Wavelet-Based Image Estimation: An Empirical Bayes Approach Using Jeffreys' Noninformative Prior

Mário A. T. Figueiredo, *Senior Member, IEEE,* and Robert D. Nowak, *Member, IEEE*

***Abstract*—The sparseness and decorrelation properties of the discrete wavelet transform have been exploited to develop powerful denoising methods. However, most of these methods have free parameters which have to be adjusted or estimated. In this paper, we propose a wavelet-based denoising technique without any free parameters; it is, in this sense, a "universal" method. Our approach uses empirical Bayes estimation based on a Jeffreys' noninformative prior; it is a step toward *objective* Bayesian wavelet-based denoising. The result is a remarkably simple fixed nonlinear shrinkage/thresholding rule which performs better than other more computationally demanding methods.**

***Index Terms*—Bayesian estimation, empirical Bayes, hierarchical Bayes, image denoising, image estimation, invariance, Jeffreys' priors, noninformative priors, shrinkage, wavelets.**

I. Introduction

A. Background

WAVELETS and other multiscale analysis tools underlie many recent advances in key areas of signal and image processing, namely, approximation/representation, estimation, and compression (see, e.g., Mallat's [24] recent book and the many references therein). In these applications, two important properties of the discrete wavelet transform (DWT) of real-world signals and images are exploited: 1) it is *sparse*, i.e., a few large coefficients dominate the representation and 2) the coefficients tend to be much less correlated than the original data. These properties, together with the existence of fast implementations, make the DWT an excellent tool for many tasks (see [24]) and also for statistical applications (see [27] and [30], and the references therein). The basic approach to DWT-based signal/image processing consists in manipulating the DWT coefficients, rather than the signal samples themselves. This is done by following a three step program:

1) compute the DWT coefficients of the signal;
2) perform some specified processing on these coefficients;
3) compute the inverse DWT to obtain the "processed" signal.

Stimulated by the seminal work of Donoho and Johnstone [12], many authors have proposed denoising (or signal/image estimation) methods adopting this standard three step approach (see, for example, Mallat [24], Mihçak *et al.* [25], Moulin and Liu [26], Ogden [30], Vidakovic [34], Krim and Schick [21]). In particular for detail-preserving (or discontinuity-preserving) image estimation/denoising (the subject of this paper), wavelet-based approaches provide a very efficient alternative to Markov random field (MRF) based techniques (see [15] and references therein).

In the denoising context, the decorrelation property suggests processing the coefficients independently of each other; the sparseness (or "heavy-tailedness") property paves the way to the use of threshold/shrinkage estimators aimed at removing/attenuating those coefficients that are "small" relative to the noise level. The classical choices for performing thresholding/shrinkage of each DWT coefficient (proposed by Donoho and Johnstone [12], [13]) are the hard and soft thresholding functions; letting ω denote an arbitrary DWT coefficient of the observed signal/image, these functions are defined, respectively, as

$$\delta_\lambda^{\text{hard}}(\omega) = \begin{cases} 0, & \Leftarrow |\omega| \leq \lambda \\ \omega, & \Leftarrow |\omega| > \lambda \end{cases} \tag{1}$$

$$\delta_\lambda^{\text{soft}}(\omega) = \begin{cases} 0, & \Leftarrow |\omega| \leq \lambda \\ \text{sgn}(\omega)(|\omega| - \lambda), & \Leftarrow |\omega| > \lambda \end{cases} \tag{2}$$

where $\text{sgn}(\cdot)$ is the sign function [$\text{sgn}(x) = 1$, if $x \geq 0$, and $\text{sgn}(x) = -1$, if $x < 0$] and λ a threshold level. In Donoho and Johnstone's classical techniques, λ depends on the known (or estimated) noise standard deviation. Their simplest approach (*VisuShrink*) uses a common value of λ for all levels (scales) of the DWT decomposition, which is based on the so-called "universal threshold." More sophisticated level-dependent adaptive schemes have also been proposed (namely, Donoho and Johnstone's *SureShrink* [13]); adaptive techniques tend to outperform fixed rules at the cost of a higher computational burden.

Recently, wavelet-based denoising/estimation has been addressed using Bayesian methods. The basic idea is to formally model the relevant properties of the DWT coefficients with prior probability distributions [7], [10], [26], [34]. These priors, together with the likelihood function (noise model), produce posterior distributions. Estimation rules can then be derived via the

Manuscript received May 10, 2000; revised May 2, 2001. M. A. T. Figueiredo was supported in part by the Science and Technology Foundation (Portugal) under Grant TIT-1580. R. Nowak was supported in part by the National Science Foundation under Grant MIP-9701692, the Army Research Office under Grant DAAD19-99-1-0349, and the Office of Naval Research under Grant N00014-00-1-0390. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pierre Moulin.

M. A. T. Figueiredo is with the Instituto de Telecomunicações and the Department of Electrical and Computer Engineering, Instituto Superior Técnico, 1049-001 Lisboa, Portugal (e-mail: mtf@lx.it.pt).

R. D. Nowak is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77001 USA (e-mail: nowak@ece.rice.edu).

Publisher Item Identifier S 1057-7149(01)07472-3.

standard Bayesian decision-theoretic approach, after the specification of a loss function [32]. Bayesian techniques usually outperform other methods and are the state-of-the-art in wavelet-based denoising [10], [27], [34].

There are several open issues in wavelet-based denoising. In threshold/shrinkage methods, the choice of the particular nonlinearity (e.g., hard or soft) is somewhat arbitrary. Thresholds are often chosen for mathematical convenience, rather than motivated by physical or inferential considerations. Moreover, the standard choices of nonlinearity have certain drawbacks. The soft thresholding function yields systematically biased estimates because it shrinks coefficients regardless of how large they are. The hard thresholding function, on the other hand, produces less biased but higher variance estimates; it can also be unstable due to its discontinuous nature. To avoid these drawbacks, several other *ad-hoc* rules have been proposed. Let us mention Gao and Bruce's [19] "firm" rule which tries to retain the best of the hard and soft functions (requiring two threshold values, thus computationally much more expensive in terms of threshold selection) and, recently, the "nonnegative garrote" function (as suggested by Gao [18]), defined as

$$\delta_{\lambda}^{\text{garrote}}(\omega) = \begin{cases} 0, & \Leftarrow \ |\omega| \leq \lambda \\ \omega - \dfrac{\lambda^2}{\omega}, & \Leftarrow \ |\omega| > \lambda \end{cases} \tag{3}$$

which we will return to in Section V.

Bayesian methods do not use a fixed arbitrary nonlinearity; the priors on the wavelet coefficients are chosen with the goal of matching empirical coefficient distributions or obtaining Bayesian estimators that mimic the conventional nonlinear rules. However, Bayesian methods are usually computationally intensive and require either careful hand-tuning of the prior parameters or signal-adaptive schemes.

