are continuous on α, provided that the wavelet coefficients are not initialized to zero. This is illustrated in Section VII.

1) Eigenvalues λ_1 and λ_n: The optimal iteration parameters α, β, and β_0 depend on the extreme eigenvalues $\lambda_1(\mathbf{C}_t^{-1}\mathbf{A}_t)$ and $\lambda_N(\mathbf{C}_t^{-1}\mathbf{A}_t)$. Given that $\mathbf{C}_t^{-1}\mathbf{A}_t = \mathbf{I} - \mathbf{C}_t^{-1}[\mathbf{I} - \mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}] \succ 0$, then $0 < \lambda_i(\mathbf{C}_t^{-1}\mathbf{A}_t) = 1 - \lambda_i(\mathbf{C}_t^{-1}[\mathbf{I} - \mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}])$, for $i = 1, \ldots, N$. Noting that $0 < \lambda_N(\mathbf{C}_t^{-1}) \leq 1$ and that $0 \leq \lambda_N([\mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}]) \leq 1$ (we are assuming that the blur is normalized to unit volume), we have $0 \leq \lambda_i(\mathbf{C}_t^{-1}[\mathbf{I} - \mathbf{P}^T\mathbf{H}^T\mathbf{H}\mathbf{P}])$ and $0 < \lambda_i(\mathbf{C}_t^{-1}\mathbf{A}_t) \leq 1$, for $i = 1, \ldots, N$.

We take the approximation $\widetilde{\lambda}_1 = \text{mean}\,[\lambda_i^2(\mathbf{H}^T\mathbf{H})]$ and $\widetilde{\lambda}_N = 1$ for $\lambda_1(\mathbf{C}_t^{-1}A_t)$ and $\lambda_N(\mathbf{C}_t^{-1}A_t)$, respectively. Given that $\lambda_N \leq \widetilde{\lambda}_N$, the convergence of (22) is assured [see condition (23)]. In fact, $\widetilde{\lambda}_1$ is a measure of the bandwidth of the blur filter and exhibits a behavior similar to λ_1, as function of the blur strength: When the blur decreases ($\mathbf{H}$ approaches $\mathbf{I}$), the eigenvalues of $\mathbf{H}^T\mathbf{H}$ approaches 1 implying that $\widetilde{\lambda}_1$ also approaches 1; when the blur increases, most eigenvalues of $\mathbf{H}^T\mathbf{H}$ approaches 0 implying that $\widetilde{\lambda}_1$ approaches 0.

Since the GEM iterations (25) shrink many wavelet coefficients to zero or nearly to zero, we could have chosen a smaller value for $\widetilde{\lambda}_1$, thus closer to λ_1. We note, however, that when a wavelet coefficient gets close to zero, the linear system $\mathbf{A}_t\boldsymbol{\theta} = \mathbf{y}'$ can be solved by eliminating the correspondent line and the column in the system matrix $\mathbf{A}_t$ and the correspondent element in the observed data $\mathbf{y}'$. Therefore, we are interested in finding an estimate of $\lambda_1(\mathbf{C}_t^{-1}\mathbf{A}_t)$ with respect to a reduced system matrix, where lines and columns of $\mathbf{C}_t^{-1}\mathbf{A}_t$ correspondent to small wavelet coefficients were eliminated. We have found out empirically that $\widetilde{\lambda}_1 = \text{mean}\,[\lambda_i^2(\mathbf{H}^T\mathbf{H})]$ leads systematically to good results.

It should be stressed that, although the approximation for λ_1 and for λ_N might be rough, it assures that inequalities (23) are satisfied and it is good enough to boost the converge rate by an order of magnitude, when comparing with *the first-order iterative method* obtained by setting $\alpha = 1$ in (22) (see [48, Ch. 5]). This aspect is shown in Fig. 1, where a simulated image of size 64×64 composed by squares of different dimensions multi-

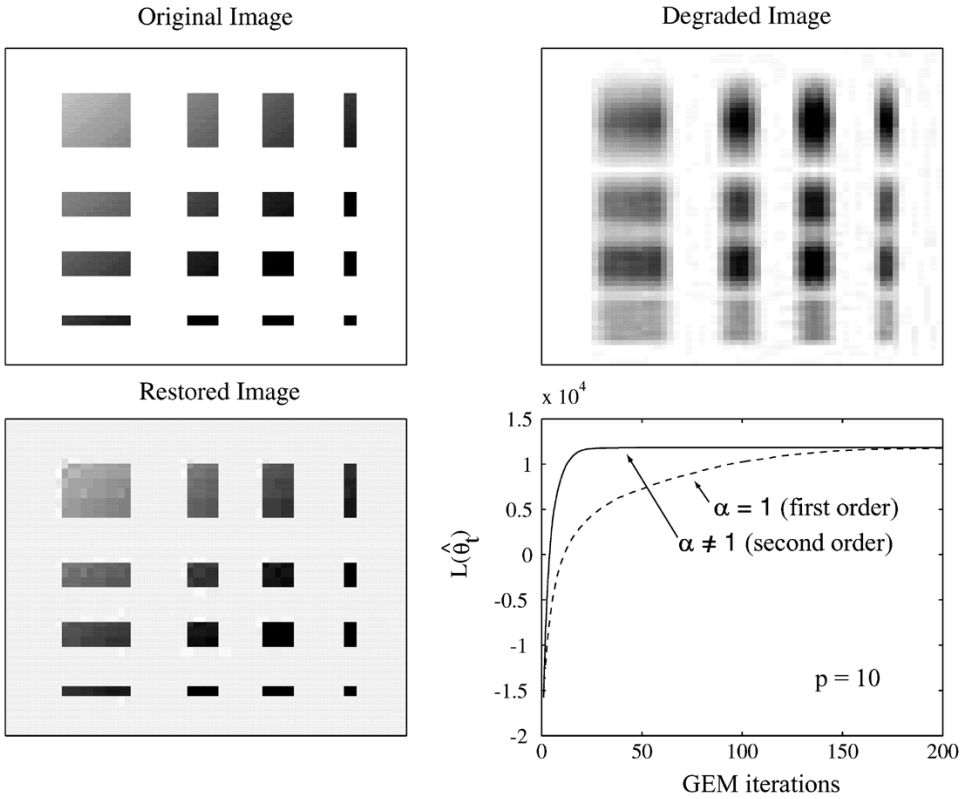

Fig. 1. Impact of the iteration parameters (α, β) on the WaveGSM rate of convergence. Top left: original image. Top right: Degraded image (uniform blur 3×9, $\mathrm{BSNR} = 40$ dB). Bottom left: WaveGSM restored image (DWT, Haar wavelets, Garrote prior, O-step iterations $p = 10$). Bottom right: Log-likelihood $L(\widehat{\boldsymbol{\theta}}_t)$ for $(\alpha = 1, \beta = \beta_0)$ and for (α, β) given by (24), corresponding to first and second-order iterative methods, respectively.

