PROC. THIRTY-SECOND ASILOMAR CONF. SIGNALS, SYSTEMS, AND COMP., PACIFIC GROVE, CA, NOV. 1998

BAYESIAN WAVELET-BASED SIGNAL ESTIMATION USING NON-INFORMATIVE PRIORS

Mário A. T. Figueiredo

Instituto de Telecomunicações, and
Dept. of Electrical and Computer Engineering.
Instituto Superior Técnico
1049-001 Lisboa, PORTUGAL

Robert D. Nowak

Dept. 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 signal denoising methods. Most existing schemes involve arbitrary thresholding nonlinearities and *ad hoc* threshold levels, or computationally expensive signal-adaptive procedures. Furthermore, because the DWT is not a translation-invariant (TI) transform, results of processing depend on the relative alignment between data and wavelets in a complicated manner. In the context of denoising, this non-stationarity can produce undesirable ("pseudo-Gibbs" or "blocking") artifacts. To overcome these deficiencies, we propose a new wavelet-based signal denoising technique derived using the theory of non-informative Bayesian priors. The resulting estimator is TI and employs a very simple fixed non-linear shrinkage/thresholding rule. Remarkably, our new approach is very computationally efficient and performs better than standard methods that are more computationally demanding.

1. INTRODUCTION

1.1. Background

The discrete wavelet transform (DWT) of real-world signals and images exhibits two important features: *sparseness*, *i.e.,* a few large coefficients dominate the representation, and *decorrelation*, *i.e.,* the transform coefficients have lower correlation than the original signal samples. These properties make the DWT ideal for many tasks including signal estimation and compression; see, *e.g.*, [1]. The idea is to process the DWT coefficients, rather than the signal samples, according to a three step program: **(i)** compute the DWT of the signal, **(ii)** perform some processing on the DWT coefficients, **(iii)** compute the inverse DWT of the processed coefficients to obtain the processed signal.

Stimulated by the seminal work in [2], a variety of denoising methods (following this approach) have been proposed; see [3, 4] and references therein. In this context, the decorrelation property justifies independent processing of each DWT coefficient; the sparseness property supports the adoption of threshold/shrinkage estimators. Standard choices are the *hard* and *soft* thresholding functions, with a fixed threshold level proportional to the noise standard deviation. "VisuShrink" is a well known method, based on the so-called "universal threshold" [2]; more sophisticated adaptive schemes have also been proposed, such as "SureShrink" [5], generally outperforming fixed rules.

Recently, wavelet-based estimation has been addressed within the Bayesian framework (see [4] and references therein), where the expected decorrelation and sparseness properties of the DWT are formally captured by an *a priori* probability distribution. This prior, combined via Bayes law with the likelihood function (noise model), leads to *a posteriori* distributions of the unknown signal conditioned on the observed one. An estimation rule can then be obtained after specifying a loss function, via standard Bayesian decision-theory [6]. Bayesian rules have been shown to outperform other methods and are the state-of-the-art in wavelet-based denoising [4, 13]. Moreover, the Bayesian formalism requires explicit modeling of all assumptions, thus providing insight into the mechanisms and trade-offs involved.

There are several problematic issues in existing schemes: in thresholding methods, the choice of the particular nonlinearity (*e.g.,* hard or soft) is usually arbitrary; fixed threshold levels are *ad hoc* and often have to be "tweaked" to yield good practical results; adaptive threshold selection methods also involve an arbitrary choice of nonlinearity and are computationally demanding. In many Bayesian methods previously proposed, the priors on the wavelet coefficients are chosen with the goal of matching empirical coefficient distributions or obtaining rules that mimic the conventional nonlinearities. Moreover, Bayesian methods are generally very computationally intensive.

Another important concern is the non-stationary nature of the DWT. Since the DWT is not a translation-invariant (TI) transform, processing results do depend on the relative alignment between the data set and the wavelets in a complicated manner. In the context of denoising, this non-stationarity can produce undesirable artifacts (*e.g.*, see Figure 4) . This is an instance of a more general question: how to perform TI data analysis using non-TI bases? The common "fix" in wavelet-based denoising is to perform standard (*i.e.*, non-TI) denoising for all possible relative shifts and then average the results [7, 8, 9]. This (TI) estimator can be implemented efficiently and generally outperforms non-TI methods: reduced artifacts, better quantitative performance (under a variety of error measures), more regular estimates (in approximation-theoretic sense [7, 10]). Furthermore, from a Bayesian viewpoint, TI methods have been recently shown to correspond to priors with smoother correlation

Partially supported by NATO through grant CRG–960010, NSF through grant MIP–9701692, and the Portuguese PRAXIS XXI program through grants BPD-14129-97 and 2/2.1/TIT-1580.