B. Contributions

We tackle the fundamental issues raised above by adopting a Bayesian perspective supported on noninformative Jeffreys' priors (see, for example, [3] or [32]).

Our approach can be seen as a step toward *objective* Bayesian wavelet-based denoising; the term "objective" means the use of priors that do not require any subjective input. If we can find a prior distribution that, in a certain sense, does not favor one signal over another, then any inferences derived from the resulting posterior distribution are solely due to the data. Accordingly, our approach mitigates the subjectiveness associated with other (Bayesian and non-Bayesian) denoising schemes. The type of noninformativeness we invoke expresses *amplitude-scale*[1] invariance, meaning that the units in which an image/signal is measured do not directly influence any inference made from it [32], [3]. In other words, the inference procedure tries to be invariant under changes of amplitude-scale. Maybe surprisingly, the result of our approach is a fixed nonlinear shrinkage/threshold rule which, nevertheless, clearly outperforms both VisuShrink and SureShrink; actually, it performs nearly as well as (sometimes even better than) much more computationally expensive Bayesian denoising methods in standard benchmark problems. Remarkably, in view of its very good performance, our rule is fixed (with no free parameters), thus it is as computationally inexpensive as possible (e.g., as simple as VisuShrink).

Our results seem to carry an important message in terms of natural image modeling. The good results achieved with our noninformative prior seem to suggest the presence of a type of invariance which has not been previously exploited in image denoising: amplitude-scale invariance. Other types of invariance, namely spatial-scale invariance (or self similarity), however, have received considerable attention (see Field [14] and Ruderman [33]).

In Section II, the denoising problem is described and notation introduced. In Section III, a new noninformative prior is proposed, based on which we derive, in Section IV, a novel empirical-Bayes denoising procedure that we call *amplitude-scale-invariant Bayes estimation* (ABE). In Section V we discuss the relation of our method with other approaches. The performance of the new rule is compared to that of other methods in Section VI. Section VII presents a Bayesian interpretation of the nonnegative garrote. Conclusions are given in Section VIII.

II. Problem Formulation

A. Wavelet-Based Denoising and the Sparseness Property

Suppose $\mathbf{y}$ is a noisy observed signal or image, modeled as

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \tag{4}$$

where $\mathbf{x}$ is the underlying original signal (or image) and $\mathbf{n}$ contains independent samples of a zero-mean Gaussian variable of variance[2] σ^2; that is, $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, with $\mathbf{I}$ denoting an identity matrix of appropriate size. The goal of denoising (signal/image estimation) is to recover $\mathbf{x}$ from the observed $\mathbf{y}$.

In wavelet-based denoising, the orthogonal DWT, denoted by $\mathcal{W}$ (either 1-D or 2-D; see, e.g., [24] for details) is applied to the noisy data yielding the noisy *wavelet coefficients* $\boldsymbol{\omega}$; these are described by an analogous model

$$\boldsymbol{\omega} \equiv \mathcal{W}\mathbf{y} = \mathcal{W}\mathbf{x} + \mathcal{W}\mathbf{n} = \boldsymbol{\theta} + \mathbf{n}' \tag{5}$$

where $\boldsymbol{\theta} = \mathcal{W}\mathbf{x}$, and $\mathbf{n}' = \mathcal{W}\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, since $\mathcal{W}$ is orthogonal (i.e., $\mathcal{W}\mathcal{W}^T = \mathbf{I}$).

As mentioned above, the wavelet transforms of most real-world signals and images tend to be dominated by a few large coefficients [12]. This is the so-called *sparseness* property which, in probabilistic terms, corresponds to a wavelet coefficient density function with a marked peak at zero and heavy tails; that is, a strongly non-Gaussian density (also called super-Gaussian). Interestingly, it has recently been found that the human visual system exploits this sparseness property by using wavelet-like representations (see, for example, the recent work by Olshausen and Field [31], and Hyvärinen [20], and references therein). On the other hand, the DWT of Gaussian

[1]Throughout this paper, we use the term "*amplitude-scale*," in place of simply *scale*, to clearly distinguish it from the more common usage of the term *scale* (meaning spatial-scale) in wavelet theory.

[2]In this paper, we assume known noise variance; this is not a shortcoming because excellent estimates can be easily obtained directly from the noisy data using, e.g., the MAD scheme [13].

white noise produces i.i.d. Gaussian distributed coefficients; with high probability, these are bounded in magnitude by a suitable threshold proportional to their standard deviation. Therefore, a natural denoising criterion results from this statistical difference between the coefficients of the signal and the noise: if the magnitude of an observed wavelet coefficient is large, its signal component is probably much larger than the noise and it should be kept; conversely, if a coefficient has small absolute value, it is probably due to noise and it should be attenuated or even removed. This (together with the decorrelation property that suggests processing the coefficients independently of each other) is the rationale underlying the now classical thresholding methods introduced by Donoho and Johnstone [12].

The choice of wavelet basis does not effect the procedure we develop in this paper, but it does play a significant role in wavelet denoising performance. Suppose that the signal or image under consideration belongs to a function space with smoothness parameter α, e.g., a Besov space. If the underlying wavelet has $r > \alpha$ vanishing moments, then the best n-term approximation (i.e., keep only the n largest terms in the signal's wavelet expansion) converges at a rate of $O(n^{-\alpha})$ [11]. The smoother the target function, the more vanishing moments we require of the wavelet. There is also interesting connection between smoothness spaces and the choice of informative Bayesian priors [1].

Finally, we mention that there is also a conceptual link between wavelet-based denoising and *independent component analysis* (ICA); see Cardoso [6], Comon [9], Bell and Sejnowski [2], and the recent book by Lee [22]. The goal of ICA is to recover independent sources (signals) given only unknown (memoryless) linear combinations of them; ICA is possible only if no more than one of the mixed signals is Gaussian, and all the others are non-Gaussian. From an ICA perspective, wavelet-based denoising may be seen as a way of separating two sources (signal and noise) by representing them on a basis where one becomes strongly non-Gaussian (the signal) and the other remains Gaussian (the noise). However, while wavelet-based denoising usually adopts fixed bases, ICA adaptively looks for bases that best reveal the non-Gaussian nature of the source(s).

B. Bayesian Formulation

The likelihood functions resulting from the observation models in the signal and wavelet domains, respectively (4) and (5), are both Gaussian with covariance $\sigma^2\mathbf{I}$:

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}\left(\mathbf{x}, \sigma^2\mathbf{I}\right) \tag{6}$$

$$\boldsymbol{\omega}|\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{I}\right) \tag{7}$$

that is, the noise is white and Gaussian both in the signal and wavelet domains. To build a Bayesian framework that exploits the sparseness and decorrelation properties of the DWT, a prior $p_\Theta(\boldsymbol{\theta})$ is formulated with respect to the wavelet coefficients. Of course, this prior $p_\Theta(\boldsymbol{\theta})$ induces a signal prior given by $p_X(\mathbf{x}) = p_\Theta(\mathcal{W}\mathbf{x})$, because $\mathcal{W}$ is an orthogonal transformation, thus possessing a unit Jacobian (i.e., $|d\boldsymbol{\theta}| = |d\mathbf{x}|$).