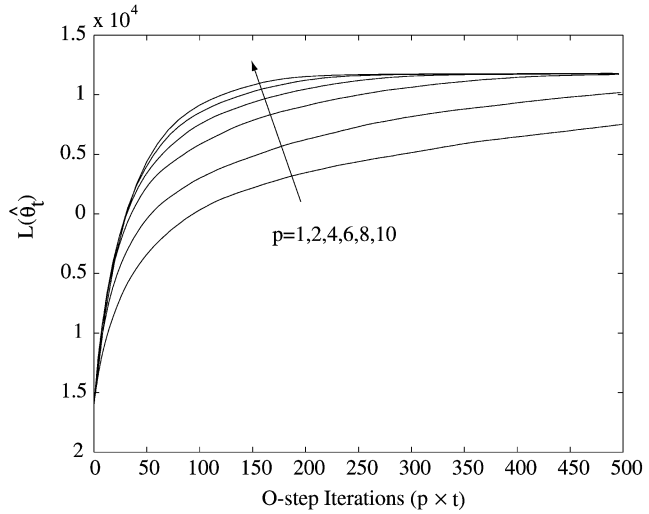

Fig. 2. Evolution of $L(\widehat{\boldsymbol{\theta}}_t)$ as function of the total number of iterations $p \times t$, parameterized by p (number of iterations in the O-step). Note that $\widehat{\boldsymbol{\theta}}_t$ takes $p \times t$ iterations to be computed.

TABLE III
BLUR, NOISE STANDARD DEVIATION, AND BSNR

	blur	σ	BSNR (dB)
Exp1	9×9 uniform	0.56	40
Exp2	$h_{ij} = (1 + i^2 + j^2),\ i, j = -7, \ldots, 7$	$\sqrt{2}$	31.85
Exp3	$h_{ij} = (1 + i^2 + j^2),\ i, j = -7, \ldots, 7$	$\sqrt{8}$	25.85
Exp4	$[1, 4, 6, 4, 1]^T [1, 4, 6, 4, 1]/256$	7	17

plied by the plane $i + j$, for $i, j = 1, \ldots, 64$, was restored by the WaveGSM algorithm using Haar wavelets; the blur is 3×9 uniform and $\mathrm{BSNR} = 40$ dB (BSNR is the SNR with respect to the blurred image). The second-order iterative algorithm converges in about 20 iterations, whereas the first-order one takes about 200 iterations.

2) Number of Inner Iterations P: The O-step of WaveGSM algorithm runs, at least, p times iteration (22). How to set p is a pertinent question. We have found out that the total number of iterations the algorithm takes to converge is not very sensitive to p, as long as $p \gtrapprox 4$. This is illustrated in Fig. 2, where the evolution of $L(\widehat{\boldsymbol{\theta}}_t)$ is plotted, parameterized by p. The abscissa axis represents the total number of iterations, $p \times t$, necessary to compute $\widehat{\boldsymbol{\theta}}_t$.

The explanation for this behavior is that each time WaveGSM runs the O-step it implements a first-order iteration and $p - 1$ second-order iterations. Therefore, given a fixed number of iterations $p \times t$, higher values of p means less first-order iterations and, thus, higher rates of convergence of the O-step.

As stated in Section V, the rate of convergency of the O-step depends on the extreme values of matrix $\mathbf{C}_t^{-1}\mathbf{A}_t$. In the previous section, we give evidence that these extreme eigenvalues are strongly related to the extreme eigenvalues of matrix $\mathbf{H}^T\mathbf{H}$. If the image width and length are larger than the blur width and

TABLE IV
SNR IMPROVEMENTS (ISNR) OBTAINED IN THE FIRST EXPERIMENT (CAMERAMAN, 9×9 UNIFORM BLUR, $\mathrm{BSNR} = 40$ dB, $\sigma = 0.56$). COLUMNS σ, $\widehat{\sigma}_{\mathrm{IE}}$, AND $\widehat{\sigma}_{\mathrm{MAD}}$ CORRESPOND TO THE WAY NOISE VARIANCE IS DEALT WITH (RESPECTIVELY, TRUE VALUE, ITERATIVELY ESTIMATED, AND MAD ESTIMATE)

	WaveGSM			WaveGSM_TI			WaveGSM_TIR		
ISNR(dB)	σ	$\widehat{\sigma}_{IE}$	$\widehat{\sigma}_{MAD}$	σ	$\widehat{\sigma}_{IE}$	$\widehat{\sigma}_{MAD}$	σ	$\widehat{\sigma}_{IE}$	$\widehat{\sigma}_{MAD}$
Garrote	6.45	5.80	5.45	8.14	7.65	7.08	8.10	8.01	7.65
Laplace	6.78	6.70	6.25	7.56	7.19	6.96	7.79	7.80	7.30
Hardy	6.85	6.81	6.30	7.53	6.83	7.52	7.79	7.80	7.30
Jeffreys	6.65	6.50	6.20	8.53	8.54	8.14	7.75	7.90	8.00
GM	6.70	6.53	5.7	8.16	8.49	8.0	7.75	7.80	7.40

length, respectively, then the extreme eigenvalues of $\mathbf{H}^T\mathbf{H}$ depend little on the image size N (see [51] for the unidimensional case). Therefore, under the these circumstances, the O-step convergency rate depends very little on N.

