PARSIMONY AND WAVELET METHODS FOR DENOISING

H. Krim, J.-C. Pesquet and I.C. Schick†*

Stochastic Systems Group, LIDS,
MIT, Cambridge, MA, USA
* LSS/Univ. Paris Sud, ESE,
Plateau de Moulon, 91192 Gif sur Yvette, France
† GTE Internetworking and
DEAS, Harvard University, Cambridge, MA, USA

ABSTRACT

Some wavelet-based methods for signal estimation in the presence of noise are reviewed in the context of the parsimonious representation of the underlying signal. Three approaches are considered. The first is based on the application of the MDL principle. The robustness of this method is improved in the second approach, by relaxing the assumption of known noise distribution following Huber's work. In the third approach, a Bayesian strategy is adopted in order to incorporate prior information pertaining to the signal of interest; this method is especially useful at low signal-to-noise ratios.

1. INTRODUCTION

Model parsimony has been of growing interest to researchers in recent years, motivated by factors as diverse as storage in computer memory, computational efficiency, and communication.

The proposed techniques are many, each entailing heuristics and allowing interpretations proper to a particular application. As a result, the common theme uniting these different approaches sometimes seems hopelessly inaccessible. Nevertheless, it is possible to cast many of these application-specific methodologies as problems of "regularization".

It is often desired to limit the number of degrees of freedom in inverse problems by assuming a prior and thereby mitigating their ill-posedness. In pattern recognition, one is typically interested in the most parsimonious model that captures whatever information in the data is deemed essential, while a penalty for model mismatch plays the role of a prior for model parameters.

In this paper, we discuss several wavelet-based methods for signal estimation in the presence of noise, within the context of the parsimonious representation of the underlying signal. We show in particular that Rissanen's Minimum Description Length (MDL) principle can be applied to wavelet reconstructions to determine the complexity of the signal representation, i.e. to choose which coefficients to include in the reconstruction, and which to dismiss as noise.

The paper is organized as follows: Section 2 presents the problem statement. In Section 3, we highlight the importance of model parsimony to the signal denoising problem. In Section 4, we describe two information-theoretic methods for signal estimation via MDL. Finally, in Section 5, we discuss a statistical approach that permits the introduction of prior information on the signal of interest when its is embedded in high-intensity noise.

2. PROBLEM STATEMENT AND NOTATION

The estimation problem of interest in this paper assumes the following observation model:

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ is an unknown signal corrupted by the zero-mean noise process $n(t)$.

The underlying signal is modeled as an orthonormal basis representation,

$$s(t) = \sum_i C_i^s \psi_i(t),$$

which in turn leads to the working model

$$C_i = C_i^s + C_i^n, \quad i \in \{1, \cdots, K\}, \tag{2}$$

where C_i are the corrupted coefficients. In many cases, the noise coefficients C_i^n can be assumed independent; they share the same second-order statistical properties as $n(t)$, when this is a white noise sequence. Our problem is to recover or reconstruct $s(t)$ from the orthogonal transform of the observed process $x(t)$.

3. PARSIMONY AND DENOISING

The unitary transformation of a process afforded by the wavelet decomposition provides a complete statistical characterization of that process in the transform domain. The fact that the properties of the underlying signal and of the contaminating noise are well characterized, together with the orthogonality of the transform (which maximally removes any redundancy), suggest the potential efficiency of this approach for the statistical separation of signal and noise. An additional feature of this transformation, which in many cases turns out to be crucial, is the property of vanishing moments of the basis functions. This property tends to concentrate energy into very few dimensions. If the noise is white, then a subset of the dimensions will represent mostly signal, and the identification of this subset is very reminiscent of the model order identification problem, where space is partitioned into what might be referred to as the signal subspace and the noise subspace,

$$\mathcal{C} = \mathcal{C}_s \oplus \mathcal{C}_n \ . \tag{3}$$

This subspace identification can also be carried out objectively through the likelihood of $\boldsymbol{C}$. A model prior on the parameters must now be assigned first to reduce the class of possible models, and to account for model mismatch [7]:

$$\mathcal{L}\,(\boldsymbol{C}, K, P) = -\log p(\boldsymbol{C} \mid \boldsymbol{C}_s) + \alpha(K, P),$$

where K and P are respectively the data length and the number of signal dimensions. Rissanen refers to $\mathcal{L}$ as *description length* which, upon minimization, represents the coding length of the observed series $\{C_i\}$. This coding parsimony, together with the model summarizing the pertinent information underlying the process, form the basis of an interesting methodology developed over the last few years and retraced below with the rationale and hindsight afforded by time.