Email: mtf@lx.it.pt, nowak@egr.msu.edu
Web: www.img.lx.it.pt/∼mtf, www.egr.msu.edu/∼nowak

behavior (relative to non-TI methods) [11]. Nevertheless, until now, this TI method has remained an *ad hoc* solution, lacking an estimation-theoretic justification.

1.2. Contributions

This paper tackles the fundamental issues raised above (arbitrary threshold rules and translation invariance) using non-informative Bayesian priors [6]. We derive a nonlinear shrinkage/threshold rule which outperforms both VisuShrink and SureShrink and performs nearly as well as (sometimes better than) the best denoising methods in standard benchmark problems. Remarkably, since it is a universal-type fixed rule (no free parameters requiring tuning), our method is as computationally inexpensive as the simplest ones (*e.g.*, VisuShrink). Additionally, we derive a Bayesian TI denoising criterion based on a non-informative translation prior coupled with a quadratic loss function. To our knowledge, this is the first formal estimation-theoretic derivation of TI denoising.

The paper is organized as follows. In Section 2, the denoising problem is presented, notation is introduced, and the Bayesian formulation is described. Section 3 addresses TI denoising in the Bayesian framework. A new non-informative prior is proposed in Section 4. In Section 5, a novel procedure is derived which is an empirical Bayes approach based on the proposed non-informative prior. Section 6 contains experimental comparisons of the performance of the new algorithm versus other methods. A final discussion and some conclusions are given in Section 7.

2. PROBLEM FORMULATION

2.1. Wavelet-based Denoising

Suppose $\mathbf{y} = [y_1, \ldots, y_N]^T$ is a vector of noisy observations of a discrete signal $\mathbf{x} = [x_1, \ldots, x_N]^T$,

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

where the noise vector $\mathbf{n}$ is comprised of N independent samples from a zero-mean Gaussian variable of variance σ^2, that is, $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, where $\mathbf{I}$ denotes an identity matrix. The goal of a denoising procedure is to recover $\mathbf{x}$ from $\mathbf{y}$.

In wavelet-based denoising, the DWT $\mathcal{W}$ is applied to the noisy data yielding the noisy *wavelet coefficients* $\boldsymbol{\omega} = \mathcal{W}\mathbf{y}$ which are described by an similar observation model

$$\boldsymbol{\omega} = \mathcal{W}\mathbf{x} + \mathcal{W}\mathbf{n} = \boldsymbol{\theta} + \mathbf{n}', \tag{2}$$

where $\mathbf{n}' \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, since $\mathcal{W}$ is an orthonormal transform (matrix), *i.e.*, $\mathcal{W}\mathcal{W}^T = \mathbf{I}$. The wavelet transforms of the majority of meaningful signals tend to be sparse, *i.e.*, a few large coefficients dominate the representation [2]. On the other hand, $\mathbf{n}' = \mathcal{W}\mathbf{n}$ is a set of i.i.d. Gaussian distributed coefficients thus, with high probability, bounded in magnitude by some level (proportional to their standard deviation). Therefore, if the magnitude of a wavelet coefficient in $\boldsymbol{\omega}$ exceeds a specified threshold, then its signal component is probably much larger than the noise. This observation, together with the decorrelation property justifying independent processing of each coefficient, pointed the way to simple denoising schemes based on threshold operations. This simple rationale underlies the (now classical) method proposed in [2] and all its variants [3]. Formally, the estimate of the i-th coefficient θ_i is $\widehat{\theta}_i = \delta_\nu(\omega_i)$, where δ_ν is either the hard or soft thresholding function (see Fig. 2) and ν is the threshold level. From the estimates $\widehat{\boldsymbol{\theta}} = \{\widehat{\theta}_i\}$, the inverse DWT yields a signal estimate $\widehat{\mathbf{x}} = \mathcal{W}^{-1}\widehat{\boldsymbol{\theta}}$.

One degree of freedom which is not explicit in the above formulation is the relative alignment (shift) between the data $\mathbf{y}$ and the wavelet basis implicit in $\mathcal{W}$. To be explicit, we now write $\mathcal{W}_s$, for $s \in S$, where $S = \{0, \ldots, N-1\}$ is the set of all possible shifts[1] between the wavelet basis and the data. This notation allows expressing the TI method discussed in the introduction [7, 8, 9] as

$$\widehat{\mathbf{x}} = \frac{1}{N} \sum_{s \in \mathcal{S}} \mathcal{W}_s^{-1} \left(\boldsymbol{\delta} \left(\mathcal{W}_s \mathbf{y} \right) \right), \tag{3}$$

where $\boldsymbol{\delta}(\cdot)$ stands for the element-wise application of the adopted threshold or shrinkage rule (*e.g.*, of $\delta_\nu(\cdot)$).