VII. EXPERIMENTAL RESULTS

We now present a set of five experiments illustrating the performance of the WaveGSM algorithm and its TI variants. Daubechies wavelets are used in all experiments. Periodic boundary is assumed. Original images are cameraman (experiments 1–3) and lena (experiment 4) both of size 256×256. Table III displays the blur, the noise, and the BSNR for each of the four experiments. These scenarios replicate those used in the evaluation of state-of-the-art methods [14], [19], [23], [24], [43] with which we compare the proposed approach.

Five GSM priors are compared: garrote, Laplace, Hardy, Jeffreys, and GM. Garrote prior is parameterized (see Tables I and II) with $\lambda = \sqrt{3}\sigma$ leading to the denoising thresholding rule (10) with threshold $\lambda = \sqrt{3}\,\sigma$. This rule yields very good results as shown in [33]. Laplace prior is parameterized with $\lambda = \beta/\sigma$ corresponding to the soft-threshold denoising rule with threshold $\sigma\beta$. Hardy prior is also parameterized with $\lambda = \beta/\sigma$ and $\alpha = 1$. Owing to problems at the origin, the Jeffreys prior is in fact a nonsparse version of the original one (i.e., $1/\sqrt{\theta^2 + \alpha^2}$ instead of $1/|\theta|$), with $\alpha = 1$.

The GM contains three zero-mean modes (Gaussian densities). Model parameters (variances and weights) are updated, as described in last paragraph of Section V-A, after step 14 of Algorithm 2. The updating corresponds exactly to an iteration of the EM algorithm for GMs [52], [53] with the mean of the modes forced to zero. In practice, we obtained better results by updating only the mode weights and, therefore, keeping constant the mode variances. What is necessary is to have a mode with small variance, accounting for most wavelet coefficients, and another with larger variance, accounting, for a few large wavelet coefficients. We set $\sigma_i^2 = 1,\ 100,\ 3000$ for $i = 1, 2, 3$ and 8-bit images. The mode with variance 100 increases the degrees of freedom of the model.

The soft-threshold denoising rule with threshold $\sigma\beta$ implicit in the Laplace prior is formally identic to the *universal threshold* $\sigma\sqrt{2\log N}$ proposed in [49]; however, since the quantity $\sqrt{2\log N}$ is too large for most real word images leading to oversmoothing [29], [34], we have chosen the value of β yielding the best estimate in terms of mean-squared error. Of course, this procedure can not be followed with real data and we adopt it only for comparison purposes: as illustrated below, even fine tuning β for each experiment, garrote and Jeffreys priors (with fixed parameters) yield similar or better estimates than Laplace and Hardy priors.

The stop criterion (i.e., *StopRule* in the WaveGSM algorithm) is

$$\frac{L(\widehat{\boldsymbol{\theta}}_t) - L(\widehat{\boldsymbol{\theta}}_{t-1})}{|L(\widehat{\boldsymbol{\theta}}_t)|} \leq \delta \tag{27}$$

with $\delta = 0.01$. The number of O-step iterations is set to $p = 10$. The shifts in WaveGSM_TI algorithm take values in the set[4] $\{(i,j) \in \mathbb{N}_0^2 |\, i,j = 0,1,\ldots,\sqrt{N_s} - 1\}$. The total number of shifts is set to $N_s = 16$.

The estimation results presented in this section are based on the mean squared error per pixel of the image estimate. The squared error per pixel has very small variance, as it is based on sample averages taken over images of size $N = 65\,536$. For this reason, we compute experimental results based on only one run of the respective algorithm.

A. First Experiment

Table IV shows the obtained signal-to-noise improvements ($\mathrm{ISNR} \equiv E\|\mathbf{y} - \mathbf{x}\|^2 / E\|\widehat{\mathbf{x}} - \mathbf{x}\|^2$) using Harr wavelets (Daubechies-2). Columns σ, $\widehat{\sigma}_{\mathrm{IE}}$, and $\widehat{\sigma}_{\mathrm{MAD}}$ correspond to the way noise variance is dealt with (respectively, true value, iteratively estimated, and MAD estimate). We see from this table that translation invariant methods outperform the non-translation ones by, approximately, 2 dB. This conclusion is in line with many recent findings on this matter. The best ISNR is achieved by the WaveGSM_TI algorithm using the Jeffreys prior, although WaveGSM_TIR with garrote, Jeffreys, or GMs priors yields comparable performance. Laplace and Hardy priors perform only a little worse than garrote and Jeffreys ones. Notice that the WaveGSM_TIR algorithm yields exactly the same ISNR when Laplace or Hardy priors are used. This raises the question whether sparseness is a desirable feature of priors in image restoration. In fact, as shown in Section IV, any sparse prior can be converted into a nonsparse one, with a very small modification of the provided MAP estimates.

[4]The couple (i, j) means a shift of i lines and of j columns.

