Minimax Description Length for Signal Denoising and Optimized Representation

Hamid Krim, *Senior Member, IEEE,* and Irvin C. Schick

***Abstract*—Approaches to wavelet-based denoising (or signal enhancement) have generally 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 (MMDL) 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, together constitute a saddle point. The MMDL approach results in a thresholding scheme that is resistant to heavy tailed noise. We further extend this framework and propose a novel approach to selecting an adapted or best basis (BB) that results in optimal signal reconstruction. Finally, 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.**

***Index Terms*—Best basis, denoising, MDL, minimax, robust.**

I. Introduction

THE concept of scale has emerged in recent years as an important characteristic for signal analysis. Wavelet theory has played a particularly important role in this regard, due to the fact that the basis functions are well suited to the analysis of local (whether in the spatial or temporal sense) scale phenomena. This property also endows wavelets with a remarkable aptitude for denoising by means of a simple nonlinear thresholding filter. Mallat and Hwang [1] first showed 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. A wavelet reconstruction that uses only large-magnitude coefficients has also 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 have 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 $|C_i| \leq \sigma\sqrt{2 \log N}$, where σ is the standard deviation of the noise, and N is the length of the observation. Krim and Pesquet [3] have given an alternative derivation for this threshold, using Rissanen's minimum description length (MDL) criterion [4] and the assumption of normally distributed noise. 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.

However it is derived, this procedure is nonrobust. Although wavelets, thanks to their compactness and localization properties, do provide an unconditional basis for a large smoothness class of signals and do offer a simple framework for nonlinear filtering, many of the procedures derived to date have been based upon the assumption of normal noise, and are therefore sensitive to outliers, i.e., to noise distributions whose tails are heavier than the Gaussian distribution.

Neumann and collaborators [6], [7] have addressed the question of nonnormality by having recourse to asymptotics. Bruce *et al.* [8] pass the wavelet coefficients through a median filter which, though effective in suppressing outliers, significantly smooths out the reconstruction. In this paper, we adopt the minimax approach due to Huber [9] to derive a filtering technique that is resistant to spurious observations. We introduce a notion of robust description length which leads to a nonlinear thresholding technique that performs remarkably well and enjoys intuitive appeal.

Denoising may be interpreted as a quest for parsimony in the representation of a process. Wavelets have the property of energy compaction, particularly at or near singularities. A given wavelet basis is not, however, universally optimal for all processes; this difficulty can be circumvented by making the representation adaptive, that is by searching in a dictionary of bases for the optimum basis. The notion of optimality, however, may take a variety of forms. A number of optimization techniques have been proposed since the entropy criterion of Coifman and Wickerhauser [10], including error criteria [11]–[13] as well as energy compaction measures [14]–[17], all of which are equivalent to compressing the representation of a process in the dual domain. In this paper, we use description (or coding) length as a measure of the parsimony of the candidate representation, and derive a best basis search criterion that minimizes it over all the bases in the dictionary.

Manuscript received February 15, 1998; revised October 15, 1998. The work of H. Krim was supported in part under Grants AFOSR-F49620-98-1-0190 and ONR-MURI-Grant-WUHT-072298-S2, and by the North Carolina State College of Engineering. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, MIT, Cambridge, MA, August 16–21, 1998.

H. Krim is with ECE/CACC, North Carolina State University, Raleigh, NC 27695-7914 USA (e-mail: ahk@eos.ncsu.edu).

I. C. Schick is with Division of Engineering and Applied Sciences, Harvard University and with GTE Internetworking, Cambridge, MA 02138 USA (e-mail: ischick@bbn.com).

Publisher Item Identifier S 0018-9448(99)02264-6.

In the section that follows, we give a concise statement of the problem. In Section III, we introduce the minimax description length (MMDL) criterion and derive the least favorable distribution and corresponding best estimator. In Section IV, we obtain an outlier-resistant thresholding technique based on this principle. In Section V, we propose a new criterion based on MMDL for selecting an optimal basis from a dictionary. Some numerical examples appear in Section VI. Finally, we provide some concluding remarks in Section VII.

