Bayesian wavelet-based image estimation using noninformative priors

Mário A. T. Figueiredo
Instituto de Telecomunicações, and
Instituto Superior Técnico
1049-001 Lisboa,
Portugal

Robert D. Nowak
Department of Electrical and Computer Engineering
Michigan State University
East Lansing, MI 48824,
USA

ABSTRACT

The sparseness and decorrelation properties of the discrete wavelet transform have been exploited to develop powerful denoising methods. Most schemes use arbitrary thresholding nonlinearities with *ad hoc* parameters, or employ computationally expensive adaptive procedures. We overcome these deficiencies with a new wavelet-based denoising technique derived from a simple empirical Bayes approach based on Jeffreys' non-informative priors. Our approach is a step towards *objective* Bayesian wavelet-based denoising. The result is a remarkably simple fixed non-linear shrinkage/thresholding rule which performs better than other more computationally demanding methods.

Keywords: image denoising, image estimation, wavelets, Bayesian estimation, non-informative priors, Jeffreys' priors, invariance, hierarchical Bayes, empirical Bayes, shrinkage.

1. INTRODUCTION

1.1. Background

Wavelets and other multiscale analysis tools underlie many recent advances in key areas of signal and image processing; namely, approximation (or representation), estimation, and compression (for example, see Mallat's[1] 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: **(a)** it is *sparse*, meaning that a few large coefficients dominate the representation, and **(b)** 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 signal/image processing tasks (see Mallat[1]) and also for statistical applications (see Ogden[2] 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 of the signal,
2. perform some specified processing on the DWT coefficients, and
3. compute the inverse DWT to obtain the "processed" signal.

M. Figueiredo was partially supported the (Portuguese) Science and Technology Foundation, grants PRAXIS XXI BPD-14129-97 and TIT-1580; R. Nowak was partially supported by the (US) National Science Foundation, grant no. MIP–9701692.

E-mail addresses: `mtf@lx.it.pt` and `nowak@egr.msu.edu`

Stimulated by the seminal work of Donoho and Johnstone,[3] many denoising (or signal/image estimation) methods adopting this standard three step approach have been proposed (see Mallat,[1] Ogden,[2] Vidakovic,[4] Krim and Schick,[5] and other references cited by those authors). In particular for detail-preserving (or edge-preserving) image estimation/denoising (the subject of this paper) these approaches provide a very efficient alternative to Markov random field (MRF) based techniques (see Figueiredo and Leitão[7] 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[3,8]) 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

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

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

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*[8]); 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 (see Vidakovic[4] and references therein). These priors, together with the likelihood function (noise model), produce posterior distributions. Estimation rules can then be derived via the standard Bayesian decision-theoretic approach, after the specification of a loss function (see Robert[9]). Bayesian techniques usually outperform other methods and are representative of the state-of-the-art in wavelet-based denoising (see Vidakovic[4] and Crouse, Nowak, and Baraniuk[10]).

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 often arbitrary. 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[11] "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, very recently, the "non-negative garrote" function (as suggested by Gao[12]), defined as

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

which we will return to in Section 5. Concerning threshold levels, in practice they often have to be "tweaked" to produce best results on a case by case basis. Even signal-adaptive techniques can still be criticized since they involve an arbitrary choice of nonlinearity and because they are computationally demanding.

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

1.2. Contributions

We tackle the fundamental issues raised above by adopting a Bayesian perspective supported on non-informative Jeffreys' priors (see Bernardo and Smith's[13] and Robert's[9] books).

Our approach can be seen as a step towards *objective* Bayesian wavelet-based denoising; the term "objective" means the use of priors that do not require any subjective input. Accordingly, our approach mitigates the arbitrariness/subjectiveness associated with other (Bayesian and non-Bayesian) denoising schemes. The type of non-informativeness we invoke expresses *amplitude-scale** invariance, meaning that the units in which an image/signal is measured do not directly influence any inference made from it (see Bernardo and Smith[13] or Robert[9]). 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 modelling. The good results achieved with our non-informative 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 a great deal of attention (see Field[14] and Ruderman[15])

In Section 2, the denoising problem is described and notation introduced. In Section 3, a new non-informative prior is proposed, based on which we derive, in Section 4, a novel empirical-Bayes denoising procedure that we call *amplitude-scale-invariant Bayes estimation* (ABE). In Section 5 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 6. Conclusions are given in Section 7.

2. PROBLEM FORMULATION

2.1. 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† 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 $\mathbf{y}$.