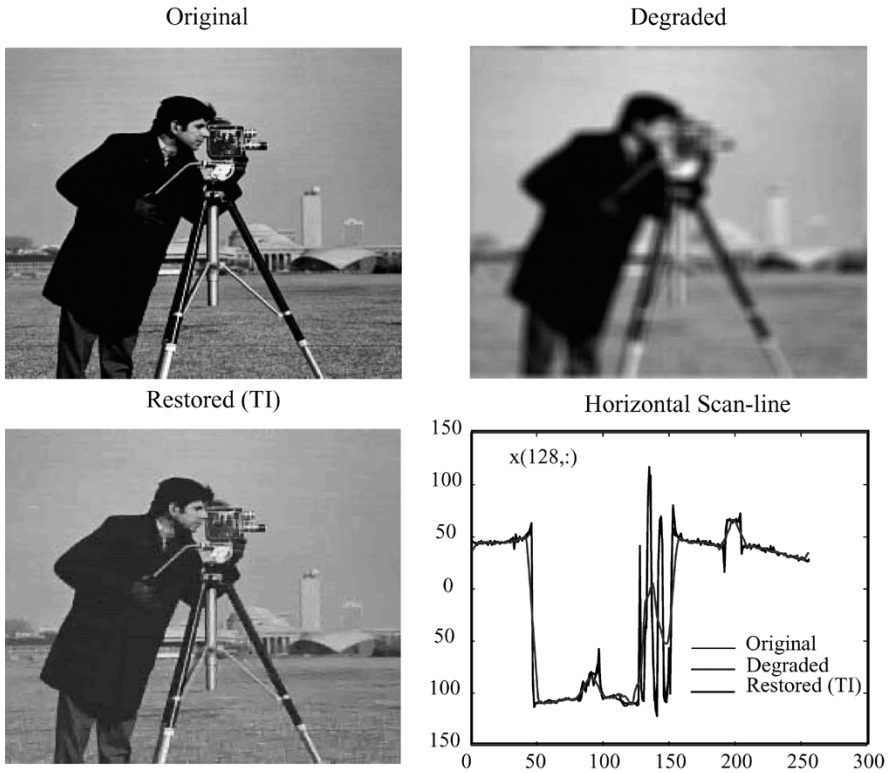

Fig. 3. First experiment. Top left: Original image. Top right: Blurred noisy image [blur (9×9) uniform, $\mathrm{BSNR} = 40$ dB]. Bottom left: Restored image with WaveGSM_TI algorithm and the Jeffreys prior ($\mathrm{ISNR} = 8.54$ dB). Bottom right: Scan-line showing original, degraded, and estimated data.

Fig. 3 shows the original cameraman image (top left), the blurred noisy image (top right), the WaveGSM_TI restored image using Jeffreys prior, corresponding to $\mathrm{ISNR} = 8.54$ dB (bottom left), and a scan-line of the original, degraded, and estimated data (bottom right). Notice that the restored image exhibits almost no ringing and that sharp transitions, as those in the camera neighborhood, are preserved. This is a result of using heavy-tailed priors on the wavelet coefficients, since sharp transitions are well described by a few large wavelet coefficients.

We repeated Experiment 1 with Daubechies-4 and Daubechies-6 wavelets. As with Daubechies-2, the best results were obtained using WaveGSM_TI and WaveGSM_TIR algorithms. However, the values of ISNR were between 0.3 and 0.8 dB below those obtained with Daubechies-2.

Based on the above results, in the remaining experiments we only use Daubechies-2 (Harr) wavelets and compare WaveGSM_TI and WaveGSM_TIR algorithms using Garrote, Jeffreys, and GM priors. The noise is assumed known, since the difference in performance among the three ways of dealing with noise is very little.

B. Second Experiment

Table V shows the ISNR obtained in the second experiment. The best results are obtained with Garrote and Jeffreys priors and the WaveGSM_TIR algorithm. In comparison with these priors, the GM performs a little better with the WaveGSM_TI algorithm and a little worse with the WaveGSM_TIR algorithm.

TABLE V
ISNR OBTAINED IN THE SECOND EXPERIMENT (CAMERAMAN, POINT-SPREAD FUNCTION OF THE BLUR $h_{ij} = (1 + i^2 + j^2)$, FOR $i, j = -7, \ldots, 7$, $\sigma^2 = 2$ CORRESPONDING TO A $\mathrm{BSNR} = 31.85$ dB)

ISNR(dB)	WaveGSM_TI	WaveGSM_TIR
Garrote	7.06	7.40
Jeffreys	7.04	7.32
GM	7.20	7.00

Fig. 4 shows the blurred noisy image (top left), the WaveGSM_TIR restored image using garrote prior, corresponding to $\mathrm{ISNR} = 7.40$ dB (top right), the WaveGSM_TI restored image using garrote prior, corresponding to $\mathrm{ISNR} = 7.06$ dB (bottom left), and scan-lines of the original and of the estimated data. Notice that the estimate determined by the WaveGSM_TI algorithm is a little bit more smooth on the textured areas.

Fig. 5 left shows the ISNR as function of the number of shifts in the WaveGSM_TI algorithm. The larger increments in the ISNR occurs in the first eight shifts. We have observed experimentally that a number of shifts higher than 16 improve the ISNR, but only by a small factor. Fig. 5 right shows the ISNR relative first difference of $L(\widehat{\boldsymbol{\theta}}_t)$ as function of the GEM number of iterations in the WaveGSM_TIR algorithm. The ISNR increases up to its maximum and then decreases slowly. This is not a surprise as we are not maximizing the ISNR but rather the log-posterior. The evolution of $[L(\widehat{\boldsymbol{\theta}}_t) - L(\widehat{\boldsymbol{\theta}}_{t-1})]/|L(\widehat{\boldsymbol{\theta}}_t|$ and