II. Problem Statement

Consider the additive noise model

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

where $s(t)$ is an unknown but deterministic signal corrupted by the zero-mean noise process $n(t)$, and $x(t)$ is the observed, i.e., noisy, signal. The objective is to recover the signal $\{s(t)\}$ based on the observations $\{x(t)\}$.

The underlying signal is modeled with an orthonormal basis representation

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

and, similarly, the noise is represented as

$$n(t) = \sum_i C_i^n \psi_i(t).$$

By linearity, the observed signal can also be represented in the same fashion, its coefficients given by

$$C_i^x = C_i^s + C_i^n.$$

A key assumption we make is that for certain values of i, $C_i^s = 0$; in other words, the corresponding observation coefficients C_i^x represent "pure noise," rather than signal corrupted by noise. Though this is not necessarily true in an exact sense, it is, as shown by Krim and Pesquet [3], a reasonable assumption in view of the spectral and structural differences between the underlying signal $s(t)$ and the noise $n(t)$ across scales. Given this assumption, wavelet-based denoising consists in determining which wavelet coefficients represent primarily signal, and which mostly capture noise. Using Rissanen's information-theoretic approach in this context results in a simple thresholding algorithm, whereby coefficients that exceed a certain value are retained, while those that fall below that value are set to zero, or "truncated." In other words, the MDL criterion is utilized for resolving the tradeoff between model complexity (each retained coefficient increases the number of model parameters) and goodness-of-fit (each truncated coefficient decreases the fit between the received—i.e., noisy—signal and its reconstruction).

III. The Minimax Description Length Criterion

Wavelet thresholding is essentially an order estimation problem, one of balancing model accuracy against overfitting, of capturing as much of the "signal" as possible, while leaving out as much of the "noise" as possible. This is a notoriously difficult problem, for which many techniques have been proposed, but no definitive solution exists. Our approach here does not necessarily break new ground, but it does, we believe, constitute an interesting framework which results in some intuitively sensible and mathematically tractable techniques, by bringing together Rissanen's work on stochastic complexity and coding length [4], [18], and Huber's work on minimax statistical robustness [9], [19].

We follow Rissanen in adopting data description length as the criterion of choice for quantifying the tradeoff between goodness-of-fit and model complexity. In other words, we seek the data representation that results in the shortest encoding of both observations and constraints. This approach involves the construction of probabilistic model classes, and for that, we turn to Huber [9]. Perfect information, of course, would allow for perfect models, but such information is in short supply, and too often this obvious fact is simply ignored for the sake of analytical expediency. Here we follow Huber in making the limits of our knowledge explicit, and adopt a minimax approach. We seek the shortest encoding under the most pessimistic assumptions, thus obtaining an estimator that, while potentially suboptimal under known conditions, provides significant safeguards against catastrophic breakdown when modeling assumptions prove to be incorrect.

Rissanen's MDL criterion is essentially the difference between the likelihood, which measures deviation from the model, and a penalty term that quantifies the amount of information that must be encoded to represent the model. The representation of the data is viewed as "correct" when this difference is minimized, thus indicating that it achieves the right balance between fit and structure, data and model. In most applications, the data are assumed to be conditionally normal, and the likelihood is therefore Gaussian. Here, we follow Huber in assuming that the noise distribution f is a (possibly) scaled version of an unknown member of the family of ε-contaminated normal distributions

$$\mathcal{P}_\varepsilon = \{(1-\varepsilon)\Phi + \varepsilon G \colon G \in \mathcal{F}\}$$

where Φ is the standard normal distribution, $\mathcal{F}$ is the set of all suitably smooth distribution functions, and $\varepsilon \in (0, 1)$ is the known fraction of contamination. (Sensitivity to the parameter ε is discussed in Section VI.) 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 $\mathbb{R}$.