4. AN INFORMATION-THEORETIC APPROACH

In what follows, we assume that the underlying signal $s(t)$ is a deterministic but unknown signal in $\mathrm{L}^2(\mathbb{R})$. For the contaminating noise, we consider two cases:

- The probability density function of the contaminating noise is assumed to be known.
- The probability density function of the contaminating noise is unknown, but belongs to a known class; the worst-case scenario is sought within a minimax framework [8].

4.1. Coding for Denoising

The property of wavelets of concentrating energy into relatively a few coefficients and its inability to achieve that with noise, simplifies our formulation of the denoising problem as one of compression. The efficiency of the resulting solution is qualitatively and quantitatively reflected by the MDL, whose rationale is to seek and determine the shortest coding length of a data sequence $\{C_i\}_{1\le i\le K}$ which best summarizes the relevant information embedded in the observed process. Recall the coefficients are assumed independent. It then follows that their joint probability density function (pdf) is,

$$\exp\left(-\sum_{i=1}^{K}\varphi_i(C_i - C_i^s)\right)$$

where φ_i is a known "potential" function. For instance, by choosing

$$\varphi_i(u) = \frac{|u|^{\beta_i}}{\gamma_i^{\beta_i}} - \log\left(\frac{\beta_i}{2\gamma_i\Gamma(1/\beta_i)}\right) , \tag{4}$$

an exponential-power distribution is obtained with $(\beta_i, \gamma_i) \in (\mathbb{R}_+^*)^2$. The Gaussian distribution corresponds to $\beta_i = 2$ and the Laplacian distribution to $\beta_i = 1$. Note that different functions φ_i can be chosen so as to take into account, for example, different statistics of the noise at each scale. The above pdf can be viewed as a function $p(C_1, \ldots, C_K \mid \boldsymbol{\zeta})$ where the parameter vector is given by

$$\boldsymbol{\zeta} = (i_1, \ldots, i_P, C_{i_1}^s, \ldots, C_{i_P}^s) \ , \tag{5}$$

P being the number of "principal directions" of the sequence $\{C_i^s\}_{1\le i\le K}$, which is assumed to satisfy

$$C_{i_l}^s \neq 0 \quad \text{iff} \ \ 1 \le l \le P \ . \tag{6}$$

The unknown parameters are the P coefficients $\{C_{i_l}^s\}_{1\le l\le P}$ and their respective locations $\{i_l\}_{1\le l\le P}$ for which one could search the maximum of the likelihood hypersurface. While a direct and naive approach of maximizing the likelihood function would generally maximize P, the solution provided by the MDL criterion attaches a regularizing penalty to lead to

$$\mathcal{L}(C_1, \ldots, C_K, \boldsymbol{\zeta}, P) = -\log p(C_1, \ldots, C_K \mid \boldsymbol{\zeta}) + \frac{1}{2}(2P)\log K \ . \tag{7}$$

Proposition 1. *If the functions φ_i are such that*

$$\forall u, \quad \varphi_i(u) \ge \varphi_i(0) \ ,$$

The P coefficients $C_{i_1}^s, \ldots, C_{i_P}^s$ which, based upon the MDL method, give the optimal coding length of x(t), are determined by the components C_i which satisfy the following inequality:

$$\varphi_i(C_i) > \log(K) + \varphi_i(0)$$

In the exponential-power case, the above inequality reduces to a hard thresholding policy:

$$\mid \mathcal{C}_i \mid > \gamma_i \, (\log(K))^{1/\beta_i} \ . \tag{8}$$

Furthermore, the resulting minimal code length is

$$\mathcal{L}^*(\mathcal{C}_1,\ldots,\mathcal{C}_K) = \sum_{i=1}^{K}\left(\min\left(\frac{|C_i|^{\beta_i}}{\gamma_i^{\beta_i}}, \log(K)\right) - \log\left(\frac{\beta_i}{2\gamma_i\Gamma(1/\beta_i)}\right)\right) . \quad (9)$$

This provides an interesting criterion for best basis search of signals embedded in (possibly non-Gaussian) noise.