The standard Bayesian version of the three step wavelet-based denoising program is:

1) compute the DWT of the data $\boldsymbol{\omega} = \mathcal{W}\mathbf{y}$;
2) obtain a Bayes estimate $\hat{\boldsymbol{\theta}}$, given $\boldsymbol{\omega}$;
3) reconstruct the signal estimate $\hat{\mathbf{x}} = \mathcal{W}^{-1}\hat{\boldsymbol{\theta}}$.

To interpret this procedure from a Bayesian decision theory perspective, let us explicitly write down $\hat{\boldsymbol{\theta}}$ as the minimizer of the *a posteriori* expected loss (see [32] or [3]); then

$$\hat{\mathbf{x}} = \mathcal{W}^{-1} \arg\min_{\tilde{\boldsymbol{\theta}}} \int L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) p(\boldsymbol{\theta}|\boldsymbol{\omega})\, d\boldsymbol{\theta}. \tag{8}$$

In (8), $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is the adopted loss function that penalizes the "discrepancy" between $\boldsymbol{\theta}$ and any candidate estimate $\tilde{\boldsymbol{\theta}}$, while $p(\boldsymbol{\theta}|\boldsymbol{\omega})$ is the *a posteriori* probability density function obtained via Bayes law $p(\boldsymbol{\theta}|\boldsymbol{\omega}) = p(\boldsymbol{\omega}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\boldsymbol{\omega})$. Now, recalling that $|d\boldsymbol{\theta}| = |d\mathbf{x}|$, and since

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p_X(\mathbf{x}) &= p(\boldsymbol{\omega}|\theta)p_X\left(\mathcal{W}^{-1}\boldsymbol{\theta}\right) \\ &= p(\boldsymbol{\omega}|\theta)p_\Theta(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}|\boldsymbol{\omega}). \end{aligned}$$

Equation (8) is equivalent to

$$\hat{\mathbf{x}} = \arg\min_{\tilde{\mathbf{x}}} \int L(\mathcal{W}\mathbf{x}, \mathcal{W}\tilde{\mathbf{x}}) p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}. \tag{9}$$

In other words, the estimate $\hat{\mathbf{x}} = \mathcal{W}^{-1}\hat{\boldsymbol{\theta}}$ does correspond to a Bayesian criterion in the signal domain, under the loss $L(\mathcal{W}\mathbf{x}, \mathcal{W}\tilde{\mathbf{x}})$, which is induced by the loss $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ that is adopted in the wavelet domain.

In some cases, this loss is invariant under orthogonal transformations, in the sense that

$$L(\mathcal{W}\mathbf{x}, \mathcal{W}\tilde{\mathbf{x}}) \propto L(\mathbf{x}, \tilde{\mathbf{x}}) \tag{10}$$

as a consequence, (9) can be further simplified to

$$\hat{\mathbf{x}} = \arg\min_{\tilde{\mathbf{x}}} \int L(\mathbf{x}, \tilde{\mathbf{x}}) p(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \tag{11}$$

meaning that $\hat{\mathbf{x}} = \mathcal{W}^{-1}\hat{\boldsymbol{\theta}}$ is a Bayes estimate under the same loss function as $\hat{\boldsymbol{\theta}}$.

It happens that the two most commonly used loss functions do verify (10):

- With the squared error loss, for which the optimal Bayes rule is the *posterior mean* [32] (PM), we can write

$$\begin{aligned} L_2(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) &= \left\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\right\|_2^2 = \|\mathcal{W}\mathbf{x} - \mathcal{W}\tilde{\mathbf{x}}\|_2^2 \\ &= \|\mathcal{W}(\mathbf{x} - \tilde{\mathbf{x}})\|_2^2 = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 = L_2(\mathbf{x}, \tilde{\mathbf{x}}) \end{aligned}$$

  (where $\|\cdot\|_2^2$ denotes squared Euclidean norm) as a trivial consequence of the orthogonality of $\mathcal{W}$; the DWT is an Euclidean norm preserving transformation (Parseval's relation). It can then be stated that the inverse DWT of the PM estimate of the coefficients coincides with the PM estimate of $\mathbf{x}$.
- For the 0/1 loss, which leads to the *maximum a posteriori* (MAP) criterion [32], $L_{0/1}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = L_{0/1}(\mathcal{W}\mathbf{x}, \mathcal{W}\tilde{\mathbf{x}}) = L_{0/1}(\mathbf{x}, \tilde{\mathbf{x}})$, simply because $\mathcal{W}^{-1}$ exists (i.e., $\mathcal{W}$ is bijective). In conclusion, the inverse DWT of the MAP estimate of the coefficients is the MAP estimate in the signal domain.

Notice that this is not true in general. It is easy to come up with loss functions that do not satisfy this condition; for example, $L_\infty(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \|\mathcal{W}\mathbf{x} - \mathcal{W}\tilde{\mathbf{x}}\|_\infty \neq L_\infty(\mathbf{x}, \tilde{\mathbf{x}})$ (where $\|\mathbf{v}\|_\infty$ stands for the *infinity norm*, $\|\mathbf{v}\|_\infty = \max\{|v_i|\}$). Of course, as seen above, the resulting rule is still a valid Bayes rule, but no simple and clear relation exists between the estimates in the signal and wavelet domains.

III. New Prior for Wavelet Coefficients

The decorrelation property supports that we model the coefficients as mutually independent

$$p(\boldsymbol{\theta}) = \prod_i p(\theta_i). \tag{12}$$

Of course, decorrelation does not imply independence, but this is a good first approximation often followed, and we adopt it here. Furthermore, recall that the likelihood function describes the observed coefficients as conditionally independent. As a consequence, the unknown coefficients are *a posteriori* conditionally independent,

$$p(\boldsymbol{\theta}|\boldsymbol{\omega}) \propto \prod_i p(\omega_i|\theta_i) \prod_j p(\theta_j) \propto \prod_i p(\theta_i|\omega_i)$$

where $p(\theta_i|\omega_i) \propto p(\omega_i|\theta_i)p(\omega_i)$, with $\omega_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$. Finally, under the MAP or the PM criterion (see above), the Bayes rule can be computed separately with respect to each coefficient,

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathrm{PM}} &= E[\boldsymbol{\theta}|\boldsymbol{\omega}] \\ &= [E[\theta_1|\omega_1], \ldots, E[\theta_N|\omega_N]]^T \\ \hat{\boldsymbol{\theta}}_{\mathrm{MAP}} &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\omega}) \\ &= \left[\arg\max_{\omega_1} p(\theta_1|\omega_1), \ldots, \arg\max_{\omega_N} p(\theta_N|\omega_N)\right]^T\end{aligned}$$

where N is the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$.