To optimize the location parameter estimator, Huber used aymptotic variance as a measure of performance, and exploited the relationship (through the Cramér–Rao lower bound) between this quantity and the Fisher information. We note an analogous relationship between description length and entropy: specifically, for fixed model order, the expectation of the MDL criterion is the entropy plus a penalty term that 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 maximized over all distributions in $\mathcal{P}_\varepsilon$, and the description length is

minimized over all estimators in $\mathcal{S}$. The saddle point, provided one can be shown to exist, yields a minimax robust version of MDL, which we call the minimax description length (MMDL) criterion.

The least favorable distribution in $\mathcal{P}_\varepsilon$, that is, 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): it is Gaussian in the center and Laplacian ("double exponential") in the tails, and switches from one to the other at a point whose value depends on the fraction of contamination ε, larger fractions corresponding to smaller switching points and vice versa.

Proposition 1: The distribution $f_H \in \mathcal{P}_\varepsilon$ that minimizes the negentropy is

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

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

$$2\left(\frac{\phi_\sigma(a)}{a/\sigma^2} - \Phi_\sigma(-a)\right) = \frac{\varepsilon}{1-\varepsilon}. \tag{3}$$

Proof: See Appendix A. □

This distribution leads to a solution for our minimax problem: for any given distribution, the MDL criterion is minimized by the maximum-likelihood estimate (MLE): since the negentropy is the expectation of the log likelihood, it follows that $E[\log f(C, \hat{\theta}_{\mathrm{MLE}}(C))]$ is maximum among all estimators $\theta \in \mathcal{S}$. Thus we have:

Proposition 2: Huber's distribution f_H, together with the MLE based on it, $\hat{\theta}_H$, result in a minimax description length, that is, they satisfy a saddle-point condition.

Proof: See Appendix B. □

IV. Minimax Thresholding

Let the set of wavelet coefficients obtained from the observed signal be denoted by $\mathcal{C}^N = \{C_1^x, C_2^x, \cdots, C_N^x\}$. Let exactly K of these coefficients contain signal information, while the remainder only contain noise. If necessary, we reindex these coefficients so that

$$C_i^x = \begin{cases} C_i^s + C_i^n, & i = 1, 2, \cdots, K \\ C_i^n, & \text{otherwise.} \end{cases} \tag{4}$$

By assumption, the set of noise coefficients $\{C_i^n\}$ is a sample of independent and identically distributed (i.i.d.) random variates drawn from Huber's distribution f_H. It follows, by (4), that the observed coefficients C_i^x obey the distribution $f_H(c - C_i^s)$ when $i = 1, 2, \cdots, K$, and $f_H(c)$ otherwise. Thus the likelihood function is given by

$$\ell(\mathcal{C}^N; K) = \prod_{i \le K} f_H(C_i^x - C_i^s) \prod_{i > K} f_H(C_i^x).$$

Since f_H is symmetric and unimodal with a maximum at the origin, the above expression is maximized (with respect to the signal coefficient estimates $\{\hat{C}_i^s\}$) by setting

$$\hat{C}_i^s = C_i^x$$

for $i = 1, 2, \cdots, K$. It follows that the maximized likelihood (given K) is

$$\ell^*(\mathcal{C}^N; K) = \prod_{i \le K} f_H(0) \prod_{i > K} f_H(C_i^x).$$

To write down the associated description length, we identify within the N coefficients $2K$ free parameters corresponding to the assumed K signal-bearing coefficients together with their positions. The problem is hence reduced to choosing the optimal value of K, in the sense of minimizing the MDL criterion

$$\begin{aligned} \mathcal{L}(\mathcal{C}^N; K) &= -\log \ell^*(\mathcal{C}^N; K) + \tfrac{1}{2}(2K)\log N \\ &= -\sum_{i \le K} \log f_H(0) - \sum_{i > K} \log f_H(C_i^x) + K \log N. \end{aligned} \tag{5}$$

Neglecting terms independent of K, this is equivalent to minimizing

$$\tilde{\mathcal{L}}(\mathcal{C}^N; K) = \frac{1}{2\sigma^2} \sum_{i > K} \eta(C_i^x) + K \log N$$

where

$$\eta(c) = \begin{cases} c^2, & \text{if } |c| < a \\ a|c| - a^2, & \text{otherwise} \end{cases}$$

is proportional to the exponent in Huber's distribution f_H. We now show that this can be achieved by a thresholding scheme.