2.2. Bayesian Formulation

The likelihood function resulting from the observation model (1) is Gaussian, mean $\mathbf{x}$, and covariance $\sigma^2\mathbf{I}$,

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I}), \tag{4}$$

or equivalently, in the wavelet domain,

$$\boldsymbol{\omega}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I}). \tag{5}$$

To capture the sparseness and decorrelation properties of the DWT, the (signal) prior $p_X(\mathbf{x})$ is formulated on the wavelet coefficients $\boldsymbol{\theta} = \mathcal{W}\mathbf{x}$; that is, a $p_\Theta(\boldsymbol{\theta})$ is specified which induces[2] $p_X(\mathbf{x}) = p_\Theta(\mathcal{W}\mathbf{x})$. The Bayesian version of the three step programme is: **(i)** compute the DWT of the data $\boldsymbol{\omega} = \mathcal{W}\mathbf{y}$; **(ii)** obtain an optimal (according to some loss function) estimate $\widehat{\boldsymbol{\theta}}$ given $\boldsymbol{\omega}$; **(iii)** obtain a signal estimate $\widehat{\mathbf{x}} = \mathcal{W}^{-1}\widehat{\boldsymbol{\theta}}$. To see under which conditions this procedure does yield a *Bayes-optimal* signal estimate, let us explicitly write the estimation rule as the minimizer of the *a posteriori* expected loss [6]; specifically,

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}^*} \int L(\mathbf{x}, \mathbf{x}^*) p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}, \tag{6}$$

where $L(\mathbf{x}, \mathbf{x}^*)$ is the adopted loss function measuring the "discrepancy" between $\mathbf{x}$ and a candidate estimate $\mathbf{x}^*$; $p(\mathbf{x}|\mathbf{y})$ is the *a posteriori* probability function, as usually obtained via Bayes law as $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})$. Inserting $\mathcal{W}^{-1}\boldsymbol{\theta} = \mathbf{x}$ in (6), after noticing that[2] $d\mathbf{x} = d\boldsymbol{\theta}$, $d\mathbf{y} = d\boldsymbol{\omega}$, and $p(\mathbf{x}|\mathbf{y}) = p(\boldsymbol{\omega}|\boldsymbol{\theta})$, and since

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x})\, p_X(\mathbf{x}) \\ &= p(\boldsymbol{\omega}|\boldsymbol{\theta})\, p_X(\mathcal{W}^{-1}\boldsymbol{\theta}) \\ &= p(\boldsymbol{\omega}|\boldsymbol{\theta})\, p_\Theta(\boldsymbol{\theta}) \;\propto\; p(\boldsymbol{\theta}|\boldsymbol{\omega}), \end{aligned} \tag{7}$$

we can write

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

[1] Throughout this paper we assume the use of the periodic (or circular) DWT and all shifts are taken to be circular as well.

[2] Notice that $d\boldsymbol{\theta} = d\mathbf{x}$ and $d\mathbf{y} = d\boldsymbol{\omega}$ because $\mathcal{W}$ is an orthonormal transformation (matrix), thus possessing a unit Jacobian.

Now, if $L(\mathcal{W}^{-1}\boldsymbol{\theta}, \mathcal{W}^{-1}\boldsymbol{\theta}^*) = L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, then

$$\begin{aligned}\widehat{\mathbf{x}} &= \mathcal{W}^{-1} \arg\min_{\boldsymbol{\theta}^*} \int L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) p(\boldsymbol{\theta}|\boldsymbol{\omega})\, d\,\boldsymbol{\theta}; \\ &= \mathcal{W}^{-1} \delta_{\text{Bayes}}(\mathcal{W}\mathbf{y}) \end{aligned} \tag{9}$$

where $\delta_{\text{Bayes}}(\cdot)$ is the resulting optimal Bayes estimation rule in the wavelet domain. This is a formal derivation of the standard programme followed (often without clear justification) in all Bayesian approaches to wavelet-based denoising. It happens that the two most common loss functions do verify the sufficient condition:

- For squared error loss, which leads to the posterior mean estimate, $L_2(\mathcal{W}^{-1}\boldsymbol{\theta}, \mathcal{W}^{-1}\boldsymbol{\theta}^*) = \|\mathcal{W}^{-1}\boldsymbol{\theta} - \mathcal{W}^{-1}\boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2 = L_2(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is a trivial consequence of the orthonormality of the DWT.
- For 0/1 loss, which produces the *maximum a posteriori* (MAP) estimate when $\varepsilon \to 0$,

$$L_{0/1}^{\varepsilon}(\mathbf{x}, \mathbf{x}^*) = \begin{cases} 0, & \|\mathbf{x} - \mathbf{x}^*\|_2 \leq \varepsilon \\ 1, & \text{otherwise} \end{cases}$$

which is obviously equal to $L_{0/1}^{\varepsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$, again due to the orthonormality of $\mathcal{W}$.

Notice that this is not true in general; it is easy to come up with many loss functions that do not satisfy this condition.

3. BAYESIAN TRANSLATION INVARIANT DENOISING

Let us now revisit TI denoising under a Bayesian perspective. Like the classical wavelet denoising schemes, the Bayes estimators described above are implicitly dependent on the alignment between data and wavelets. Equation (9), with explicit reference to s, becomes

$$\widehat{\mathbf{x}}_s = \mathcal{W}_s^{-1} \delta_{\text{Bayes}}(\mathcal{W}_s \mathbf{y}). \tag{10}$$

The Bayesian approach is to consider s as an additional unknown characterized by a prior $p(s)$. A uniform probability mass function $p(s) = \frac{1}{N}$ on the set of all shifts $\mathcal{S}$, is a non-informative prior expressing that no one shift is *a priori* favoured. A moments thought reveals that this is the only prior that can possibly lead to a TI estimate.

We are now in a position to derive a TI Bayes estimator. To begin, notice that we have a pair of unknowns $(\mathbf{x}, s)$, and therefore a new loss function is needed. The fact that we are only interested in estimating $\mathbf{x}$ is formalized by adopting a squared error loss function independent of s,

$$L((\mathbf{x}, s), (\mathbf{x}^*, s^*)) = L_2(\mathbf{x}, \mathbf{x}^*). \tag{11}$$

Accordingly, the Bayes estimate is the posterior mean

$$\begin{aligned}\widehat{\mathbf{x}} &= \sum_{s\in\mathcal{S}} \int \mathbf{x}\, p(\mathbf{x}, s|\mathbf{y})\, d\mathbf{x} \\ &= \sum_{s\in\mathcal{S}} p(s|\mathbf{y}) \underbrace{\int \mathbf{x}\, p(\mathbf{x}|\mathbf{y}, s)\, d\mathbf{x}}_{\widehat{\mathbf{x}}_s}, \end{aligned} \tag{12}$$

where $\widehat{\mathbf{x}}_s$ is simply the shift-dependent posterior mean, *i.e.*, the estimate in (10), under squared error loss. So, the TI estimate is simply a weighted average of all shift-dependent estimates[3]. What remains, is to identify the weights $p(s|\mathbf{y})$.

Using Bayes rule and the fact that $p(s)$ is constant (flat),

$$p(s|\mathbf{y}) = \frac{p(\mathbf{y}|s)p(s)}{\sum_{s\in S} p(\mathbf{y}|s)p(s)} = \frac{p(\mathbf{y}|s)}{\sum_{s\in S} p(\mathbf{y}|s)}, \tag{13}$$

showing that the weights in (12) are proportional to the marginal likelihoods $p(\mathbf{y}|s)$. This shows that, in general, the optimal TI estimator weights each shift-dependent estimator proportionally to the *evidence* $p(\mathbf{y}|s)$ given by the data in favor of the corresponding shift s. That is, the Bayes-optimal TI rule is

$$\widehat{\mathbf{x}} = \sum_{s\in\mathcal{S}} p(s|\mathbf{y})\, \mathcal{W}_s^{-1} \delta_{\text{Bayes}}(\mathcal{W}_s \mathbf{y}), \tag{14}$$

where $\delta_{\text{Bayes}}(\cdot)$ stands for the posterior mean wavelet based estimation rule (derived from a specific wavelet domain prior). Only if $p(\mathbf{y}|s) =$ const. would this optimal TI estimator coincide with the standard TI method in (3).

4. A NEW PRIOR FOR WAVELET COEFFICIENTS

Since (under certain loss functions) Bayes-optimal signal denoising can be carried out in the wavelet domain, let us focus on the choice of a prior for the wavelet coefficients. With the decorrelation property giving support to modeling the coefficients independently, the standard approach is to model each coefficient with an *informative* prior that attempts to explicitly capture the sparseness property; *i.e.*, using heavy-tailed densities [12, 13, 4]. Here, we take a different approach; the coefficients are still modeled as independent, but we attempt to remain non-informative, letting the data speak for themselves. Formally, we adopt a hierarchical Bayesian framework with the following levels.