Let us then focus on choosing a prior for each wavelet coefficient, which we will now simply denote as θ. The usual approach is to explicitly capture the sparseness property with heavy-tailed priors. For example, Chipman *et al.* [7] and Crouse *et al.* [10] consider $p(\theta)$ as a mixture of two zero mean Gaussians: one with small variance and the other with large variance. Abramovich *et al.* [1] take this approach to an extreme by considering the small variance component as a point mass at zero. Student-t distributions were adopted by Vidakovic [35]. Other variants of these approaches are reviewed by Vidakovic [34].

Finally, recall that a Laplacian prior $p(\theta) \propto \exp\{-\nu|\theta|\}$ leads to the soft thresholding function [see equation (2)] as the MAP Bayes rule. A Bayesian interpretation of the hard threshold was presented by Moulin and Liu [26].

Here, we follow a different route based on the notion of "noninformativeness" or "invariance." The type of noninformativeness we are seeking must express *amplitude-scale* invariance; this means that the units in which a quantity is measured do not influence any conclusions drawn from it (see [32] or [3]). In other words, the inference procedure must be invariant under changes of amplitude-scale. For a positive parameter, say α, this kind of invariance is expressed by the well-known (noninformative) amplitude-scale-invariant Jeffreys' prior $p(\alpha) \propto 1/\alpha$ (see [32] or [3]). Now, our θ can be positive or negative; the corresponding amplitude-scale-invariant prior is then

$$p(\theta) \propto \frac{1}{|\theta|}. \tag{13}$$

This happens to be an extremely heavy-tailed "density," thus in accordance with the expected behavior of wavelet coefficients. In fact, it is so heavy-tailed that it is improper.[3] Notice that the simple invocation of amplitude-scale invariance leads to a heavy-tailed prior.

Let us clearly show how this noninformative prior exhibits amplitude-scale invariance. Say we change the measurement units (amplitude-scale) in which θ and all other quantities are expressed. This defines a new unknown $\beta = K\theta$, where K is the constant expressing the change of units/amplitude-scale. Then, by applying the rule for the change of variable in a pdf to $p(\theta) = |\theta|^{-1}$, we retain the same prior $p(\beta) \propto |\beta|^{-1}$. It is in this sense that the prior (13) is said to be amplitude-scale-invariant. Other priors for Bayesian denoising (based on Laplacian, Gaussian mixture, or other heavy-tailed densities) do not share this invariance property, and hence they must be tuned/adapted to the amplitude-scale of the wavelet coefficients of each particular signal/image at hand.

IV. Hierarchical/Empirical Bayes Approach

The prior $p(\theta) = |\theta|^{-1}$, together with the simple Gaussian observation model $\omega|\theta \sim \mathcal{N}(\theta, \sigma^2)$, leads to an improper (nonintegrable) *a posteriori* pdf $p(\theta|\omega)$. This *a posteriori* pdf also has a singularity at the origin, thus being unclear how to derive a simple inference rule from it. Consequently, we have to look for an alternative to a fully Bayesian approach. This alternative is provided by the identification of a hierarchical Bayesian model that is equivalent to the prior $p(\theta) = |\theta|^{-1}$; the goal is to facilitate the use of an empirical-Bayes-type approach. The equivalent hierarchical model is as follows.

- Each (unknown) coefficient is conditionally zero-mean Gaussian, with variance ϕ^2, $\theta|\phi^2 \sim \mathcal{N}(0, \phi^2)$, for $\phi^2 \geq 0$.
- Again, amplitude-scale invariance with respect to ϕ^2 is expressed by the noninformative improper Jeffreys' (hyper) prior $p(\phi^2) \propto 1/\phi^2$.

With these assumptions,

$$\int_0^\infty p\left(\theta|\phi^2\right) p\left(\phi^2\right) d\phi^2 = |\theta|^{-1} \tag{14}$$

showing that $p(\theta) = |\theta|^{-1}$ can be decomposed into a continuous mixture of zero-mean Gaussians, weighted according to the Jeffreys' noninformative hyper-prior $p(\phi^2) \propto 1/\phi^2$. Since this hyper-prior is the limiting case of the conjugate inverse-Gamma family, the prior $p(\theta) = |\theta|^{-1}$ is itself a limiting case of a family of Student-t densities [3]. Student-t densities are common robust substitutes for Gaussian priors [32], [3], which have been

[3]A prior is *improper* if it is not normalizable (its integral is not finite). Improper priors are common in Bayesian inference since only the relative weighting expressed by their shape impacts the *a posteriori* density [32], [3].

used in wavelet-based denoising with specially selected parameter settings (see Vidakovic [34]). Our (noninformative) prior leaves us with **no** free parameters to adjust.

This hierarchical Bayesian model opens the door to the use of an empirical-Bayes-type technique [32]; i.e., we break the fully Bayesian analysis chain as follows:

- First, a variance estimate $\widehat{\phi^2}$ is obtained with the MAP criterion based on the marginal likelihood $p(\omega|\phi^2)$ and the corresponding (amplitude-scale-invariant) Jeffreys' prior.
- Given $\widehat{\phi^2}$, both the MAP and the posterior mean criteria lead to the well known shrinkage estimator, resulting from a Gaussian likelihood (of variance σ^2) and a $\mathcal{N}(0, \widehat{\phi^2})$ prior,

$$\hat{\theta} = \frac{\widehat{\phi^2}}{\widehat{\phi^2} + \sigma^2}\,\omega. \tag{15}$$

Notice that this is a nonlinear estimator because, although not clearly expressed by the notation, $\widehat{\phi^2}$ depends on ω.

The MAP estimate of the variance, $\widehat{\phi^2}$, is derived as follows. Since $\omega = \theta + n'$, the marginal likelihood is very simply $\omega|\phi \sim \mathcal{N}(0, \phi^2 + \sigma^2)$, and the corresponding Jeffreys' prior is now $p(\phi^2) \propto 1/(\phi^2 + \sigma^2)$. Notice that this Jeffreys' prior respects our amplitude-scale-invariance desideratum. To see this, consider again the change the measurement units expressed by defining a new variable $\xi^2 = K\phi^2$. Applying the rule for a change of variable to the prior $p(\phi^2)$, we obtain $p(\xi^2) \propto 1/(\xi^2 + K\sigma^2)$, which is the same prior, with the noise variance adequately re-scaled. The resulting MAP estimate of ϕ^2 is

$$\begin{aligned}\widehat{\phi^2} &= \arg\max_{\phi^2 \geq 0} \left\{ \left(\phi^2 + \sigma^2\right)^{-3/2} e^{-(\omega^2/2(\phi^2+\sigma^2))} \right\} \\ &= \left(\frac{\omega^2}{3} - \sigma^2\right)_+ \end{aligned} \tag{16}$$

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

Let us also point out another interpretation of the Bayesian (variance) estimator in (16). Ignoring the $(\cdot)_+$ function (which is necessary simply because we are estimating ϕ^2 from an estimate of $\phi^2 + \sigma^2$, and the valid parameter space is $\mathbb{R}_0^+$), this is an instance of the following problem: given n i.i.d. $\mathcal{N}(0, \gamma^2)$ observations, $x_1, \ldots, x_n$, what is the best estimate of the variance, with the form $\widehat{\gamma^2} = c(x_1^2 + \cdots + x_n^2)$, in a mean squared error (MSE) $E[(\gamma^2 - \widehat{\gamma^2})^2]$ sense? It is known that the value $c = 1/(n+2)$ (in our case, $n = 1$, thus $c = 1/3$) yields the minimum MSE (although biased) estimate of γ^2 (see [23]). This coincides with the MAP rule with a Jeffreys' prior on γ^2.