Proposition 3:

1) When $\log N > a^2/2\sigma^2$, the coefficient $|C_i^x|$ is truncated if
$$|C_i^x| < \frac{a}{2} + \frac{\sigma^2}{a} \log N.$$
2) When $\log N \le a^2/2\sigma^2$, the coefficient $|C_i^x|$ is truncated if
$$|C_i^x| < \sigma\sqrt{2 \log N}.$$

Proof: See Appendix C. □

In practice, the noise variance and mixture parameter often have to be estimated; in this, however, our approach is no different than others, and there are a variety of techniques to address this problem. (For example, see [20] and [21].)

To illustrate the two distinct cases in the above Proposition, consider Fig. 1, a plot of the exponent in Huber's distribution $f_H(c)$ for $c > 0$. The graph clearly indicates a linear region and a quadratic region, respectively, corresponding to Cases 1 and 2. Note that Case 2 corresponds to the assumption of normality, and is identical to the threshold proposed by [3] and [2].

When $\sigma^2 \to 0$, the thresholding scheme reduces to Case 2, and C_i^x is never truncated; since this represents the no-noise case, it stands to reason that all coefficients should be retained in the reconstruction. On the other hand, for large σ^2, the thresholding scheme reduces to Case 1, which is the more conservative threshold. For $\sigma^2 \to \infty$, the signal-to-noise ratio

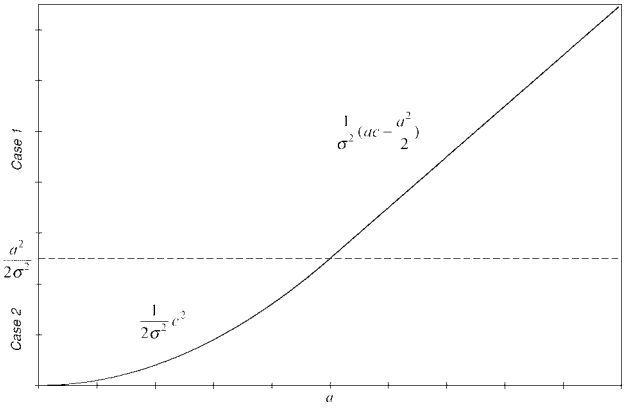

Fig. 1. Plot of the negative exponent of $f_H(c)$ against c, indicating the two regions corresponding to Cases 1 and 2 of Proposition 3.

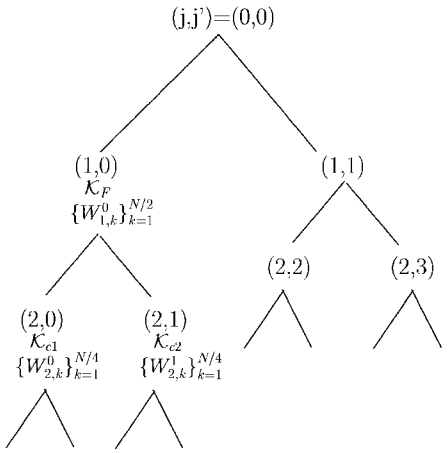

Fig. 2. Tree structure of wavelet packet bases.

(SNR) becomes zero and the best one can do is to estimate the signal as identically zero.

Similarly, when $a \to \infty$, the noise distribution becomes purely Gaussian, and the thresholding scheme reduces to Case 2, as expected. On the other hand, when $a \to 0$, the noise distribution becomes purely Laplacian, and the thresholding scheme reduces to Case 1.

Finally, when $N \to 1$, the thresholding scheme reduces to Case 2, suggesting that outliers are unlikely to occur in a small sample, and it is, therefore, more reasonable to assume purely Gaussian noise. On the other hand, for large N, the thresholding scheme reduces to Case 1, since outliers are highly likely to occur in a large sample.