- According to the noise model above, $\omega_i|\theta_i \sim \mathcal{N}(\theta_i, \sigma^2)$, where σ^2 is assumed known.[4]
- Each (unknown) coefficient is modeled as conditionally zero-mean Gaussian $\theta_i|\tau_i^2 \sim \mathcal{N}(0, \tau_i^2)$, for $\tau_i^2 \geq 0$.
- Total ignorance about each (unknown) variance is expressed by a non-informative improper[5] Jeffreys (hyper) prior $p(\tau_i^2) \propto \frac{1}{\tau_i^2}$.

This non-informative prior exhibits the following important scale invariance property: if the data $\mathbf{y}$ (equivalently $\{\omega_i\}$ and noise standard deviation σ) are re-scaled by an arbitrary factor (*e.g.*, corresponding to a change in measurement units), then any inference results are not effected, apart from the corresponding re-scaling of the estimated signal. For more details on Jeffreys priors and invariance, see, *e.g.*, [6]. Other Bayesian denoising methods that the authors are aware of (based on Laplacian, Gaussian mixture, or other heavy-tailed densities, for example) do not share this desirable invariance property.

[3] This can be identified as a *model averaging* procedure [14].

[4] Assuming known noise variance is not a shortcoming; excellent estimates are easily obtained, *e.g.*, from the MAD scheme [2].

[5] A density function is called *improper* if it is not normalizable because its integral is not finite. Improper priors are common in Bayesian inference [6]; in fact, only the relative weighting expressed by the shape of the prior impacts the *a posteriori* density [6, 14].

To gain some insight into our new prior, let us consider the marginal *a posteriori* density $p(\theta_i|\omega_i)$; its defining expression,

$$p(\theta_i|\omega_i) = \frac{p(\omega_i|\theta_i)}{p(\omega_i)} \underbrace{\int p(\theta_i|\tau_i^2)p(\tau_i^2)d\tau_i^2}_{p(\theta_i)}, \qquad (15)$$

reveals the presence of an equivalent prior $p(\theta_i)$ which is a continuous mixture of zero-mean Gaussians, weighted according to the Jeffreys prior $p(\tau_i^2) \propto \frac{1}{\tau_i^2}$. Since this prior is the limiting case of the conjugate inverse-Gamma family [6], the resulting prior $p(\theta_i)$ is itself a limiting case of a family of Student-t densities; this is so because t densities can be seen as mixtures of Gaussians of common mean with an inverse-Gamma weighting of the variance [14]. In particular, the integration indicated in (15) yields $p(\theta_i) \propto \frac{1}{|\theta_i|}$. Interestingly, this prior is itself scale-invariant, symmetric, extremely heavy-tailed, and improper.

Both continuous and finite Gaussian mixtures have been used before by several authors, as informative priors for wavelet-based denoising. Finite mixtures of Gaussians were considered in [12, 13]; these require (hyper) parameter specification or estimation from the data, which is a crucial issue due to the non-invariant nature of these priors. Student-t densities (*i.e.*, continuous mixtures of Gaussians which are common robust substitutes for Gaussian priors [14]) have been used in wavelet-based denoising with specially selected parameter settings [4]. Our (non-informative) prior leaves **no** free parameters to adjust.

5. A NEW WAVELET-BASED DENOISING ALGORITHM

It turns out that the hierarchical Bayesian setup built in the previous section leads to an improper *a posteriori* probability density function (15). This fact is well known from other applications where similar hierarchical Bayes formulations are used (see, *e.g.*, [14], pages 139-140).

To bypass this difficulty we adopt a *parametric empirical Bayes*-type approach [15], *i.e.*, we break the fully Bayesian analysis chain as follows:

- First, an estimate $\widehat{\tau_i^2}$ is obtained according to the MAP criterion based on the marginal likelihood $p(\omega_i|\tau_i^2)$ and on the corresponding Jeffreys prior.
- Given the estimate $\widehat{\tau_i^2}$, both the MAP and the posterior mean estimates of θ_i are given by the well known shrinkage estimator, resulting from a Gaussian likelihood (of variance σ^2) together with a $\mathcal{N}(0, \widehat{\tau_i^2})$ prior,

$$\widehat{\theta}_i = \frac{\widehat{\tau_i^2}}{\widehat{\tau_i^2} + \sigma^2}\omega_i. \qquad (16)$$