Now, by plugging the estimate (16) into (15), we have our new nonlinear rule, which we call the *amplitude-scale-invariant Bayes estimator* (ABE)

$$\hat{\theta} = \delta^{\mathrm{ABE}}(\omega) = \frac{\left(\omega^2 - 3\sigma^2\right)_+}{\omega} \tag{17}$$

which is plotted in Fig. 1, together with the classical soft and hard thresholding functions (for the same threshold value). Notice how the proposed rule places itself between those two functions: it is close to the soft rule for small ω, thus effectively behaving like a shrinkage rule; it approaches the hard rule (and consequently the identity line) for large ω, avoiding the undesirable bias incurred with the soft rule.

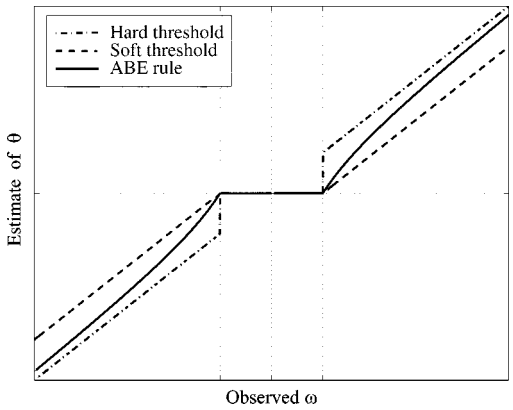

Fig. 1. ABE nonlinearity versus the hard and soft thresholding rules (for the same threshold).

Computationally, our denoising method is as simple as any other one that uses some fixed thresholding/shrinkage nonlinearity depending on a fixed threshold proportional to the noise standard deviation σ (e.g., VisuShrink); that is, the only needed input is σ. Remarkably, however, it achieves the performance of (more computationally demanding) Bayesian methods (see the experimental results in Section VI) without requiring any tuning or adaptive estimation of parameters of the prior.

V. Relation with the Non-Negative Garrote

As mentioned in the Introduction, Gao [18] has very recently proposed the use of the so-called "nonnegative garrote" function, defined in (3), for wavelet-based denoising. Notice that the ABE rule [equation (17)] happens to be a "nonnegative garrote" with a fixed threshold $\lambda = \sqrt{3\sigma^2}$:

$$\delta^{\mathrm{ABE}}(\omega) = \delta^{\mathrm{garrote}}_{\sqrt{3\sigma^2}}(\omega).$$

In [18], this function is shown to outperform both the hard and soft nonlinearities when the threshold is optimally selected with the help of the underlying true function. Gao credits the nonnegative garrote to Breiman [4] who introduced it in the context of subset selection for regression problems. Brillinger [5] has also briefly mentioned a similar function [in fact $\delta^{\mathrm{garrote}}_{\sigma}(\omega)$] as a possible alternative to the hard and soft rules; according to him, this nonlinear function had been proposed by Tukey (in unpublished work of 1979), also in a regression context.

Finally, the nonnegative garrote [specifically, $\delta^{\mathrm{garrote}}_{\sigma}(\omega)$] also arises naturally in certain cross-validation methods, as used by Nowak [28] and Nowak and Baraniuk [29].

VI. Experimental Results

A. Signal Denoising

We have evaluated the ABE rule versus the standard SureShrink and VisuShrink methods (based on soft thresholding), using Donoho and Johnstone's [12] well known test signals: "Blocks," "Doppler," "HeaviSine," and "Bumps." We have also included in our comparison a Bayesian approach based on mixture priors (see [10]) which, to our knowledge, is

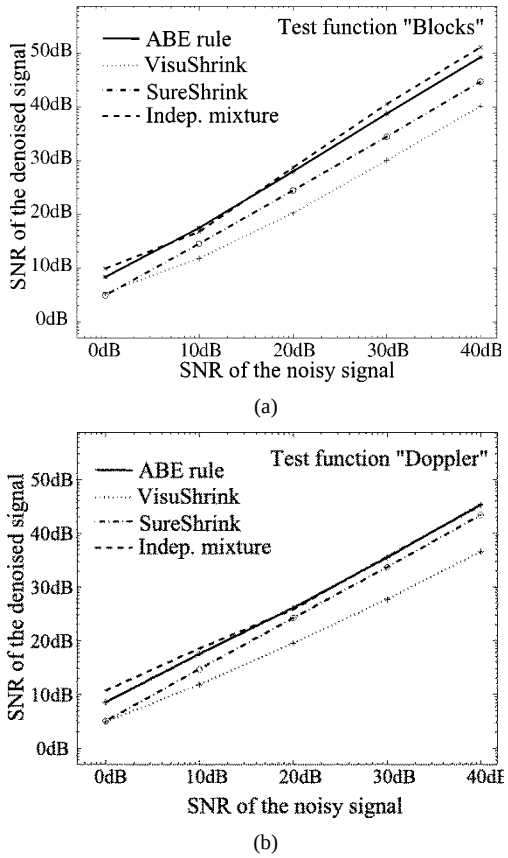

Fig. 2. Input and output SNR for various wavelet denoising schemes applied to two standard test signals: "blocks" and "Doppler" (wavelets: Daubechies-2 for Blocks, Daubechies-8 for Doppler).