V. Robust Basis Selection

The signal denoising method discussed thus far focuses on achieving robustness under the worst case scenario, that is, in the presence of noise having the least favorable distribution in the sense of maximizing entropy. It only makes a weak assumption about the noise distribution, namely, that it belongs to the family of ε-contaminated Gaussian distributions. It also uses minimal prior knowledge about the underlying signal, namely, that it is deterministic and unknown. In these respects, it is quite general.

At the same time, using a single basis to represent the entire signal implicitly assumes a certain amount of stationarity, in the sense that if the properties of the signal (e.g., its smoothness or oscillatory behavior) were to vary significantly over time, the chosen basis may be unable to represent it with uniform efficiency and fidelity. Thus further generality may be achieved by making the basis adaptive, that is, by allowing it to change over time so as to represent a nonstationary signal in the best possible way. Adaptivity of the basis thus enriches the class of signals that can be analyzed, while ensuring additional breakdown robustness [22].

Adapting a basis to a process may be efficiently achieved by pruning a tree-structured dictionary of bases using a properly constructed additive cost measure [10]. In Fig. 2, each row (i.e., each layer in the tree) represents a particular scale, the resolution going from coarse to fine as we proceeds downward. Within each layer, different nodes correspond to different segments of the signal or its spectrum. A cost is assigned to each node, and a systematic and highly efficient comparison is carried out from bottom up, effectively pruning an overcomplete set of functions down to a basis. The construction of such measures remains a topic of active research [15], centered largely on the notion of parsimony. The synergy between an adapted representation (or best approximation) of a signal and noise removal is of central importance, as is further discussed below.

For a signal with complex features, such as a piecewise-polynomial function belonging to some smoothness class $\mathcal{M}$, adapted wavelet bases offer greater flexibility than the Karhunen–Loève (principal components) expansion, and various criteria have been proposed to search for a basis which results in the most parsimonious signal representation [11], [12], [24], [25]. In this section, we again turn to description length for quantifying parsimony, and use it to search for the representation best adapted to the signal, while achieving robustness against heavy-tailed noise.

Formally, we assume a dictionary of bases $\mathcal{D}=\{\mathcal{B}^\pi|\pi\in\Pi\}$, and wish to select a basis $\mathcal{B}^\pi = \{W_k^\pi\}_{k=1,\cdots,N} \in \mathcal{D}$ for the observation space, where π is a particular partition of the unit interval. As shown in Fig. 2, this basis can be written as

$$\mathcal{B}^\pi = \bigcup_{\{(j,j'):\, I_{j,j'}\in\pi\}} \mathcal{B}_j^{j'}$$

where $\{j, j'\}$ specifies the spectral partition j' at scale j (with $j \in \{0, \cdots, J\}$ and $j' \in \{0, \cdots, 2^j - 1\}$), and

$$I_{j,j'} = [2^{-j}j', 2^{-j}(j'+1)[$$

is an interval in the partition $\pi \in \Pi$ of $[0, 1[$. The efficiency of pruning the tree is due to the latter's binary structure, which yields the best partition or best basis (BB) in $\mathcal{D}$. Each basis $\mathcal{B}_j^{j'} = \{W_{j,k}^{j'}\}_{k=1,\cdots,N/2^j}$ is then an orthonormal basis of a vector subspace of $\ell^2(\{1, \cdots, N\})$. Each basis function at scale j is the result of a linear transformation of the parent basis, which implies the following important property:

$$\text{Span}\left\{\mathcal{B}_j^{j'}\right\} = \text{Span}\left\{\mathcal{B}_{j+1}^{2j'}\right\} \overset{\perp}{\oplus} \text{Span}\left\{\mathcal{B}_{j+1}^{2j'+1}\right\} \qquad (6)$$

where $\overset{\perp}{\oplus}$ denotes the direct sum of orthogonal subspaces. The three subspaces in (6) correspond to a triplet consisting of one parent and two child nodes on the tree, each with an associated cost $\mathcal{K}(\cdot)$ as shown in Fig. 2.