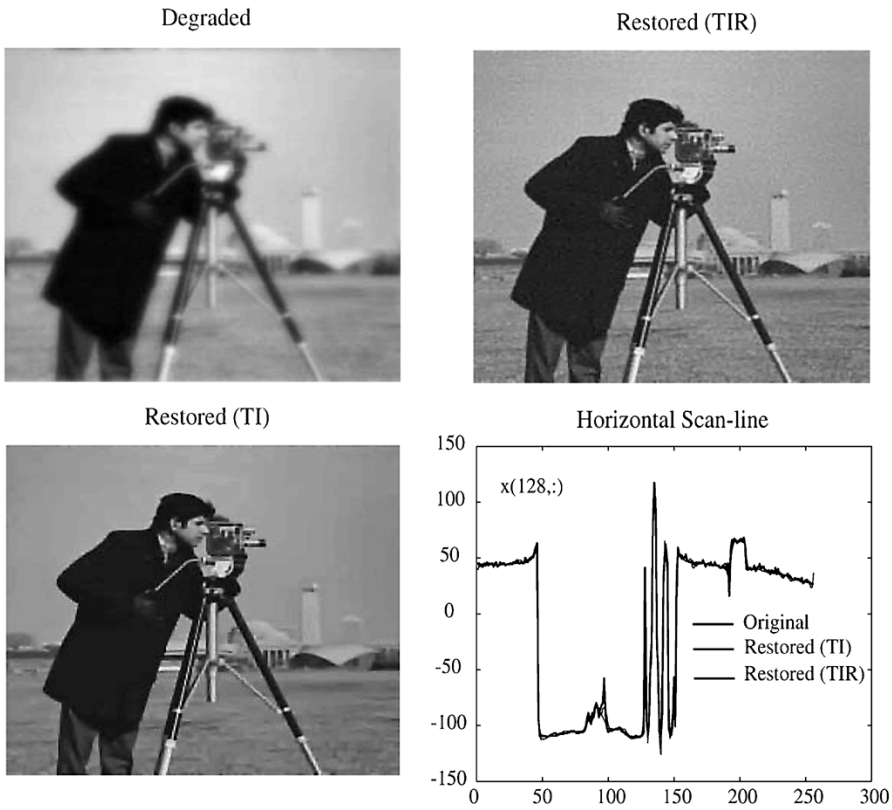

Fig. 4. Second experiment. Top left: Blurred noisy image [blur $h_{ij} = (1+i^2+j^2)$, for $i, j = -7, \ldots, 7$, and $\sigma^2 = 2$ corresponding to a $\mathrm{BSNR} = 31.85$ dB]. Top right: Restored image with WaveGSM_TIR (garrote prior) algorithm ($\mathrm{ISNR} = 7.40$ dB). Bottom left: Restored image with WaveGSM_TI (garrote prior) algorithm ($\mathrm{ISNR} = 7.06$ dB). Bottom right: Scan-line showing original data and estimated data with WaveGSM_TIR and WaveGSM_TIR algorithms.

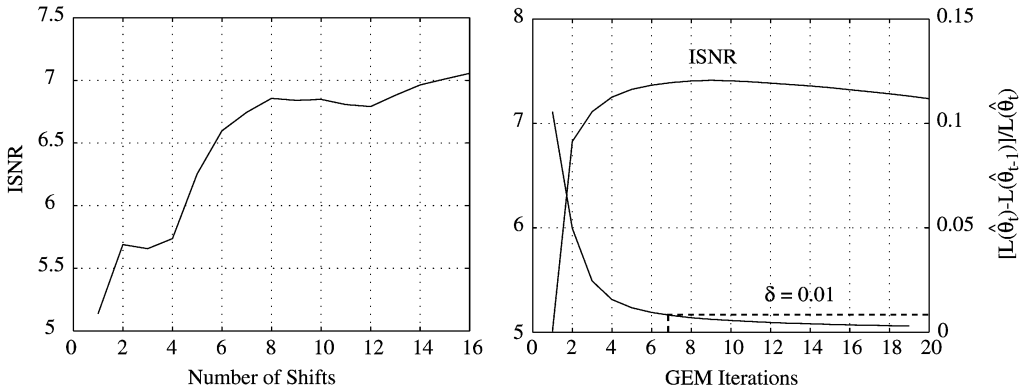

Fig. 5. Second experiment. Left: ISNR as function of the number of shifts in the WaveGSM_TI algorithm. Right: ISNR and relative first difference of $L(\widehat{\boldsymbol{\theta}}_t)$ as function of the GEM number of iterations in the WaveGSM_TIR algorithm.

the coordinates at which it decreases below the threshold δ are also plotted.

C. Third Experiment

Table VI shows the ISNR obtained in the third experiment. As in the previous experiments, garrote and Jeffreys priors originate identical results. WaveGSM_TIR algorithm performs a little better than WaveGSM_TI one.

D. Fourth Experiment

The blur used in this experiment is the weakest among the four experiments and is not far from a denoising only problem. For this reason, we initialize the deblurring algorithms with the degraded image $\mathbf{y}$, instead of the Wiener estimate, this leading to fewer GEM iterations.

TABLE VI
ISNR OBTAINED AS IN THE SECOND EXPERIMENT, EXCEPT FOR THE NOISE STANDARD DEVIATION WHICH IS SET TO $\sigma^2 = 8$, CORRESPONDING TO A $\mathrm{BSNR} = 25.85$ dB

ISNR(dB)	WaveGSM_TI	WaveGSM_TIR
Garrote	4.90	5.15
Jeffreys	5.00	5.30
GM	5.11	4.50

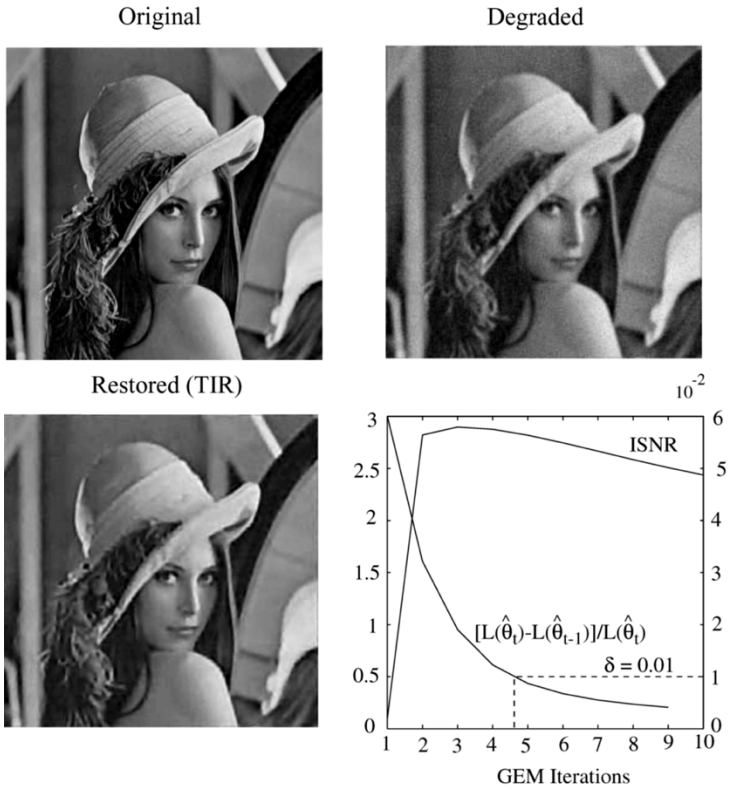

Fig. 6. Fourth experiment. Top left: Original image. Top right: Blurred noisy image (blur $[1,4,6,4,1]^T[1,4,6,4,1]/256$ and $\sigma = 7$, corresponding to a $\mathrm{BSNR} = 17$ dB). Bottom left: Restored image with WaveGSM_TIR (garrote prior) algorithm ($\mathrm{ISNR} = 2.85$ dB). Bottom right: ISNR and relative first difference of $L(\widehat{\boldsymbol{\theta}}_t)$ as function of the GEM number of iterations in the WaveGSM_TIR algorithm.