In wavelet-based denoising, the orthogonal DWT, denoted by $\mathcal{W}$ (either 1-D or 2-D; see, *e.g.*, Mallat's book[1]) is applied to the noisy data yielding the noisy *wavelet coefficients* $\boldsymbol{\omega}$; these are described by an analogous observation 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 (see Donoho and Johnstone[3]). 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, Olshausen and Field,[16] Hyvärinen, Hoyer and Oja,[17] and references therein). On the other hand, the DWT of Gaussian 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

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

†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.[8]

the rationale underlying the now classical thresholding methods introduced by Donoho and Johnstone[3] and all their descendants.

Finally we mention that there is also a conceptual link between wavelet-based denoising and *independent component analysis* (ICA); see Cardoso,[18] Comon,[19] Bell and Sejnowski,[20] and the recent book by Lee.[21] 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).

2.2. 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 $\sigma^2\mathbf{I}$:

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

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

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 an optimal Bayes estimate $\widehat{\boldsymbol{\theta}}$, given $\boldsymbol{\omega}$;
3. reconstruct the signal estimate $\widehat{\mathbf{x}} = \mathcal{W}^{-1}\widehat{\boldsymbol{\theta}}$.

To interpret this procedure from a Bayesian decision theory perspective, let us explicitly write down $\widehat{\boldsymbol{\theta}}$ as the minimizer of the *a posteriori* expected loss (see Bernardo and Smith[13] or Robert[9]); then

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

In (8), $L(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}})$ is the adopted loss function that penalizes the "discrepancy" between $\boldsymbol{\theta}$ and any candidate estimate $\widetilde{\boldsymbol{\theta}}$, 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

$$p(\mathbf{x}|\mathbf{y}) \;\propto\; p(\mathbf{y}|\mathbf{x})\, p_X(\mathbf{x}) \;=\; p(\boldsymbol{\omega}|\theta)\, p_X(\mathcal{W}^{-1}\boldsymbol{\theta}) \;=\; p(\boldsymbol{\omega}|\theta)\, p_\Theta(\boldsymbol{\theta}) \;\propto\; p(\boldsymbol{\theta}|\boldsymbol{\omega}), \tag{9}$$

equation (8) is equivalent to

$$\widehat{\mathbf{x}} = \arg\min_{\widetilde{\mathbf{x}}} \int L(\mathcal{W}\mathbf{x}, \mathcal{W}\widetilde{\mathbf{x}}) p(\mathbf{x}|\mathbf{y})\, d\,\mathbf{x}. \tag{10}$$

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

In some cases, this loss is invariant under orthogonal transformations (maybe up to a constant), that is

$$L(\mathcal{W}\mathbf{x}, \mathcal{W}\widetilde{\mathbf{x}}) \propto L(\mathbf{x}, \widetilde{\mathbf{x}}); \tag{11}$$

as a consequence, (10) can be further simplified to

$$\widehat{\mathbf{x}} = \arg\min_{\widetilde{\mathbf{x}}} \int L(\mathbf{x}, \widetilde{\mathbf{x}}) p(\mathbf{x}|\mathbf{y})\, d\,\mathbf{x}, \tag{12}$$

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

It happens that the two most commonly used loss functions do verify (11):

- With the squared error loss, for which the optimal Bayes rule is the posterior mean[9] (PM), we can write $L_2(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2^2 = \|\mathcal{W}\mathbf{x} - \mathcal{W}\widetilde{\mathbf{x}}\|_2^2 = \|\mathcal{W}(\mathbf{x} - \widetilde{\mathbf{x}})\|_2^2 = \|\mathbf{x} - \widetilde{\mathbf{x}}\|_2^2 = L_2(\mathbf{x}, \widetilde{\mathbf{x}})$ (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,[9] $L_{0/1}(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = L_{0/1}(\mathcal{W}\mathbf{x}, \mathcal{W}\widetilde{\mathbf{x}}) = L_{0/1}(\mathbf{x}, \widetilde{\mathbf{x}})$, 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, $L_\infty(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) = \|\mathcal{W}\mathbf{x} - \mathcal{W}\widetilde{\mathbf{x}}\|_\infty \neq L_\infty(\mathbf{x}, \widetilde{\mathbf{x}})$ (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.

