Robust Wavelet Thresholding for Noise Suppression

I.C. Schick† *and H. Krim*‡
†Network Engineering, BBN; DEAS, Harvard University, Cambridge, MA 02138
schick@bbn.com
‡Stochastic Systems Group, LIDS, MIT, Cambridge, MA 02139
ahk@mit.edu

Abstract

Approaches to wavelet-based denoising (or signal enhancement) have so far relied on the assumption of normally distributed perturbations. To relax this assumption, which is often violated in practice, we derive a robust wavelet thresholding technique based on the *Minimax* Description Length principle. We first determine the least favorable distribution in the ε-contaminated normal family as the member that maximizes the entropy. We show that this distribution and the best estimate based upon it, namely the Maximum Likelihood Estimate, constitute a saddle point. This results in a threshold that is more resistant to heavy-tailed noise, but for which the estimation error is still potentially unbounded. We address the practical case where the underlying signal is known to be bounded, and derive a two-sided thresholding technique that is resistant to outliers and has bounded error. We provide illustrative examples.

1 Introduction

The concept of "scale" has emerged in recent years as an important characteristic for signal analysis, particularly with the advent of wavelet theory.

Wavelets provide a powerful tool for non-linear filtering of signals contaminated by noise. Mallat and Hwang [1] have shown that effective noise suppression may be achieved by transforming the noisy signal into the wavelet domain, and preserving only the local maxima of the transform. Alternatively, a reconstruction that uses only the large-magnitude coefficients has been shown to approximate well the uncorrupted signal. In other words, noise suppression is achieved by thresholding the wavelet transform of the contaminated signal.

To choose the appropriate threshold, Donoho and Johnstone [2] have taken a minimax approach to characterizing the signal (rather than the disturbance, which they assume to be Gaussian). They derived a threshold that is approximately minimax (in the sense that its sample size dependence is of the same order as that of the true minimax): a coefficient C_i is excluded from the reconstruction if $\mid C_i \mid \leq \sigma\sqrt{2\log K}$, where σ is the standard deviation of the noise, and K is the length of the observation. Krim and Pesquet [3] have used Rissanen's Minimum Description Length (MDL) criterion [4] together with the assumption of normally distributed noise, and derived an identical threshold. Another feature that makes this threshold compelling is that it is asymptotically equivalent to the *maximum* of a sample of independent normally distributed variates [5], suggesting the intuitively pleasing interpretation that anything larger in magnitude is extremely unlikely to be pure noise and must therefore contain signal.

Nevertheless, the procedure remains *non-robust*. Although wavelets, thanks to their compactness and localization properties, do provide an unconditional basis for a large smoothness class of signals and offer a simple framework for nonlinear filtering, the procedures derived to date have been based upon the assumption of normality of the noise, and are therefore sensitive to outliers, i.e. to noise distributions whose tails are heavier than the Gaussian distribution. In this paper we adopt the minimax approach due to Huber [6] to derive a thresholding technique that is resistant to spurious observations.

2 Problem Statement

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

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

with $x(t) \in \mathrm{L}^2(\mathbb{R})$, and where $s(t)$ is a deterministic but unknown signal corrupted by the noise process $n(t)$.

In nonparametric estimation, the underlying signal model is often assumed to be induced by an orthonormal basis representation,

$$s(t) = \sum_i C_i^s \psi_i(t), \tag{2}$$

which in turn leads to the working model

$$C_i = C_i^s + C_i^n, \;\; i = 1, \cdots, K, \tag{3}$$

where the independent noise component has the same statistical properties as $n(t)$. Our problem is to recover/reconstruct $s(t)$ from the orthogonal transform of the observed process of $x(t)$. This can be achieved by using the MDL principle to determine which coefficients C_i contain signal information, and which are primarily noise and can therefore be left out of the reconstruction. However, since MDL is the maximum log likelihood minus a penalty term proportional to the number of parameters (i.e. the number of signal-containing coefficients C_i) in the model, it is strongly dependent on the distributional assumptions that characterize the noise. This paper addresses the derivation of a filtering technique that is resistant to heavy-tailed noise.

3 The Minimax Description Length (MMDL) Criterion

Following Huber [6], we assume that the noise distribution f is a (possibly) scaled version of a distribution belonging to the family of ε-contaminated normal 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 the known fraction of contamination. We cast our signal estimation problem as one of location parameter estimation, and thus assume the estimators to be in $\mathcal{S}$, the set of all integrable mappings from $\mathbb{R}$ to $\mathbb{R}$.