Since $\omega_i = \theta_i + n_i'$, the marginal likelihood is very simply $\omega_i|\tau_i \sim \mathcal{N}(0, \tau_i^2 + \sigma^2)$. The Jeffreys prior is now $p(\tau_i^2) \propto 1/(\tau_i^2 + \sigma^2)$, with the corresponding MAP estimate being

$$\widehat{\tau_i^2} = \arg\max_{\tau_i^2 \geq 0} \frac{e^{-\frac{\omega_i^2}{2(\tau_i^2+\sigma^2)}}}{(\tau_i^2+\sigma^2)^{3/2}} = \left(\frac{\omega_i^2}{3} - \sigma^2\right)_+, \qquad (17)$$

where $(\cdot)_+$ stands for "the positive part of", *i.e.*, $(x)_+ = x$, if $x \geq 0$, and $(x)_+ = 0$, if $x < 0$. By plugging this estimate into (16), we obtain our final shrinkage/thresholding rule,

$$\widehat{\theta}_i = \frac{(\omega_i^2 - 3\sigma^2)_+}{\omega_i}, \qquad (18)$$

plotted in Figure 1. In Figure 2, the new 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 these two rules, behaving close to the soft rule for small ω_i, and close to the hard rule for large ω_i. An important feature of our rule is that, unlike the soft threshold, it approaches identity as the observed value becomes large (see Figure 1).

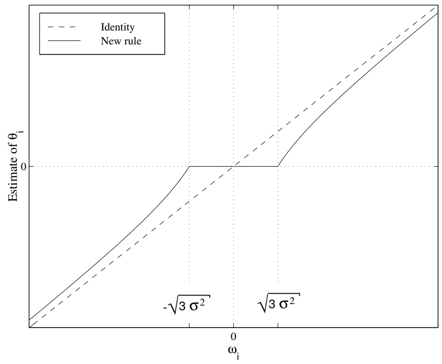

Figure 1: New non-linear shrinkage/thresholding rule, with its fixed (with respect to the noise variance) threshold.

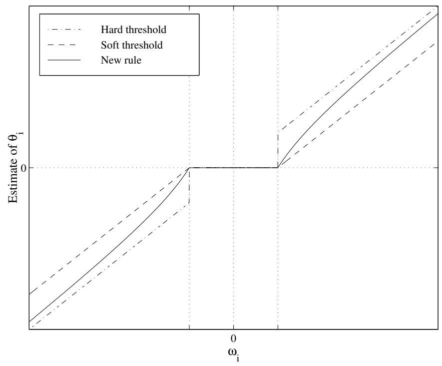

Figure 2: This figure illustrates how the new rule, when the same threshold is adopted, exhibits a behavior between those of the hard and soft thresholding rules.

The Bayesian (variance) estimator in (17) has an interesting (frequentist) interpretation. Ignoring the $(\cdot)_+$ function (necessary simply because we are estimating τ_i^2 from an estimate of $\tau_i^2 + \sigma^2$, and the valid parameter space is $I\!R_0^+$), this is an instance of the problem: given n i.i.d. $\mathcal{N}(0, \gamma^2)$ observations, $x_1, ..., x_n$, what is the best estimator of the form $\widehat{\gamma^2} = k(x_1^2 + \cdots + x_n^2)$, in a mean squared error (MSE) $E[(\gamma^2 - \widehat{\gamma^2})^2]$ sense? It is well known that $k = 1/(n+2)$ (in our case, $n = 1$, $k = 1/3$) yields the minimum MSE (although biased) estimate of γ^2 [16]. This coincides with the MAP rule with a Jeffreys prior on γ^2.

6. PERFORMANCE COMPARISON

Here we compare the performance of various TI wavelet based estimators. Two non-Bayesian methods, Sureshrink and Visushrink, are tested; in these methods, the TI estimates are obtained by applying the standard (non-TI) denoising for all possible relative shifts and then taking the (un-weighted) average of the results, which is the conventional non-Bayesian approach to TI denoising [7, 8, 9]. Our empirical (non-informative) Bayesian method is also compared with another (more complicated and computationally demanding) empirical Bayesian approach in which each wavelet coefficient is modeled as an independent Gaussian mixture random variable [13]. To our knowledge, the Gaussian mixture based method is representative of the very best wavelet denoising methods available, and hence it serves as a good benchmark for our new, less computationally demanding technique. In both cases, because both approaches are empirical Bayesian methods, a means for properly calculating or estimating $p(s|\mathbf{y})$ is not readily apparent. Hence, in both cases we, again, employ the standard TI averaging (equivalent to approximating $p(s|\mathbf{y})$ as a constant). If we pursued a fully Bayesian denoising scheme, then we could easily calculate the true value of $p(s|\mathbf{y})$ and compute the optimal TI posterior mean estimate given by the weighted average in (14).