3. A NEW PRIOR FOR WAVELET COEFFICIENTS

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

$$p(\boldsymbol{\theta}) = \prod_i p(\theta_i); \tag{13}$$

of course, decorrelation does not imply independence, but this is a good a 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,

$$p(\boldsymbol{\theta}|\boldsymbol{\omega}) \propto p(\boldsymbol{\omega}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \prod_i p(\omega_i|\theta_i) \prod_j p(\theta_j) \propto \prod_i p(\theta_i|\omega_i) \tag{14}$$

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, if either the MAP or the PM criterion is adopted (see above), the Bayes rule can be computed separately with respect to each coefficient:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{PM}} = E[\boldsymbol{\theta}|\boldsymbol{\omega}] = [E[\theta_1|\omega_1], ..., E[\theta_N|\omega_N]]^T \tag{15}$$

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{\omega}) = \left[\arg\max_{\omega_1} p(\theta_1|\omega_1), ..., \arg\max_{\omega_N} p(\theta_N|\omega_N)\right]^T \tag{16}$$

where N is the dimension of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$. In conclusion, under white noise, with an independent prior as (13), the MAP or PM estimates can be obtained separately for each coefficient.

Let us then focus on choosing a prior for each single wavelet coefficient, which we will now simply denote as θ. The usual approach is to try to explicitly capture the sparseness property with heavy-tailed priors. For example, Chipman, Kolaczyk, and McCulloch[22] and Crouse, Nowak, and Baraniuk[10] consider $p(\theta)$ as a mixture of two zero mean Gaussians: one with small variance and the other with large variance. Abramovich, Sapatinas and Silverman[23] 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.[24] Other variants of these approaches are reviewed by Vidakovic.[4] Finally, it is well known that a Laplacian prior $p(\theta) \propto \exp\{-\nu|\theta|\}$, which is also heavy-tailed, coupled with 0/1 loss, leads to the soft thresholding function (see equation (2)) as the optimal Bayes rule.

Here, we follow a different route based on the notion of "non-informativeness" or "invariance". The type of non-informativeness 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 Bernardo and Smith[13] or Robert[9]). 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 (non-informative) amplitude-scale-invariant Jeffreys' prior $p(\alpha) \propto 1/\alpha$ (again, see Bernardo and Smith[13] or Robert[9]). Now, our θ can be positive or negative; the corresponding amplitude-scale-invariant prior is then

$$p(\theta) \propto \frac{1}{|\theta|}. \tag{17}$$

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[‡]. Notice that the simple invocation of amplitude-scale invariance leads to a heavy-tailed prior.

Let us clearly show how this non-informative 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 (17) 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 desirable invariance property, and hence they must be tuned/adapted to the amplitude-scale of each particular signal/image at hand.

4. A HIERARCHICAL/EMPIRICAL BAYES APPROACH

Unfortunately, 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 (non-integrable) *a posteriori* pdf $p(\theta|\omega)$, from which no simple inference rule can be derived. 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 our prior $p(\theta) = |\theta|^{-1}$; the goal is to facilitate the use of an empirical-Bayes-type approach. The equivalent hierarchical Bayesian model is:

- 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 non-informative improper Jeffreys' (hyper) prior $p(\phi^2) \propto 1/\phi^2$.

The marginal *a posteriori* density $p(\theta|\omega)$ resulting from this hierarchical model,

$$p(\theta|\omega) = \int p(\theta, \phi^2|\omega) d\phi^2 = \frac{p(\omega|\theta)}{p(\omega)} \underbrace{\int p(\theta|\phi^2) p(\phi^2) d\phi^2}_{p(\theta) = |\theta|^{-1}}, \tag{18}$$

does reveal the presence of the prior $p(\theta) = |\theta|^{-1}$. This shows that this prior can be decomposed into a continuous mixture of zero-mean Gaussians, weighted according to the Jeffreys' non-informative hyper-prior $p(\phi^2) \propto 1/\phi^2$ (see Robert[9]). 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. Student-t densities are common robust substitutes for Gaussian priors (see Bernardo and Smith[13] or Robert[9]), and have been used in wavelet-based denoising with specially selected parameter settings (see Vidakovic[4]). Our (non-informative) 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 (see Robert[25]); *i.e.*, we break the fully Bayesian analysis chain as follows:

- First, a variance estimate $\widehat{\phi^2}$ is obtained according to 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,

$$\widehat{\theta} = \frac{\widehat{\phi^2}}{\widehat{\phi^2} + \sigma^2} \, \omega. \tag{19}$$

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

[‡]A prior is said *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; see Bernardo and Smith[13] or Robert.[9]

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 automatically re-scaled in accordance with the new units. The resulting MAP estimate of ϕ^2 is

$$\widehat{\phi^2} = \arg\max_{\phi^2 \geq 0} \left\{ (\phi^2 + \sigma^2)^{-3/2}\, e^{-\frac{\omega^2}{2(\phi^2+\sigma^2)}} \right\} \tag{20}$$

$$= \left(\frac{\omega^2}{3} - \sigma^2 \right)_+, \tag{21}$$

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 (21). 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, $x_1, ..., x_n$, what is the best variance estimate of 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 well 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 Lehmann[26]). This coincides with the MAP rule with a Jeffreys' prior on γ^2.