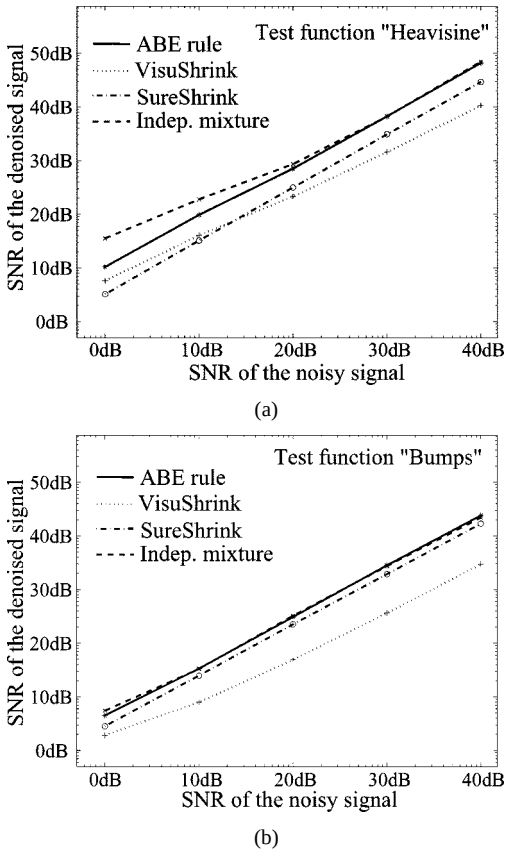

Fig. 3. Input and output SNR for various wavelet denoising schemes applied to two standard test signals: "heavisine" and "bumps" (wavelets: Daubechies-8 for HeaviSine, and Daubechies-6 for Bumps).

representative of the very best Bayesian methods. Figs. 2 and 3 report the signal-to-noise ratios (SNR) obtained by each of the methods, based on 100 runs for each original SNR value. These results show that the ABE rule consistently (for all four test signals and at all SNR levels) outperforms SureShrink; this is a remarkable fact because SureShrink is an adaptive method (more computationally demanding) while the ABE rule is fixed. With respect the VisuShrink, which has a similar computational load, ABE achieves far superior results. Finally, as is also clear in Figs. 2 and 3, the proposed technique performs comparably (except for HeaviSine at low SNR) with the much more computationally demanding mixture based method.

Our experimental results allow adding the following conclusions to those of Gao [18]: at least for the signals and SNR values considered, a nonnegative garrote with a fixed threshold $\lambda = \sqrt{3\sigma^2}$ still beats SureShrink (and also, of course, VisuShrink). This conclusion implies an important practical guideline: the ABE method should be used instead of SureShrink. Our method performs better than SureShrink, and it is much more computationally efficient.

B. Image Denoising

For image denoising, we have compared the ABE rule versus the hard, the soft, and the garrote nonlinearities for a range of threshold values. Fig. 4 shows a mosaic containing the well-known "Lena" and "Cameraman" images (and two other images) after being contaminated by noise of standard deviation $\sigma = 20$. Fig. 6 shows the mean squared error achieved by the hard, soft, and garrote rules, as a function of the respective threshold values; the horizontal dotted line represents the mean squared error of the proposed (fixed threshold) ABE. Notice how the ABE rule achieves lower MSE than both the hard and soft functions, even when these are allowed to choose ideal (clairvoyant) thresholds using the underlying true images (of course, something that in practical situations can not be done). Concerning the garrote, it is remarkable that the optimal threshold is found to be $\lambda = 35.1$ which is very close to our fixed threshold $\sqrt{3\sigma^2} \simeq 34.6$. The resulting denoised images are shown in Figs. 4 and 5.

The same test was performed for two other values of σ (10 and 40), and the results are also reported in Fig. 6. Again, our $\delta^{\mathrm{ABE}}(\omega)$ outperforms both the hard and soft rules, even with their ideal/clairvoyant thresholds. The garrote rule [of which, recall, $\delta^{\mathrm{ABE}}(\omega)$ is a particular case] is able to find thresholds with which it very slightly beats the ABE rule (however, recall that these are clairvoyant thresholds that can not be found in practical situations because they would require access to the unknown underlying images). Again, the optimal garrote rule thresholds are very near our fixed level of $\sqrt{3}\sigma$.

In Mihçak *et al.* [25], an image denoising method was proposed which has some similarities with ours. In that paper, wavelet coefficients are also modeled as zero-mean Gaussian with unknown variance. The denoising procedure obtains variance estimates and plugs these estimates in the standard linear

(a)

(b)

Fig. 4. Noisy image ($\sigma = 20$) and denoised image produced by the ABE rule (MSE = 125.2) (wavelet: Daubechies-2).

(a)

(b)

Fig. 5. Denoised images produced by the hard (MSE = 155.6) and soft (MSE = 131.8) rules, with ideal/clairvoyant thresholds, from the noisy image in Fig. 4 (wavelet: Daubechies-2).

shrinkage rule [equation (15)]. The main difference between our method and the one in [25] is that our variance estimates are obtained independently for each coefficient (under a Jeffreys' prior), thus leading to a coefficient-wise closed-form estimation rule, while in [25] the variance estimates are obtained from small windows around each coefficient. So, in a sense, our method can be seen as a limit case of the one in [25] with windows of size 1, thus being even less computationally demanding. Two variants of the method are studied in [25]: variance estimates are obtained by the ML criterion, leading to a scheme called *locally adaptive window-based denoising using ML* (LAWML); and using MAP variance estimates obtained under an exponential prior (LAWMAP). Table I shows the results of our (ABE) rule, in comparison with those reported in [25] (for the Lena and Barbara images). As expected, due to its less adaptable nature, and less robust variance estimate (based on only one coefficient), ABE performs a little worse than LAWML and LAWMAP (with 3×3 windows), but clearly better than the hard threshold.

VII. Bayesian Interpretation of the Garrote

The denoising approach proposed in this paper provides an empirical Bayes interpretation of the nonnegative garrote estimator. This fact suggests the question: is there any prior $p(\theta)$ for the wavelet coefficients that leads to the nonnegative garrote as the true MAP Bayesian estimator (rather than an empirical Bayes one)? The answer to this question would help shed some "Bayesian light" on the good performance of the nonnegative garrote shrinkage rule. It turns out that the answer is positive: with $\omega|\theta \sim \mathcal{N}(\theta, \sigma^2)$, and if we assume the prior

$$p(\theta) \propto \exp\left\{ \frac{\theta^2}{4\sigma^2} - \frac{|\theta|\sqrt{4a\sigma^2+\theta^2}}{4\sigma^2} - a\log\left(|\theta| + \sqrt{4a\sigma^2+\theta^2}\right) \right\} \quad (18)$$