As in [7], we use the coding length of the observation in Equation 3 to determine the optimality of the signal estimate. For fixed model order, the expectation of the MDL criterion is the entropy (plus the penalty term which is independent of both the distribution and the functional form of the estimator). In accordance with the minimax principle, we seek the least favorable noise distribution and evaluate the MDL criterion for that distribution. In other words, we solve a minimax problem where the entropy is simultaneously maximized over all distributions in $\mathcal{P}_\varepsilon$ and minimized over all estimators in $\mathcal{S}$.

The least favorable distribution in $\mathcal{P}_\varepsilon$, i.e. the distribution that maximizes the entropy, is precisely the same as that found by Huber to maximize the asymptotic variance (or equivalently minimize the Fisher information). Generalizing Huber's distribution to perturbations from the zero-mean normal distribution with variance σ^2, we get:

Proposition 1. *The distribution $f_H \in \mathcal{P}_\varepsilon$ that maximizes the entropy is*

$$f_H(c) = \begin{cases} (1-\varepsilon)\phi_\sigma(a) e^{\frac{1}{\sigma^2}(ac+a^2)} & c \le -a \\ (1-\varepsilon)\phi_\sigma(c) & -a \le c \le a \\ (1-\varepsilon)\phi_\sigma(a) e^{\frac{1}{\sigma^2}(-ac+a^2)} & a \le c \end{cases} \tag{4}$$

where ϕ_σ is the normal density with variance σ^2 and a is related to ε by the equation

$$2\left(\frac{\phi_\sigma(a)}{a/\sigma^2} - \Phi_\sigma(-a)\right) = \frac{\varepsilon}{1-\varepsilon} \tag{5}$$

Proof: The proof that f_H maximizes the entropy is similar to that of Huber for the Fisher information. It can be shown that the negentropy $H(f) = \mathrm{E}[\log f]$ is a convex function of f, that $\mathcal{P}_\varepsilon$ is a convex set, and that if we define $f_\lambda = (1-\lambda)f_H + \lambda f$ for any $f \in \mathcal{P}_\varepsilon$ and any $\lambda \in (0,1)$, then

$$\frac{\partial}{\partial \lambda} H(f_\lambda) \mid_{\lambda=0} \; \ge \; 0 \tag{6}$$

establishing the desired result. ■

Thus, the least favorable distribution in $\mathcal{P}_\varepsilon$ is normal in the center and Laplacian ("double exponential") in the tails. The point where the density switches from Gaussian to Laplacian is a function of the fraction of contamination, larger fractions corresponding to smaller switching points and vice versa.

For a given distribution, the entropy is minimized by the Maximum Likelihood Estimate (MLE): since the negentropy is the expectation of the log likelihood, it follows that $\mathrm{E}[\log f(C; \hat{\theta}_{MLE}(C))]$ is maximum among all functions $\theta \in \mathcal{S}$. Thus, we obtain:

Proposition 2. *Huber's distribution f_H, together with the MLE based on it, $\hat{\theta}_H$, result in a Minimax Description Length, i.e. they satisfy a saddle-point condition.*

Proof: Using a theorem due to Verdú and Poor [8], this can be shown to be equivalent to proving that

$$\int f_\lambda(c;\theta) \log \frac{f_\lambda(c;\hat{\theta}_\lambda(c))}{f_\lambda(c;\hat{\theta}_H(c))} dc \;=\; o(\lambda) \tag{7}$$

where θ is the true value of the parameter and $\hat{\theta}_\lambda$ is the MLE based upon f_λ. Setting $\hat{\theta}_\lambda = \hat{\theta}_H + \Delta(\lambda)$, it can be shown that

$$\begin{aligned} \frac{f_\lambda(c;\hat{\theta}_\lambda(c))}{f_\lambda(c;\hat{\theta}_H(c))} \;&=\; 1 \;+\; \Delta(\lambda)\lambda\varepsilon \Big(g'(c;\hat{\theta}_H(c)) \\ &\quad -\; g'_H(c;\hat{\theta}_H(c)) \Big) \;+\; o(\Delta) \end{aligned} \tag{8}$$

where, since f and f_H are in $\mathcal{P}_\varepsilon$, they can be represented as $f = (1-\varepsilon)\phi + \varepsilon g$ and $f_H = (1-\varepsilon)\phi + \varepsilon g_H$, respectively. Furthermore, since

$$\begin{aligned} f'_\lambda(c;\hat{\theta}_\lambda(c)) \;&=\; f'_H(c;\hat{\theta}_\lambda(c)) \;+\; \lambda\varepsilon \Big(g'(c;\hat{\theta}_\lambda(c)) \\ &\quad -\; g'_H(c;\hat{\theta}_\lambda(c)) \Big) \end{aligned} \tag{9}$$

Figure 1: Plot of the negative exponent vs. C, showing quadratic and linear regions.