TABLE VII
ISNR OBTAINED IN THE FOURTH EXPERIMENT (LENA, BLUR $[1,4,6,4,1]^T[1,4,6,4,1]/256$, AND $\sigma = 7$, CORRESPONDING TO A $\mathrm{BSNR} = 17$ dB

ISNR(dB)	WaveGSM_TI	WaveGSM_TIR
Garrote	2.7	2.85
Jeffreys	1.92	2.80
GM	2.46	2.8

ISNR(dB)	WaveGSM_TIR
Original	3.0
Degraded	2.84
Wiener	2.82

Table VII shows the ISNR obtained in the fourth experiment. The two algorithms display identical results with the garrote prior.

Fig. 6 shows the original image (top left), the blurred image (top right), the restored image with the WaveGSM_TIR algorithm and garrote prior corresponding to a $\mathrm{ISNR} = 2.85$ dB (bottom left) and the relative first difference of $L(\widehat{\boldsymbol{\theta}}_t)$ as function of the GEM number of iterations in the WaveGSM_TIR algorithm (bottom right).

From the four experiments presented, we see that the two algorithms yield comparable estimates. Concerning priors, although there is not a single winner, the garrote prior exhibits higher consistency than Jeffreys and GM priors.

E. Fifth Experiment

This experiment compares the performance of the WaveGSM_TIR algorithm and the garrote prior under different initializations: original image, degraded image $\mathbf{y}$, and Wiener filter. The setup is as in Experiment 4.

The obtained ISNR, shown in Table VII, depends a little bit on the initialization. This was to be expected, as the the garrote prior is not convex. Note, however, that similar results are obtained initializing the WaveGSM_TIR algorithm with the degraded image $\mathbf{y}$ or the Wiener image $\widehat{\mathbf{x}}_0 = (\sigma^2/\sigma_x^2\mathbf{I} + \mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$. As expected, the best figure is obtained initializing the algorithm with the original data.

F. Computational Complexity

The WaveGSM algorithm and its translaction invariant variants were implemented in MATLAB. Table VIII shows, per experiment, the time each algorithm took and the number of GEM iterations in the case of the WaveGSM_TIR version. The time dynamic range of the TIR version is larger than the TI one. The number of GEM iterations increases with the blur strength.

G. Comparison With State-of-the-Art Methods

Table IX shows the ISNR of the proposed WaveGSM_TIR algorithm using the garrote prior and of the methods [14], [23], [24], [43], [19], for the experiments 1–4. The proposed method yields the best ISNR in experiments 1–3. In experiment 4, the

TABLE VIII
TIME IN SECONDS AND GEM ITERATIONS

	WaveGSM_TI	WaveGSM_TIR	
	Time	Time	GEM Iter.
Exp1	260	600	60
Exp2	148	60	6
Exp3	119	97	10
Exp4	80	38	4

TABLE IX
ISNR OF THE PROPOSED WAVEGSM_TIR ALGORITHM USING THE GARROTE PRIOR AND OF THE METHODS [14], [19], [23], [24], [43] FOR THE FOUR EXPERIMENTS

	ISNR (dB)			
Method	Exp1	Exp2	Exp3	Exp4
Algoritm 1	8.10	7.40	5.15	2.85
Figueiredo & Nowak [24]	7.59	6.93	4.88	2.94
Neelamani et al. [43]	7.30	–	–	–
Banham & Katsaggelos [14]	6.70	–	–	–
Jalobeanu et al. [23]	–	6.75	4.85	–
Liu & Moulin [19]	–	–	–	1.08

ISNR is only 0.09 dB below the value obtained with the best competitor, which is the algorithm published in [24].

The algorithm in [24] belongs also to the EM class and was designed with the objective of exploiting previous MAP denoising wavelet-based approaches, which, typically, leads to wavelet thresholding/shrinking rules depending on the prior. This objective was achieved by rewriting the observation (1) as three additive terms, one of then being interpretable as missing data and the remaining ones interpretable as noise terms. This decomposition led to E and M steps implementing one Landweber iteration and a denoising step, respectively.

Our EM approach adopts a perspective rather different from that of [24]: The missing variables were designed in connection with the prior to obtain a quadratic problem in each M step. The E step consists in recomputing a diagonal matrix playing the role of weights in a reweighted least squares type iterative scheme. From Table IX, we see that the gain in ISNR of the proposed method over the method [24] increases with the blur strength. The gain in time that the algorithm takes increases also with the blur strength. For example, in Experiment 1, the proposed algorithm reaches $\mathrm{ISNR} = 7.45$ (note that this is not the final ISNR) in ten GEM iterations and 60 s, whereas the algorithm [24] takes 600 EM iterations and 3000 s to achieve the same ISNR.

VIII. CONCLUDING REMARKS

We developed a new wavelet-based algorithm to image deconvolution. The problem was formulated in the wavelet domain following a Bayesian approach. The observed images were assumed to be degraded by space-invariant blur and additive Gaussian noise. The wavelet coefficients were assumed to be independent with density given by a GSM. This set of densities contains many heavy-tailed priors adopted in image restoration of real-world images, namely the generalized Gaussian class, the Jeffreys noninformative prior, and the GM. We have shown that the prior induced by the *garrote* thresholding rule is also a GSM. Furthermore, we stated necessary and sufficient conditions under which the prior induced by a thresholding/shrinking denoising rule is a GSM.

To compute the MAP estimate, we developed an EM algorithm, termed WaveGSM, where the missing variables are the scale factors of the prior GMs. We have shown that the E-step can be directly obtained from the prior without the explicit knowledge of the GSM decomposition. The maximization step of the EM algorithm includes a huge nondiagonal linear system with unbearable computational complexity. To avoid this difficulty, we approximated the linear system solution by a few iterations of a *linear stationary second-order iterative method*. The resulting scheme was a GEM algorithm, achieving convergence in a few tens of iterations. The FFT and the DWT are the heaviest computations on each GEM step. Thus, the overall algorithm complexity is $O(N \log N)$.