Now, by plugging the estimate (21) into (19), we have our new non-linear rule, which we call the *amplitude-scale-invariant Bayes estimator* (ABE),

$$\widehat{\theta} = \delta^{\mathrm{ABE}}(\omega) = \frac{\left(\omega^2 - 3\sigma^2\right)_+}{\omega}, \tag{22}$$

which is plotted in Figure 1. Also in Figure 1, the ABE rule is shown 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.

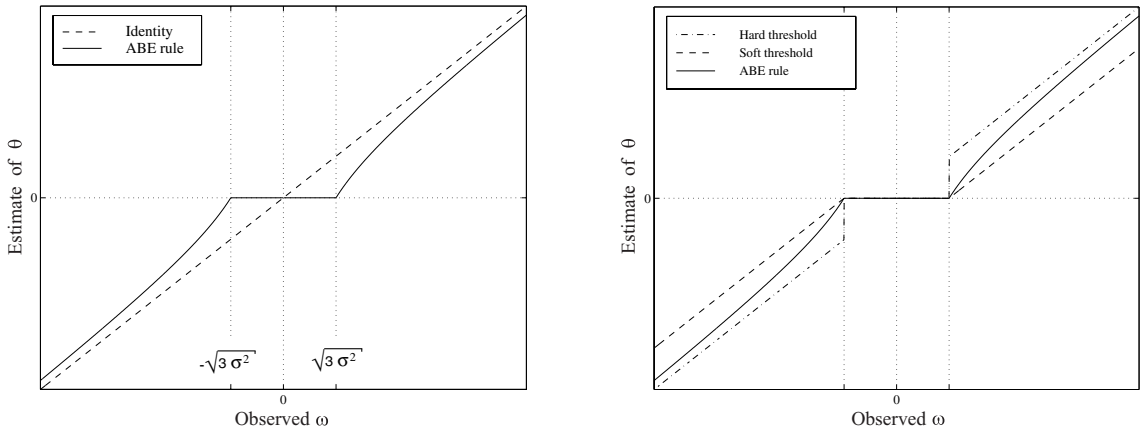

Figure 1. Left: ABE rule, with its fixed (with respect to the noise variance) threshold. Right: the ABE nonlinearity versus the hard and soft thresholding rules (here with the same threshold).

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 6) without requiring any tuning or adaptive estimation of parameters of the prior.

5. RELATION WITH OTHER APPROACHES

As we mentioned in the Introduction, Gao[12] has very recently proposed the use of the so-called "non-negative garrote" function, defined in (3), for wavelet-based denoising. Notice that the ABE rule (equation (22)) happens to be a "non-negative garrote" with a fixed threshold $\lambda = \sqrt{3\sigma^2}$:

$$\delta^{\mathrm{ABE}}(\omega) = \delta^{\mathrm{garrote}}_{\sqrt{3\sigma^2}}(\omega). \tag{23}$$

In his paper,[12] Gao credits the non-negative garrote to Breiman[27] who introduced it in the context of subset selection for regression problems. In that same paper, 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.

We also recently found that Brillinger[28] has briefly mentioned a similar function (in fact $\delta^{\mathrm{garrote}}_{\sigma}(\omega)$) as a possible alternative to the hard and soft rules. Brillinger states that this nonlinear function had been proposed by Tukey (in unpublished work of 1979), also in a regression context.

The non-negative garrote (specifically, $\delta^{\mathrm{garrote}}_{\sigma}(\omega)$) also happens to arise naturally in certain cross-validation methods, as used by Nowak[29] and Nowak and Baraniuk[30] to derive denoising rules.

6. EXPERIMENTAL RESULTS

6.1. Signal Denoising

We have evaluated our denoising rule versus the standard SureShrink and VisuShrink methods (based on soft thresholding nonlinearity), using Donoho and Johnstone's[3] well known test signals: "Blocks", "Doppler", "HeaviSine", and "Bumps". We have also included in our comparison a recent Bayesian approach based on mixture priors (as in Crouse, Nowak, and Baraniuk[10]) which, to the authors' knowledge, is representative of the very best Bayesian methods. Figure 2 reports the signal-to-noise ratios (SNR) obtained by each of the denoising methods, based on 100 runs for each original SNR value. As is clear from these results, the ABE rule performs consistently (i.e., for all four test signals and at all SNR levels) better than the 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, our rule achieves far superior results. Finally, as is also clear in Figure 2, the proposed technique performs comparably (except for the HeaviSine signal at low SNR) with the much more computationally demanding mixture based method.