This framework serves in a bottom-up comparison of the costs of children versus those of parents, and ultimately in pruning the tree. For example, the decision whether to select the local bases of nodes $(2, 0)$ and $(2, 1)$ or that of the parent node $(1, 0)$ is determined by whether or not $\mathcal{K}_{c1}+\mathcal{K}_{c2} < \mathcal{K}_F$. This optimization yields a minimal tree structure whose end leaves form an adapted orthonormal basis among a collection of orthonormal bases $\mathcal{D}$.

With no loss of generality, we restrict our discussion to wavelet packet bases [10] as the dictionary of choice; this affords an adaptive partitioning of the frequency axis. An alternative orthonormal basis corresponds to a division of the time axis into intervals of varying sizes, and is referred to as a local cosine basis [26].

For a discrete signal of length N, it can be shown that a tree of wavepacket or local cosine bases includes more than 2^N bases, and that the signal expansion in these bases is computed with algorithms that require $O(N \log N)$ operations, because these bases share many vectors. Wickerhauser and Coifman [27] have proved that for any signal x and any isotonic functional $\mathcal{K}(x)$, finding the basis $\mathcal{B}^{\pi_0}$ that minimizes, over all bases, an additive cost function of the form

$$\text{Cost}\,(x, \mathcal{B}^{\pi}) = \sum_{i=1}^{N} \mathcal{K}(|C^x_{i_\pi}|^2)$$

(where i_π indicates the πth basis indices) requires $O(K \log K)$ operations. Since our goal is to obtain the basis having the minimax coding length, our cost $\mathcal{K}(x)$ consists of description lengths (with respect to the least favorable distribution) over all the subspaces on the tree. Recall that description length is given in terms of the log-likelihood function

$$\mathcal{L}(\mathcal{C}^N, \sigma^2, K) = -\log \ell(\{C^x_i\}|\sigma^2, K) + K \log N.$$

For a given basis in some subspace, the coding length of the corresponding coefficients may be evaluated by first noting that there are three distinct possibilities, each contributing differently to the complexity.

1) Signal is present in K signal-bearing coefficients, for which the likelihood terms are of the form $f_H(0) = \frac{1-\varepsilon}{\sqrt{2\pi}\sigma}$.
2) Only noise is present, and the coefficients are below the switching point in the least favorable distribution, i.e., $|C^x_i| \le a$. There are K' such coefficients, for which the likelihood terms are of the form

$$f_H(C^x_i) = \phi_\sigma(C^x_i) = \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} e^{-\frac{(C^x_i)^2}{2\sigma^2}}.$$

3) Only noise is present and the coefficients are above the switching point in the least favorable distribution, i.e., $|C^x_i| > a$. There are K'' such coefficients, for which the likelihood terms are of the form

$$f_H(C^x_i) = f_L(C^x_i) = \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} e^{-\frac{a}{2\sigma^2}} e^{\frac{-a|C^x_i|+a^2}{\sigma^2}}.$$

(Here, f_L denotes the Laplacian distribution, and its dependence on σ is implicit.)

Rearranging indices if necessary, and by independence, this leads to the following expression for description length:

$$\begin{aligned}\mathcal{L}(\mathcal{C}^N, \sigma^2, K) &= -\log \prod_{i=1}^{K} f_H(0) - \log \prod_{i=1}^{K'} \phi_\sigma(\mathcal{C}^x_{K+i}) \\ &\quad - \log \prod_{i=1}^{K''} f_L(\mathcal{C}^x_{K+K'+i}) + K \log N' \\ &= -N' \log \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2} \sum_{i=1}^{K'} (C^x_{K+i})^2 \\ &\quad - K'' \frac{a^2}{2\sigma^2} + \frac{a}{\sigma^2} \sum_{i=1}^{K''} |C^x_{K+K'+i}| \\ &\quad + K \log N'\end{aligned}$$

where $N' = K + K' + K''$ is the length of sequence of coefficients for the local basis. The last term is the usual penalty term, which represents the coding length for the model description.