To reduce the *blocky* artifacts associated to the dyadic shifts inherent to the orthogonal DWT basis functions, we have introduced two algorithms, both based on WaveGSM, that implement translation invariance: the WaveGSM_TI averages a few WaveGSM estimates computed from shifted versions of the original degraded image; the WaveGSM_TIR replaces DWT coefficients with DWT_TI ones (nondecimated wavelet coefficient). In a series of experiments, the proposed approach outperformed or performed similarly to with state-of-the art methods, demanding comparable (in some cases much less) computational complexity. The gains in ISNR and computation time, with respect to the best competitor, increase with the blur strength.

APPENDIX

The proof of Lemma 1, basically, exploits the relation $\phi'(\theta) = [T^{-1}(\theta) - \theta]/\sigma^2$ and the fact that the product of completely monotonic functions on $(0, \infty)$ is also completely monotonic. The see this, note that $(fg)^{(l)} = \sum_{t=1}^{l+1} \alpha_t f^{(m_t)} g^{(n_t)}$, where α_t, m_t, and n_t are nonnegative integers and $m_t + n_t = l$. Therefore, $(-1)^l (fg)^{(l)} \geq 0$, since $(-1)^l f^{(m_t)} g^{(n_t)} = [(-1)^{m_t} f^{(m_t)}][(-1)^{n_t} g^{(n_t)}] \geq 0$.

Proof: (Sufficient condition) To prove that $p_\theta(\theta) \propto e^{-\phi(\theta)}$ is a GSM, it is necessary to show that 1) $\lim_{\theta \to 0^+} \phi(\theta)$ exists, with $\phi(\theta)$ given by (9), and that 2) $(-1)^l p_\theta^{(l)}(\sqrt{\theta}) \geq 0$, for $\theta \in (0, \infty)$ and all natural l.

1) Since T is nondecreasing and T^{-1} exists in $(0, \infty)$, then $\lim_{\theta \to 0^+} T^{-1}(\theta)$ is bounded and $\lim_{\theta \to 0^+} \phi(\theta)$, with $\phi(\theta)$ given by (9), is also bounded.

2) From (8), we have for $\theta > 0$, $\phi'(\theta) = [T^{-1}(\theta) - \theta]/\sigma^2$ and, therefore,[5] $\phi'(\sqrt{\theta}) = f_T(\theta)/(\sigma^2\sqrt{\theta})$. Since $1/\sqrt{\theta}$ and $f_T(\theta)$ are completely monotones in $(0, \infty)$, then $\phi'(\theta)$ is also completely monotone in the same interval, meaning that $(-1)^{l-1}\phi^{(l)}(\sqrt{\theta}) \geq 0$, for $\theta > 0$ and $l = 1, 2, \ldots$ On the other hand, given a C^∞ function ψ, and after some algebra, we have $[e^\psi]^{(l)} = e^\psi \sum_{t=1}^{l} \alpha_t [\psi^{(m_t)}]^{n_t} [\psi^{(o_t)}]^{p_t}$, where α_t, m_t, n_t, o_t, and p_t are nonnegative integers and $m_t n_t + o_t p_t = l$. Therefore, $(-1)^l [e^{-\phi(\sqrt{\theta})}]^{(l)} \geq 0$, for $l = 1, 2, \ldots$, since $(-1)^l [-\phi^{(m_t)}(\sqrt{\theta})]^{n_t} [-\phi^{(o_t)}(\sqrt{\theta})]^{p_t} = [(-1)^{m_t - 1}\phi^{(m_t)}(\sqrt{\theta})]^{n_t} [(-1)^{o_t - 1}\phi^{(o_t)}(\sqrt{\theta})]^{p_t}$ and $(-1)^{l-1}\phi^{(l)}(\sqrt{\theta}) \geq 0$.

(Necessary condition) Admit that $f_T(\theta) < 0$ for some $\theta > 0$. Then, we have $(-1)^1 [e^{-\phi(\sqrt{\theta})}]^{(1)} = e^{-\phi(\sqrt{\theta})}\phi^{(1)}(\sqrt{\theta}) = e^{-\phi(\sqrt{\theta})} f_T(\theta)/(\sigma^2\sqrt{\sigma}) < 0$. Therefore, $e^{-\phi(\theta)}$ would not be a GSM. ■

[5] We recall that $f^{(n)}(\sqrt{\theta}) \equiv (d)/(d\theta)(f(\sqrt{\theta}))$.

REFERENCES

[1] J. Dias, "A fast GEM algorithm for Bayesian wavelet-based image restoration using a class of heavy-tailed priors," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, M. Figueiredo, A. Rangarajan, and J. Zerubia, Eds. Lisbon, Portugal: Springer, Jul. 2003, pp. 407–420.

[2] ——, "Fast GEM wavelet-based deconvolution algorithm," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Barcelona, Spain, 2003, pp. 961–964.

[3] A. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[5] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Nature*, vol. 317, pp. 314–319, 1985.

[6] D. Terzopoulos, "Regularization of inverse visual problems involving discontinuities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 4, pp. 413–424, Jul. 1986.

[7] A. Katsaggelos, *Digital Image Restoration*. New York: Springer-Verlag, 1991.

[8] A. Katsaggelos, J. Biemond, R. Schafer, and R. Mersereau, "A regularized iterative image restoration algorithm," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 914–929, Apr. 1991.

[9] F. Jeng and J. Woods, "Compound gauss-markov random fields for image estimation," in *IEEE Trans. Signal Process.*, vol. 39, Mar. 1991, pp. 683–697.

[10] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, MA: MIT Press, 1987.

[11] M. Nikolova, "Thresholding implied by truncated quadratic regularization," *IEEE Trans. Signal Process.*, vol. 48, no. 11, pp. 3437–3450, Nov. 2000.

[12] ——, "Local strong homogeneity of a regularized estimator," *SIAM J. Appl. Math.*, vol. 61, pp. 3437–3450, 2000.

[13] T. Wang, X. Zhuang, and G. Pan, "Solution of inverse problems in inverse in image processing by wavelet expansions," *IEEE Trans. Image Process.*, vol. 4, no. 5, pp. 579–593, May 1995.