Our experimental results allow adding the following conclusions to those of Gao[12]: at least for the signals and SNR values considered, a non-negative 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.

6.2. Image Denoising

For image denoising, we have compared the ABE rule, with its fixed threshold, versus the hard, the soft, and the garrote nonlinearities for a range of threshold values. Figure 3 shows the well-known "Cameraman" image after being contaminated by noise of standard deviation $\sigma = 20$. Figure 4 (left plot) then 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 an ideal threshold using the underlying true image (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 $\lambda = 34.9$ which is very close to our fixed threshold $\sqrt{3\sigma^2} \simeq 34.6$. The resulting denoised images are shown in Figure 3.

The same test was performed for two other values of σ (10 and 40), and the results are also reported in Figure 4. Again, our $\delta^{\mathrm{ABE}}(\omega)$ outperforms both the hard and soft rules, even with their ideal thresholds. The garrote rule (of which, recall, $\delta^{\mathrm{ABE}}(\omega)$ is a particular case) is able to find ideal thresholds with which it very slightly beats the ABE rule (however, recall that this ideal thresholds 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 $\sqrt{3}\sigma$.

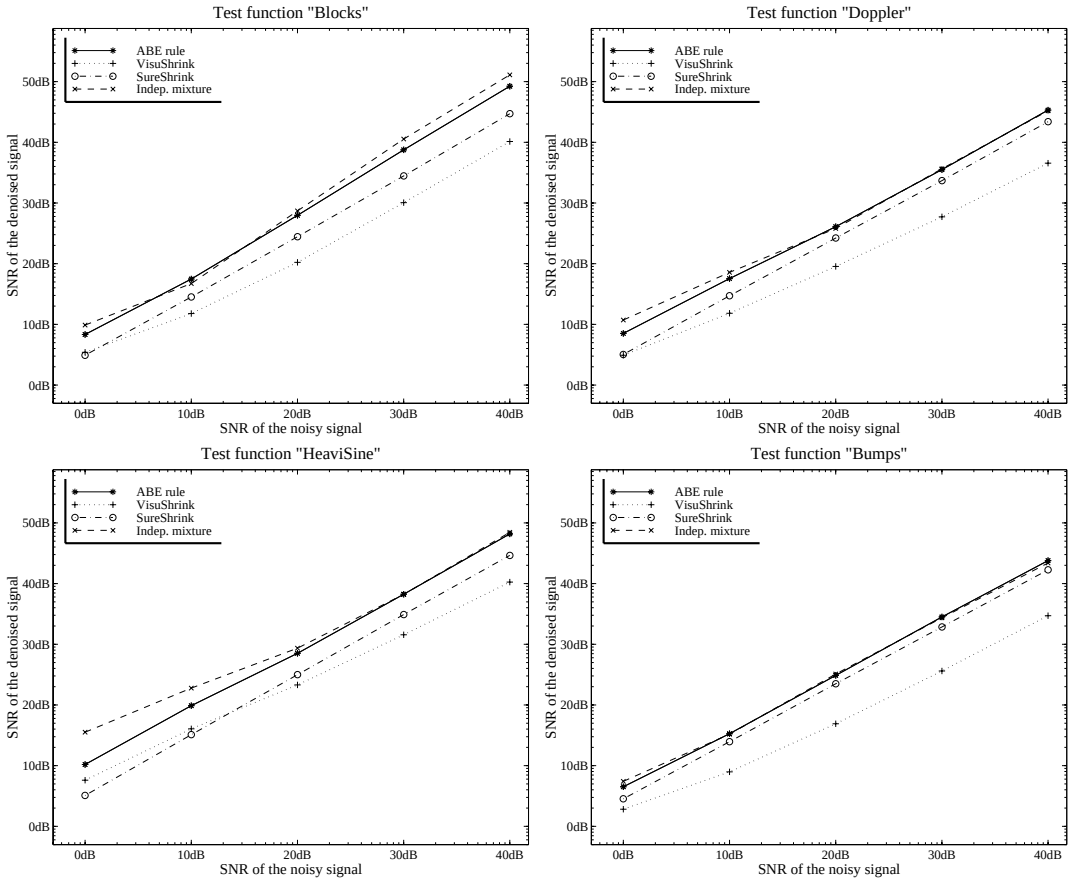

Figure 2. Input and output SNR for various wavelet denoising schemes applied to standard test signals (wavelets: Daubechies-2 for Blocks, Daubechies-8 for Doppler and HeaviSine, and Daubechies-6 for Bumps).