4.2. Robust Representation

While the assumption that all the statistical characteristics of the noise are known may hold in few practical cases, its analytical tractability and appealing closed form results have been the root casue of its popularity. To bring us closer to practical scenarios, we follow Huber's approach by assuming that our noise distribution comes from a class of distributions $\mathcal{P}_\varepsilon = \{(1-\varepsilon)\Phi + \varepsilon G : G \in \mathcal{F}\}$, where Φ is the standard normal distribution, $\mathcal{F}$ is the set of all distribution functions, and $\varepsilon \in (0,1)$ is a known fraction of contamination.

Prior to determining the coding length, we have to identify the model in $\mathcal{P}_\varepsilon$ for our observed data. For a given underlying signal whose representation has a fixed number of components, the expected MDL is the entropy plus a constant independent of the prevailing distribution and of the estimator. In accordance with the minimax principle we seek the least favorable noise distribution and evaluate the MDL. This is tantamount to simultaneously maximizing the entropy over $\mathcal{P}_\varepsilon$ and minimizing over the set of all estimators $\mathcal{S}$. Interestingly, the least favorable distribution in $\mathcal{P}_\varepsilon$ which maximizes the entropy coincides with that which maximizes the asymptotic variance and derived by Huber [2]. For a standard normal density with variance σ^2 we have the following result:

Proposition 2. *The least favorable distribution $p_H(c)$ in $\mathcal{P}_\varepsilon$ which maximizes the entropy is*

$$p_H(c) = \begin{cases} (1-\varepsilon)\phi(a)e^{ac+a^2} & c \le -a \\ (1-\varepsilon)\phi(c) & |c| < a \\ (1-\varepsilon)\phi(a)e^{-ac+a^2} & a \le c \end{cases} \quad (10)$$

where ϕ is the standard univariate normal density and a is related to ε by the equation

$$2\left(\frac{\phi(a)}{a} - \phi(-a)\right) = \frac{\varepsilon}{1-\varepsilon}. \quad (11)$$

The density is normal in the center and Laplacian on the tails. On the other hand, the Maximum Likelihood estimator minimizes the entropy which then leads to the notion of MinMax description length.

Proposition 3. *Huber's distribution p_H together with the MLE based upon it, $\hat{\theta}_H$, result in a minimax MDL, i.e. they satisfy a saddle-point condition.*

Using an exactly similar approach as that of the Gaussian distribution, the minimax description length leads to the following thresholding rule:
Case 1 When $\log K > \frac{a^2}{2\sigma^2}$, the coefficient estimate is set to zero when

$$\frac{1}{\sigma^2}\left(-a\mid C_i\mid + \frac{a^2}{2}\right) + \log K > 0 \quad (12)$$

which implies that

$$\mid C_i \mid < \frac{a}{2} + \frac{\sigma^2}{a}\log K \quad (13)$$

Case 2 When $\log K < \frac{a^2}{2\sigma^2}$, the coefficient estimate is set to zero when

$$\frac{C_i^2}{2\sigma^2} < \log K \quad (14)$$

which implies that

$$\mid C_i \mid < \sigma\sqrt{2\log K} \quad (15)$$

This is the traditional threshold proposed by [1] and [3].

5. BAYESIAN APPROACH

The above approaches have been demonstrated to lead to good results in relatively moderate noise scenarios and have been successfully applied in a variety of settings. They are, however, based upon threshold values which present two drawbacks:

- They are directly dependent upon the noise variance without regard to the signal characteristics.
- They grow without bound with the data record length.

In some applications these shortcomings may greatly reduce the performance of the forementioned methods in retrieving the underlying signal. Fortunately, some prior information about the signal is often available, and it is thus natural to investigate its utility to regularize the estimation problem at hand.