As shown in Figure 3, the new rule performs consistently (i.e., for several test signals and a wide range of SNRs) better than the widely accepted (and computationally heavier) SureShrink. With respect the standard approach using the "universal threshold" (VisuShrink) [2], which has a similar computational load, our rule achieves far superior results. Moreover, the performance of the proposed technique is comparable with the far more computationally demanding Gaussian mixture based method. We also note that, although not shown here, our experiments have confirmed that the MSE performance of TI methods is slightly better than that of their non-TI counterparts. The subjective improvement of the TI methods is apparent in the results shown in Figure 4.

7. CONCLUSIONS

In this paper we have addressed the ad hoc selection of threshold rules and TI wavelet based estimation using the theory of non-informative Bayesian priors. In particular, we have developed a new empirical Bayes wavelet-based denoising rule using a non-informative Jeffreys prior. Our new rule performs remarkably simple non-linear shrinkage/thresholding, and (unlike other Bayesian schemes) has no free parameters requiring tuning, estimation, or elicitation. Moreover, it outperforms both VisuShrink and SureShrink and performs nearly as well as (sometimes better than) the best denoising methods in standard benchmark problems (to our knowledge, the best existing techniques are Bayesian; *e.g.*, see the recent comparisons in [13]). A Bayesian TI denoising method that is optimal under squared loss was derived based on a non-informative uniform prior placed on the shift. We pointed out that this optimal TI method can be used in conjunction with *any* wavelet coefficient prior. However, we also note that it is difficult to compute the evidences $p(s \mid \mathbf{y})$ (optimal weights for averaging shift-dependent estimates) under an empirical Bayes approach like the one underlying our new rule. We are currently investigating methods for estimating $p(s \mid \mathbf{y})$ in such cases.

8. REFERENCES

[1] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *App. and Comp. Harmonic Analysis*, vol. 1, pp. 100–115, Dec. 1993.

[2] D. L. Donoho and I. M. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[3] R. T. Ogden, *Essential Wavelets for Statistical Applications and Data Analysis*. Boston, MA: Birkhäuser, 1997.

[4] B. Vidakovic, "Wavelet-based nonparametric Bayes methods," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, vol. LNS 133, pp. 133–155, Springer-Verlag, 1998. Editors Dey, Müller and Sinha.

[5] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

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

[7] R. Coifman and D. Donoho, "Translation invariant denoising," in *Lecture Notes in Statistics: Wavelets and Statistics*, vol. New York: Springer-Verlag, pp. 125–150, 1995.

[8] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Nonlinear processing of a shift invariant DWT for noise reduction," in *SPIE Symp. OE/Aerospace Sensing and Dual Use Photonics*, (Orlando, FL), 1995.

[9] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Lecture Notes in Statistics: Wavelets and Statistics*, vol. New York: Springer-Verlag, pp. 281–299, 1995.

[10] K. Berkner and J. R. O. Wells, "Smoothness estimates for soft-thresholding denoising via translation invariant wavelet transform," Tech. Rep. CML TR 98-01, Computational Mathematics Laboratory, Rice University, Houston, 1999.

[11] R. D. Nowak, "Shift invariant wavelet-based statistical models and $1/f$ processes," *Proc. IEEE Digital Signal Processing Workshop*, Bryce Canyon, UT, 1998.

[12] H. A. Chipman, E. D. Kolaczyk, and R. E. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 92, pp. 1413–1421, 1997.

[13] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, 1998.

[14] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. London: Chapman & Hall, 1995.

[15] B. Carlin and T. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall, 1996.

[16] E. L. Lehmann, *Theory of Point Estimation*. Pacific Grove, CA: Wadsworth & Brooks/Cole, 1983.