Finally, we repeated the same set of tests using a larger image composed of four different smaller sub-images; its noisy version (again for $\sigma = 20$) and the three denoised images can be seen in Figure 5. The MSE results reported in Figure 6 confirm the general behavior found in the previous tests: the ABE fixed-threshold nonlinearity (very approximately) coincides with the garrote with the ideal threshold and outperforms both the hard and the soft rules with their ideally chosen thresholds (except for $\sigma = 40$ where the soft rule with its ideal threshold yields a similar MSE; however, we stress again that this ideal threshold could not be found in a practical situation where we do not have access to the original image).

7. CONCLUSIONS AND FUTURE WORK

We have proposed an empirical-Bayes approach to wavelet-based image and signal estimation, where a non-informative (amplitude-scale invariant) prior plays a central role. A hierarchical/empirical Bayes path lead us to a simple fixed non-linear 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 Bayesian approach based on independent mixture priors (as in Crouse, Nowak, and Baraniuk[10]) which, to the authors' knowledge, 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 thresholds using the true original image. The excellent estimation performance of the non-informative 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

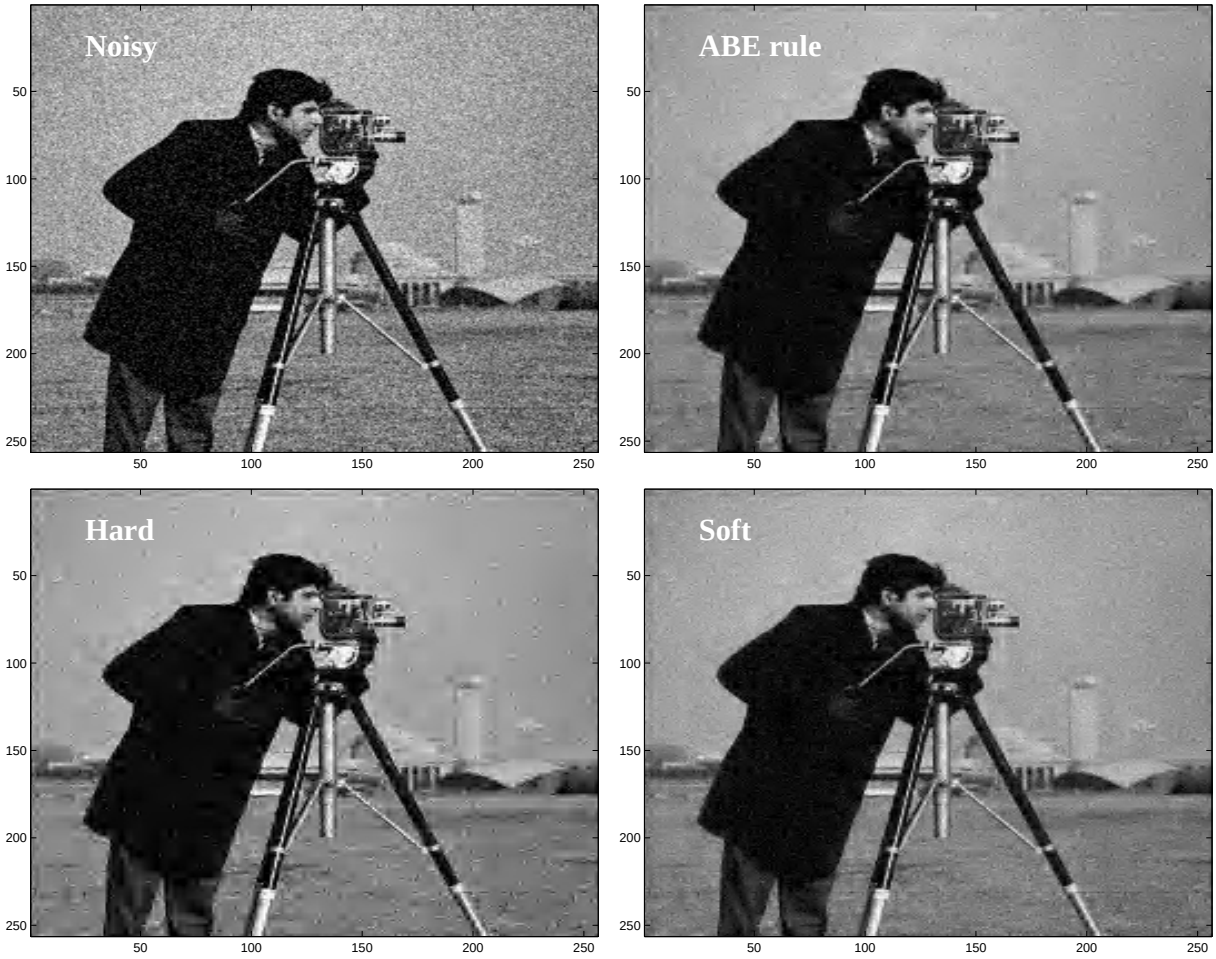

Figure 3. Cameraman noisy image ($\sigma = 20$) and denoised images produced by the ABE rule and the hard and soft rules with ideal thresholds (wavelet: Daubechies-2).