Let the probability distributions of $\boldsymbol{C}_s$ and $\boldsymbol{C}_n$ be denoted respectively by f and p where the forms of functions f and p are assumed to be known. An estimate of $\boldsymbol{C}_s$ can be obtained by the following Maximum a Posteriori (MAP) estimate

$$\widehat{\boldsymbol{C}}_s = \arg\min_{\boldsymbol{C}_s}\left[-\log p(\boldsymbol{C}-\boldsymbol{C}_s) - \log f(\boldsymbol{C}_s)\right]. \quad (16)$$

By comparing this approach with the MDL approach, we see that the regularizing term now takes a more elaborate form allowing us to account for probabilistic prior information we may have about the signal of interest. Interestingly, it can be proved that many thresholding rules may be included within this framework [9]. For instance, if the noise components are i.i.d. Gaussian and the signal components are i.i.d., zero-mean and have a Laplacian distribution, a soft thresholding policy allows us to recover the signal. The threshold value is however independent of the data length K as it is equal to $\sqrt{2}\sigma^2/\sigma_s$ where σ^2 and σ_s^2 denote respectively the variances of C_i^n and C_i^s. To better take into account the expected sparsity of the components of the signal of interest, some more appropriate priors can be chosen. Gaussian mixtures constitute such valuable statistical models. For example, in the presence of i.i.d. Gaussian noise, the Bernoulli-Gaussian distribution (which is a degenerate Gaussian mixture) leads to an estimate which is a tradeoff between a Wiener and a thresholding estimator [6]. The estimated components then read

$$\widehat{C}_i^s = \begin{cases} \frac{\sigma_i^2}{\sigma_i^2+\sigma^2} C_i & \text{if } |C_i| > \chi_i \\ 0 & \text{otherwise} \end{cases} \qquad (17)$$

where σ_i^2 is the variance of the nonzero values of C_i^s and χ_i is a threshold value depending on σ^2, σ_i^2 and the mixture parameter. The interest of this Bayesian approach is shown in Fig. 1.

An important problem when dealing with this Bayesian approach is the estimation of the parameters of the model. Different algorithms can be envisaged, such as the Generalized Maximum Likelihood method or non-standard forms of the EM algorithm [5]. A fully Bayesian approach can also be adopted where priors are introduced on the parameters and one resorts to MCMC algorithmes in order to build an ergodic Markov chain whose equilibrium is the posterior distribution of interest [4]

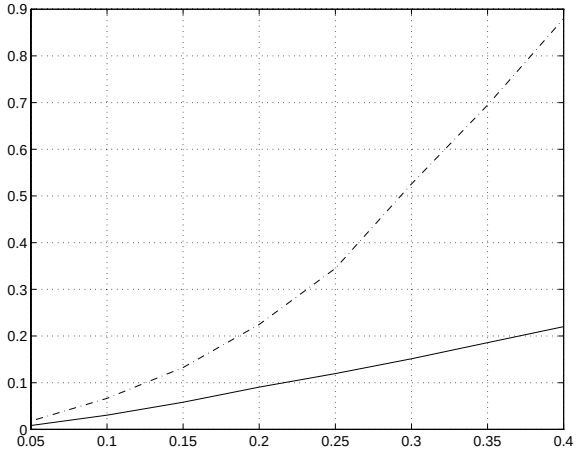

Figure 1: Comparison in terms of normalized mean square error of the MDL method (dashed line) and a Bayesian method (solid line) based on a B-G model as a function of the standard deviation of the noise (Doppler signal decomposed into wavelet packets).

6. REFERENCES

[1] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, Sept. 1994.

[2] P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Stat.*, 35:1753–1758, 1964.

[3] H. Krim and J.-C. Pesquet. On the statistics of best bases criteria. In A. Antoniadis, editor, *Wavelets and Statistics*, pages 192–207. Lecture Notes in Statistics, Springer Verlag, 1995.

[4] D. Leporini. Méthodes MCMC pour la décomposition en paquets d'ondelettes de signaux transitoires. In *Actes du colloque GRETSI*, pages 1455–1458, Grenoble, France, 15-19 septembre 1997.

[5] D. Leporini, J.-C. Pesquet, and H. Krim. Best basis representations based on prior statistical models with application to signal denoising. Tech. Rep., Laboratoire des Signaux et Systèmes, France, 1997.

[6] J.-C. Pesquet, H. Krim, D. Leporini, and E. Hamman. Bayesian approach to best basis selection. In *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, pages V.2634–V.2637, Atlanta, Georgia, USA, May 7-9 1996.

[7] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[8] I. Schick and H. Krim. Robust wavelet thresholding for noise suppression. In *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, volume V, pages 3421–3424, Munich, Germany, 1997.

[9] B. Vidakovic. Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. Internal Report, Duke University, 1995.