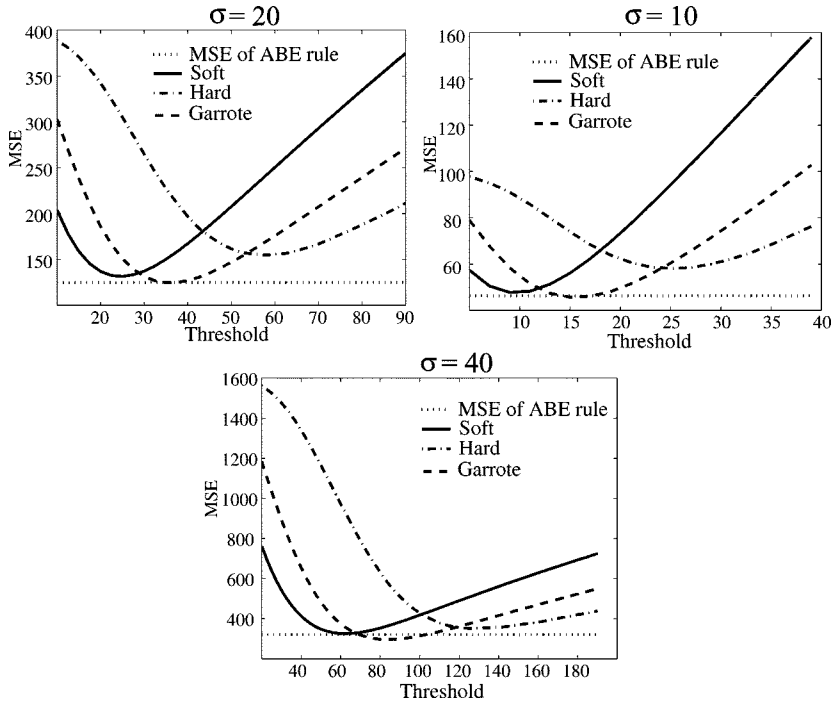

Fig. 6. MSE achieved by the hard, soft, and garrote rules, as function of threshold value, for three noise standard deviations: $\sigma = 20$, $\sigma = 10$, and $\sigma = 40$ (image of Figs. 4 and 5). The horizontal dotted line shows the MSE obtained by the ABE rule (with its fixed threshold).

TABLE I
PSNR (dB) RESULTS FOR THE ABE RULE, LAWML, LAWMAP, AND HARD THRESHOLDING

Lena				
	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 25$
hard threshold	30.34	28.52	27.24	26.34
ABE rule	32.74	30.48	28.74	27.38
LAWML	33.72	31.37	29.63	28.22
LAWMAP	34.25	32.33	31.00	29.96
Barbara				
	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 25$
hard threshold	27.29	25.01	23.65	22.83
ABE rule	31.28	28.79	27.03	25.71
LAWML	32.32	29.72	27.93	26.53
LAWMAP	32.46	30.03	28.39	27.21

(which is not improper, although the normalization constant can not be obtained in closed form, only numerically), the resulting MAP estimator of θ, $\hat{\theta}_{\text{MAP}} = \arg\max_\theta p(\theta)p(\omega|\theta)$, is given by $\hat{\theta}_{\text{MAP}} = \delta^{\text{garrote}}_{\sqrt{a\sigma^2}}(\omega)$ (with $a = 3$ corresponding to the estimator proposed in this paper). This can be confirmed by using the formulation in [20] to "reverse engineer" the nonnegative garrote shrinkage/thresholding function. It is somewhat odd to be led to a prior that depends on the noise variance, which is a

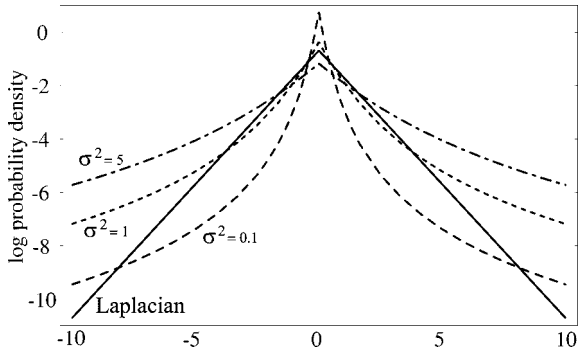

Fig. 7. Probability densities given by (18), all for $a = 3$ and for different values of σ^2, and Laplacian density.

parameter of the likelihood function (observation model). Notice, however, that the use of priors that depend on the likelihood function is not uncommon, with the Jeffreys prior being an obvious example.

In Fig. 7, we plot $p(\theta)$ as given in (18) for $a = 3$ and for three different values of σ^2 (0.1, 1, and 5); for comparison, the Laplacian density (which leads to a MAP estimate given by the soft threshold rule) is also plotted. The prior given by (18) always exhibits heavier tails than the Lapacian, as is clear in the plots.

A final question is: are the members in the family of densities defined by (18) good models for wavelet coefficient statistics of natural images? To answer this question we plot in Fig. 8 the (normalized) histogram of the wavelet coefficients of the original image used in the example of Fig. 4 together with the density of the form (18) fitted to this histogram via the maximum likelihood (ML) criterion (parameters: $a = 2.2$, $\sigma^2 = 55$). For comparison, we also plot the best (also in the ML sense) *gener-*

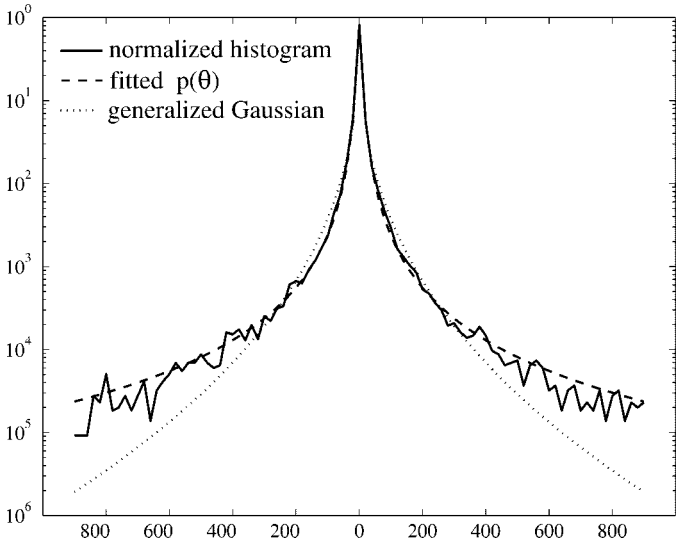

Fig. 8. Normalized histogram (solid line) of the wavelet coefficients of the original image used in the example of Fig. 4 and the density from the family (18) fitted to it (dashed line). For comparison, the generalized Gaussian density fitted to this data via the ML criterion is also plotted (dotted line).

alized Gaussian (GG) density, which is a commonly used model for heavy-tailed distributions [26]. The GG density is defined as

$$p_{\mathrm{GG}}(\theta) \propto \exp\left\{-\left(\frac{A}{\tau}\right)^{\xi}|\theta|^{\xi}\right\}$$

where

- ξ shape parameter;
- τ standard deviation;
- $A = \sqrt{\Gamma(3/\xi)/\Gamma(1/\xi)}$ [where $\Gamma(\cdot)$ denotes Euler's Gamma function].

The parameters resulting from the ML fit are $\xi = 0.42$ and $\tau = 56$. The plot reveals that the best density of the form (18) fits the histogram slightly better than the best GG density. This is confirmed by the fact that the normalized maximum likelihood for the density (18) is -0.938, versus -0.948 for the GG density.

We repeated the experiment with other images obtaining similar results. The very close agreement between empirical histograms and the functional of the prior density underlying the nonnegative garrote rule helps to explain its good performance. It is also foreseeable that this density can be used to obtain new wavelet-based coders/quantizers for image compression.