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[31]; actually, TI denoising can also be formalized through the use of non-informative priors.[32] 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. S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.
2. R. T. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhäuser, Boston, MA, 1997.
3. D. L. Donoho and I. M. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika* **81**, pp. 425–455, 1994.
4. B. Vidakovic, "Wavelet-based nonparametric Bayes methods," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, Dey, Müller, and Sinha, eds., vol. LNS 133, pp. 133–155, Springer-Verlag, 1998.
5. H. Krim and I. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Trans. on Inform. Theory* , April 1999. to appear.
6. K. Timmermann and R. Nowak, "Multiscale modeling and estimation of poisson processes with application to photon-limited imaging," *IEEE Transactions on Infomation Theory* **45**, 1999. To appear in the special issue on multiscale statistical signal analysis and its applications.
7. M. Figueiredo and J. Leitão, "Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle," *IEEE Transactions on Image Processing* **IP-6**, pp. 1089–1102, August 1997.

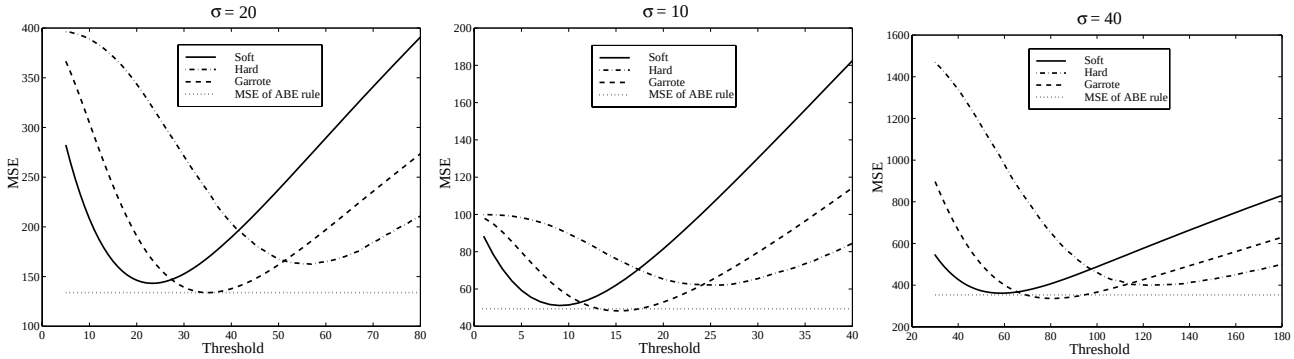

Figure 4. Mean squared errors (MSE) achieved by the hard, soft, and garrote nonlinearities, as function of the threshold values, for three noise standard deviations: $\sigma = 20$, $\sigma = 10$, and $\sigma = 40$ ("Cameraman" image). The horizontal dotted line shows the MSE obtained by the ABE rule (with its fixed threshold).