Case 1 When $\log K > \frac{a^2}{2\sigma^2}$, the coefficient estimate is set to zero if

$$\frac{1}{\sigma^2}\left(a \mid C_i \mid - \frac{a^2}{2}\right) < \log K \tag{11}$$

which implies that

$$\mid C_i \mid < \frac{a}{2} + \frac{\sigma^2}{a}\log K \tag{12}$$

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

$$\frac{1}{2\sigma^2} C_i^2 < \log K \tag{13}$$

which implies that

$$\mid C_i \mid < \sigma\sqrt{2\log K} \tag{14}$$

This is the threshold based on the assumption of normality, as proposed by [2] and [3].

A numerical example for Case 1 appears in Figure 2.

proximately 0dB SNR) in the top graph consists of two ramps with a discontinuity between them, onto which is superimposed noise from a Gaussian mixture with the following components: $\mathcal{N}(0, \sigma^2)$ with probability $1 - \varepsilon = 0.9$, and $\mathcal{N}(0, 9\sigma^2)$ with probability $\varepsilon = 0.1$. The second graph shows a reconstruction using coefficients over 4 resolutions achieved by the normal threshold; the third graph shows a similar reconstruction achieved by the proposed robust threshold. The robust threshold suppresses several instances of impulsive noise that the normal threshold misses.

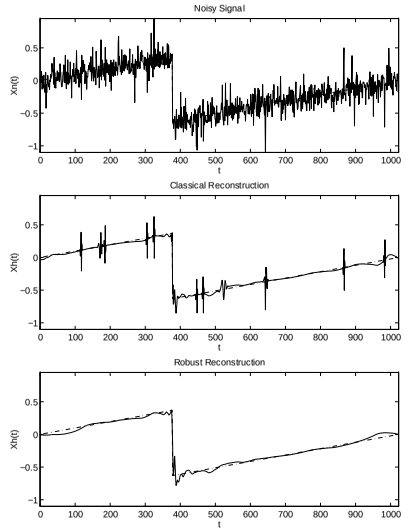

Figure 2: Noisy ramp signal, and its classical and robust reconstructions.

5 Constrained Minimax Thresholding

The procedure outlined above is somewhat more conservative than the traditional threshold in deciding whether or not a coefficient represents "primarily noise" and thus should be excluded from the reconstruction. It is therefore less vulnerable to noise that is heavier-tailed than normal. However, since it is based upon thresholding from below only, it does not result in an estimator whose error is bounded. Because the coefficients $\{C_i\}$ are assumed independent, $\hat{C}_i = C_i$ for each C_i that exceeds the threshold, and there is nothing to counterbalance the influence of a single spurious data point. Thus the estimation error can increase without bound.

In this section, we describe a method that involves

Figure 3: Absolute reconstruction error vs. outlier amplitude for three thresholds.

We proposed a *Minimax* Description Length (MMDL) principle as the criterion of choice for thresholding wavelet coefficients. We determined the least favorable distribution in the ε-contaminated normal family, which we used to derive a robust threshold that is resistant to outliers. We further assumed that the true signal has bounded amplitude and derived a thresholding technique from above and below that results in bounded estimation error. These robust thresholds yield denoising methods that are less sensitive to heavy-tailed noise than the traditional threshold based on the assumption of normality.

Acknowledgement

The work of second author was supported in part by the Army Research Office (DAAL-03-92-G-115), and the Air Force Office of Scientific Research (F49620-95-1-0083 and BU GC12391NGD). We are also grateful to D. Tucker for help with the numerical examples.

References

[1] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. on Information Theory*, vol. IT-38, pp. 617–643, 1992.

[2] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *JASA*, vol. 90, pp. 1200–1223, 1995.

[3] H. Krim and J.-C. Pesquet, "On the Statistics of Best Bases Criteria", vol. *Wavelets in Statistics* of *Lecture Notes in Statistics*, pp. 193–207. Springer-Verlag, 1995.

[4] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[5] B. Gnedenko, "Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire," *Ann. Mathematics*, vol. 44, pp. 423–453, 1943.

[6] P. Huber, "Robust estimation of a location parameter," *Ann. Math. Stat.*, vol. 35, pp. 1753–1758, 1964.

[7] H. Krim, S. Mallat, D. Donoho, and A. Willsky, "Best basis algorithm for signal enhancement," in *ICASSP'95*, (Detroit, MI), IEEE, May 1995.

[8] S. Verdú and H. V. Poor, "On minimax robustness: A general approach and applications," *IEEE Trans. on Information Theory*, vol. IT-30,