[14] M. Banham and A. Katsaggelos, "Spatially adaptive wavelet-based multiscale image restoration," *IEEE Trans. Image Process.*, vol. 5, no. 5, pp. 619–634, May 1996.

[15] M. Belge, M. Kilmer, and E. Miller, "Wavelet domain image restoration with adaptive edge-preserving regularization," in *IEEE Trans. Image Process.*, vol. 9, Apr. 2000, pp. 597–608.

[16] P. Rivaz, "Complex Wavelet Based Image Analysis and Synthesis," Ph.D. dissertation, 2000.

[17] D. Donoho, "Nonlinear solution of linear inverse problems by wavelet-vaguelette decompositions," *J. Appl. Comput. Harmon. Anal.*, vol. 1, pp. 100–115, 1995.

[18] F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy. Stat. Soc. B*, vol. 60, pp. 725–749, 1998.

[19] J. Liu and P. Moulin, "Complexity-regularized image restoration," in *Proc. IEEE Int. Conf. Image Processing*, 1998, pp. 555–559.

[20] Y. Wan and R. Nowak, "A wavelet-based approach to joint image restoration and edge detection," presented at the SPIE Conf. Wavelet Applications in Signal and Image Processing VII, vol. 3813, Denver, CO, 1999.

[21] J. Kalifa and S. Mallat, "Minimax restoration and deconvolution," in *Bayesian Inference in Wavelet Based Models*, P. Muller and B. Vidakovic, Eds. New York: Springer-Verlag, 1999.

[22] J. Portilla, V. Strelaand, M. Wainwright, and E. Simoncelli, "Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain," in *Proc. IEEE Int. Conf. Image Processing*, 2001, pp. 7–10.

[23] A. Jalobeanu, N. Kingsbury, and J. Zerubia, "Image deconvolution using hidden Markov tree modeling of complex wavelet packets," presented at the IEEE Int. Conf. Image Processing, Thessaloniki, Greece, 2001.

[24] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[25] D. Geiger and F. Girosi, "Parallel and deterministic algorithms from MRF's: Surface reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 5, pp. 401–412, May 1991.

[26] J. Zerubia and R. Chellappa, "Mean field annealing using compound Gauss-Markov random fields for edge detection and image estimation," *IEEE Trans. Neural Netw.*, vol. 4, no. 4, pp. 703–709, Jul. 1993.

[27] M. Figueiredo and L. Leitão, "Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1089–1102, Aug. 1997.

[28] T. Simchony, R. Chellappa, and Z. Lichtenstein, "Graduated nonconvexity algorithm for image estimation using compound Gauss Markov field models," in *Proc. Inf. Conf. Acoustics, Speech, Signal Processing*, Glasgow, U.K., 1989, pp. 1417–1420.

[29] R. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*. Boston, MA: Birkhauser, 1997.

[30] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.

[31] P. Muller and B. Vidakovic, Eds., *Bayesian Inference in Wavelet-Based Models*. New York: Springer-Verlag, 1999.

[32] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *J. Comput. Graph. Stat.*, vol. 7, pp. 469–488, Dec. 1998.

[33] M. Figueiredo and R. Nowak, "Wavelet-based image estimation: An empirical bayes approach using Jeffreys' noninformative prior," *IEEE Trans. Image Process.*, vol. 10, no. 9, pp. 1322–1331, Sep. 2001.

[34] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized—Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 909–919, May 1999.

[35] C. Robert, *The Bayesian Choice. A Decision-Theoritic Motivation*. New York: Springer-Verlag, 1994.

[36] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, no. 11, pp. 1338–1351, Nov. 2003.

[37] H. Xie, L. Pierce, and F. Ulaby, "SAR speckle reduction using wavelet denoising and Markov random fiels modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2196–2212, Oct. 2002.

[38] R. Neelamani, H. Choi, and R. Baraniuk, "Wavelet-based deconvolution using optimally inversion for ill-conditioned systems," *Wavelet Appl. Signal Image Process.*, vol. 3169, pp. 389–399, Oct. 2001.

[39] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Roy. Stat. Soc.*, vol. 36, no. 99–102, 1974.

[40] K. Lange and J. Sinsheimer, "Normal/independent distributions and their applications in robust regression," *J. Comput. Graph. Stat.*, vol. 2, pp. 175–198, 1993.

[41] F. Girosi, "Models of noise and robust estimates," Artificial Intelligence Lab. (Memo 1287) and Center for Biological and Computational Learning (Paper 66), Mass. Inst. Technol., Cambridge, 1991.

[42] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.

[43] R. Neelamani, H. Choi, and R. G. Baraniuk, "ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 418–433, Feb. 2004.

[44] N. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals," *J. Appl. Comput. Harmon. Anal.*, vol. 10, no. 3, pp. 234–253, 2001.

[45] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics*. New York: Springer-Verlag, 1995, pp. 125–150.
[46] J. Pesquet and D. Leporini, "Bayesian wavelet denoising: Besov priors and nongaussian noises," *Signal Process.*, vol. 81, pp. 55–67, 2001.
[47] I. Schoenberg, "Metric spaces and completely monotone functions," *Ann. Math.*, vol. 39, pp. 811–841, 1938.
[48] O. Axelsson, *Iterative Solution Methods*. New York: Cambridge Univ. Press, 1996.
[49] D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[50] C. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, no. 1, pp. 95–103, 1983.
[51] R. M. Gray, "On the asymptotic eigenvalue distribution of toeplitz matrices," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 3, pp. 725–730, May 1972.
[52] K. McLachlan and K. Basford, *Mixture Models: Inference and Application to Clustering*. New York: Marcel Dekker, 1998.
[53] K. McLachlan and T. Krishnan, *The EM Algorithm and Its Extensions*. New York: Wiley, 1997.

José M. Bioucas Dias (S'87–M'95) received the E.E., M.Sc., and Ph.D. degrees in electrical and computer engineering from Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal, in 1985, 1991, and 1995, respectively.

He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, IST. He is also a Researcher with the Communication Theory and Pattern Recognition Group, Institute of Telecommunications. His research interests include remote sensing, signal and image processing, pattern recognition, and communications.