8. D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.* **90**, pp. 1200–1224, Dec. 1995.
9. C. Robert, *The Bayesian Choice: A Decision Theoretic Motivation*, Springer-Verlag, New York, 1994.
10. M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing* **46**, pp. 886–902, 1998.
11. H. Gao and A. Bruce, "WaveShrink with firm shrinkage," *Statistica Sinica* **7**, pp. 855–874, 1997.
12. H. Gao, "Wavelet shrinkage denoising using the non-negative garrote," *Journal of Computational and Graphical Statistics* **7**, pp. 469–488, December 1998.
13. J. Bernardo and A. Smith, *Bayesian Theory*, John Wiley & Sons, Chichester, United Kingdom, 1994.
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., pp. 151–193, Oxford University Press, 1993.
15. D. Ruderman, "The statistics of natural images," *Network: Computation in Neural Systems* **5**, pp. 517–548, 1995.
16. B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, pp. 607–609, 1996.
17. A. Hyvärinen, P. Hoyer, and E. Oja, "Sparse code shrinkage for image denoising," in *Proc. of the IEEE Intern. Joint Conf. on Neural Networks*, pp. 859–864, (Anchorage, Alaska), 1998.
18. J. Cardoso, "Blind signal separation: statistical principles," *Proc. of the IEEE* **86**(10), pp. 2009–2025, 1998.
19. P. Comon, "Independent component analysis: a new concept?," *Signal Processing* **36**(3), pp. 287–314, 1994.
20. A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation* **7**(6), pp. 1004–1034, 1995.
21. T. Lee, *Independent Component Analysis*, Kluwer Academic Publishers, Dordrecht, 1998.
22. H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Statist. Assoc.* **92**, pp. 1413–1421, 1997.
23. F. Abramovich, T. Sapatinas, and B. Silverman, "Wavelet thresholding via a Bayesian approach," *Journal of the Royal Statistical Society (B)* **60**, 1998.
24. B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *J. Amer. Statist. Assoc.* **93**, pp. 173–179, 1998.
25. C. Robert, *The Bayesian Choice: A Decision Theoretic Motivation*, Springer-Verlag, New York, 1994.
26. E. L. Lehmann, *Theory of Point Estimation*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1983.
27. L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics* **37**, pp. 373–384, 1995.
28. D. Brillinger, "Uses of cumulants in wavelet analysis," *Nonparametric Statistics* **6**, pp. 93–114, 1996.
29. R. D. Nowak, "Optimal signal estimation using cross-validation," *Signal Processing Letters* **4**(1), pp. 23–25, 1997.
30. R. D. Nowak and R. G. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *Proc. SPIE, Wavelet Applications in Signal and ImageProcessing V* **3169**, pp. 55–66, 1997.
31. R. Coifman and D. Donoho, "Translation invariant de-noising," in *Wavelets and Statistics*, Lecture Notes in Statistics, pp. 125–150, Springer-Verlag, (New York), 1995.
32. M. Figueiredo and R. Nowak, "Bayesian wavelet-based signal estimation using non-informative priors," in *Proc. Asilomar Conf. Signals, Systems, and Comput.*, pp. 1368–1373, 1998.

Figure 5. Noisy image ($\sigma = 20$) and denoised images produced by the ABE rule and the hard and soft rules with ideal thresholds (wavelet: Daubechies-2).

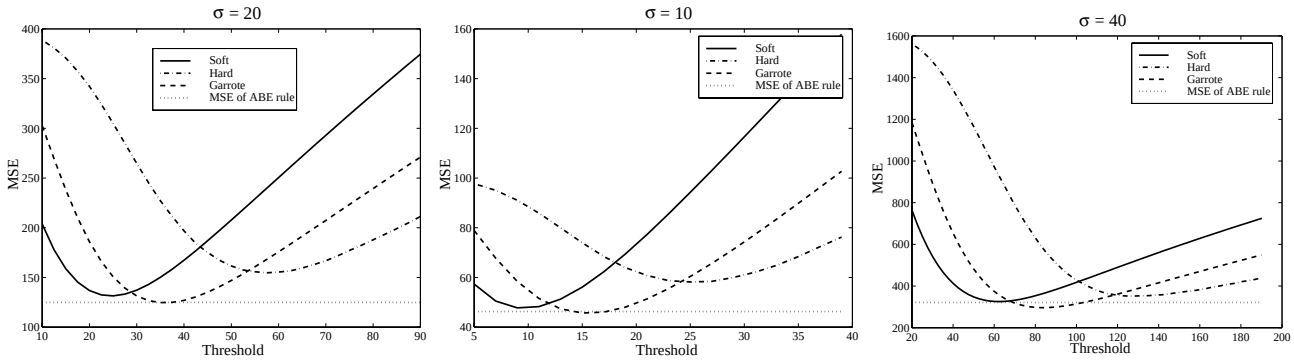

Figure 6. Mean squared errors (MSE) achieved by the hard, soft, and garrote nonlinearities, as function of the threshold values, for three noise standard deviations: $\sigma = 20$, $\sigma = 10$, and $\sigma = 40$ (image of Figure 5). The horizontal dotted line shows the MSE obtained by the ABE rule (with its fixed threshold).