VIII. Conclusions and Future Work

We have proposed an empirical-Bayes approach to wavelet-based image and signal estimation, where a noninformative (amplitude-scale invariant) prior plays a central role. A hierarchical/empirical Bayes path lead us to a simple fixed nonlinear shrinkage/thresholding rule; unlike other schemes, it has no free parameters requiring tuning or estimation. Tests based on Donoho and Johnstone's standard test signals showed that our rule outperforms both VisuShrink and SureShrink. Moreover, it performs comparably with a recent approach, based on independent mixture priors [10], which is representative of the very best Bayesian methods.

Concerning image estimation, we showed that the ABE rule achieves lower MSE than both the hard and the soft nonlinearities, even when these are allowed to find their ideal/clairvoyant thresholds using the true original image. The excellent estimation performance of the noninformative approach here described seems to support the presence a relevant characteristic for natural image modeling: amplitude-scale invariance. This feature of natural images means that they contain information at all amplitude-scales; any model that fails to take this into account will have to pay the price of adapting to the dominant amplitude-scale features of the particular image in hand, at the expense of features at other amplitude scales.

We are currently investigating the use of our rule in conjunction with translation-invariant (TI) denoising schemes [8]; actually, TI denoising can also be formalized through the use of noninformative priors [16], [17]. TI schemes mitigate undesirable (*pseudo-Gibbs* or *blocking*) artifacts and improve the performance of all methods considered above, with the ranking of their relative performances being approximately unchanged.

References

[1] F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet thresholding via a Bayesian approach," *J. R. Statist. Soc. B*, vol. 60, 1998.
[2] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1004–1034, 1995.
[3] J. Bernardo and A. Smith, *Bayesian Theory*. Chichester, U.K.: Wiley, 1994.
[4] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, pp. 373–384, 1995.
[5] D. Brillinger, "Uses of cumulants in wavelet analysis," *Nonparam. Statist.*, vol. 6, pp. 93–114, 1996.
[6] J. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, pp. 2009–2025, 1998.
[7] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 92, pp. 1413–1421, 1997.
[8] R. Coifman and D. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics*. New York: Springer-Verlag, 1995, pp. 125–150.
[9] P. Comon, "Independent component analysis: A new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.
[10] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, 1998.
[11] R. DeVore, "Nonlinear Approximation," *Acta Numer.*, pp. 51–150, 1998.
[12] D. Donoho and I. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[13] ——, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, 1995.
[14] D. Field, "Scale-invariance and self-similar wavelet transforms: An analysis of natural scenes and mammalian visual systems," in *Wavelets, Fractals, and Fourier Transforms*, M. Farge, J. Hunt, and C. Vascillios, Eds. Oxford, U.K.: Oxford Univ. Press, 1993, pp. 151–193.
[15] M. Figueiredo and J. Leitão, "Unsupervised image restoration and edge location using compound Gauss–Markov random fields and the MDL principle," *IEEE Trans. Image Processing*, vol. 6, pp. 1089–1102, 1997.
[16] M. Figueiredo and R. Nowak, "Bayesian wavelet-based signal estimation using noninformative priors," in *Proc. 32nd Asilomar Conf. Signals, Systems, Computers*, Monterey, CA, 1998, pp. 1368–1373.
[17] ——, "Bayesian wavelet-based image estimation using noninformative priors," *Proc. SPIE*, vol. 3816, pp. 97–108, 1999.
[18] H. Gao, "Wavelet shrinkage denoising using the nonnegative garrote," *J. Comput. Graph. Statist.*, vol. 7, pp. 469–488, 1998.

[19] H. Gao and A. Bruce, "WaveShrink with firm shrinkage," *Statist. Sinica*, vol. 7, pp. 855–874, 1997.
[20] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Comput.*, vol. 11, pp. 1739–1768, 1999.
[21] H. Krim and I. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Trans. Inform. Theory*, vol. 45, pp. 898–908, 1999.
[22] T. Lee, *Independent Component Analysis*. Dordrecht, The Netherlands: Kluwer, 1998.
[23] E. Lehmann, *Theory of Point Estimation*. Pacific Grove, CA: Wadsworth, 1983.
[24] S. Mallat, *A Wavelet Tour of Signal Processing*. New York: Academic, 1998.
[25] M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Processing Lett.*, vol. 6, pp. 300–303, 1999.
[26] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalizad Gaussian and complexity priors," *IEEE Trans. Inform. Theory*, vol. 45, pp. 909–919, 1999.
[27] P. Müller and B. Vidakovic, Eds., *Bayesian Inference in Wavelet-Based Models*. New York: Springer-Verlag, 1999.
[28] R. Nowak, "Optimal signal estimation using cross-validation," *IEEE Signal Processing Lett.*, vol. 4, pp. 23–25, 1997.
[29] R. Nowak and R. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *Proc. SPIE*, vol. 3169, pp. 55–66, 1997.
[30] R. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*. Boston, MA: Birkhäuser, 1997.
[31] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[32] C. Robert, *The Bayesian Choice: A Decision Theoretic Motivation*. New York: Springer-Verlag, 1994.
[33] D. Ruderman, "The statistics of natural images," *Network: Comput. Neural Syst.*, vol. 5, pp. 517–548, 1995.
[34] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *J. Amer. Statist. Assoc.*, vol. 93, pp. 173–179, 1998.
[35] ——, "Wavelet-based nonparametric Bayes methods," in *Practical Nonparametric and Semiparametric Bayesian Statistics*. Berlin, Germany: Springer-Verlag, 1998, pp. 133–155.

Mário A. T. Figueiredo (S'87–M'95–SM'00) received the E.E., M.Sc., and Ph.D. degrees in electrical and computer engineering, all from the Instituto Superior Técnico, Technical University of Lisbon (IST), Lisbon, Portugal, in 1985, 1990, and 1994, respectively.

Since 1994, he has been 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, Lisbon. In 1998, he held a visiting position with the Department of Computer Science and Engineering, Michigan State University, East Lansing. His scientific interests include image processing and analysis, computer vision, statistical pattern recognition, statistical learning, and information theory. He is on the editorial board of *Pattern Recognition Letters*.

Dr. Figueiredo received the Portuguese IBM Scientific Prize in 1995. He is co-chair of the 2001 International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition.

Robert D. Nowak (S'88–M'89) received the B.S. (with highest distinction), M.S., and Ph.D. degrees in electrical engineering, all from the University of Wisconsin-Madison in 1990, 1992, and 1995, respectively.

He spent several summers with General Electric Medical Systems' Applied Science Laboratory, where he received the General Electric Genius of Invention Award and a U.S. patent for his work in 3-D computed tomography. He was an Assistant Professor at Michigan State University from 1996 to 1999. He is now an Assistant Professor with the Department of Electrical and Computer Engineering at Rice University, Houston, TX. His research interests include statistical image and signal processing, multiscale analysis, medical imaging, and communication networks.

Dr. Nowak received the National Science Foundation CAREER Award in 1997, the Army Research Office Young Investigator Program Award in 1999, the Office of Naval Research Young Investigator Program Award in 2000, and IEEE Signal Processing Society Young Author Best Paper Award in 2000.