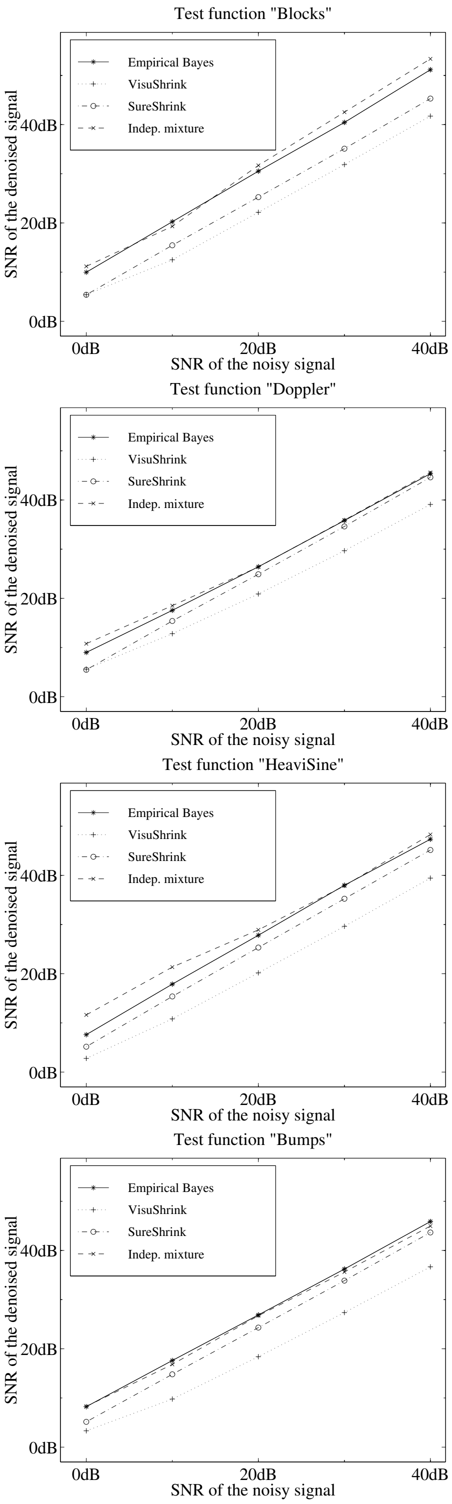

Figure 3: Input and output SNR values for various TI wavelet denoising schemes applied to Donoho and Johnstone's test signals. The wavelets used were: Daubechies-2 (Haar) for Blocks, Daubechies-8 for Doppler and HeaviSine, and Daubechies-6 for Bumps.

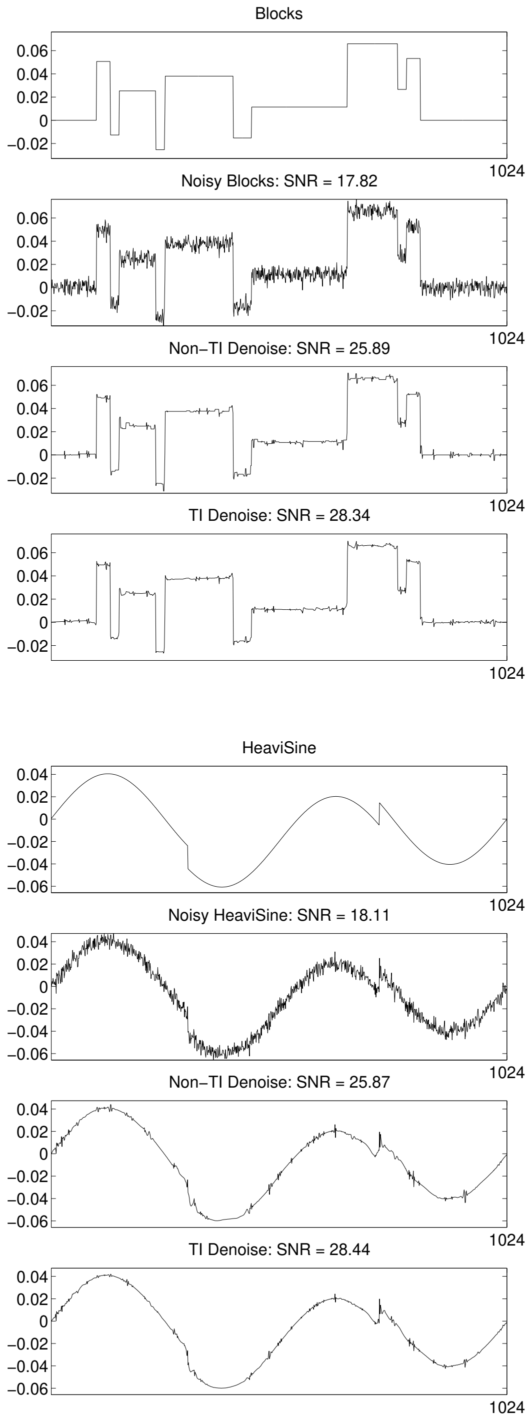

Figure 4: Comparison of TI and non-TI empirical Bayes estimator.