The costs (or description lengths) of each pair of nodes are summed and compared to that of their common parent node in order to determine the surviving scale among the triplet. The resulting basis corresponds to the minimax description length.

VI. Numerical Experiments

In the examples that follow, we demonstrate the performance of the robust thresholding procedure described herein, and compare it with that of the thresholding scheme based upon the assumption of normally distributed noise.

A. Example 1—Robust Thresholding

Using WAVELAB,[1] we synthesized a broken ramp signal of length $N = 1024$. This signal admits an efficient representation in a wavelet basis, i.e., one with very few nonzero coefficients. The noise was additive and i.i.d., obeying an $N(0, \sigma^2)$ distribution contaminated by a fraction $\varepsilon = 10\%$ of white Gaussian noise with distribution $N(0, 9\sigma^2)$. The overall SNR was maintained at 10 dB (see Fig. 3, top).

We constructed two estimators. The first was based on the assumption of purely Gaussian noise, i.e., no contamination, and used the thresholding scheme due to [2] and [3]. The second was the MMDL robust estimator described here in Section IV. The reconstructions based on each estimator appear in Fig. 3. The reconstruction based upon the assumption of purely Gaussian noise was highly susceptible to outliers; the robust reconstruction was not.

[1] Available from the Stanford Statistics Department, courtesy of D. L. Donoho and I. M. Johnstone.

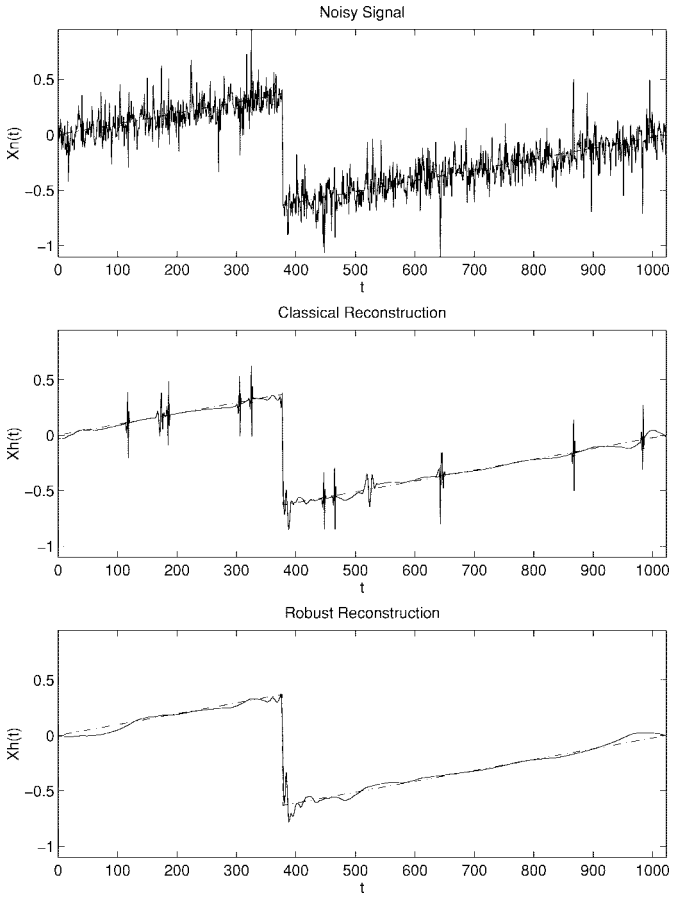

Fig. 3. Noisy ramp signal, and its Gaussian and robust reconstructions.

We also carried out a Monte Carlo simulation to evaluate the reconstruction performance over a range of SNR's. At each value of the SNR, 100 experiments were conducted, and the cumulative reconstruction error is displayed in Fig. 4. The robust estimator uniformly outperforms the classic estimator in both L_1 and L_2 errors over a wide range of SNR's. Furthermore, the performances of the Gaussian and robust estimators become indistinguishable at high SNR's, i.e., small noise variance, showing that robustness does not come at the cost of reduced efficiency.

Bounding the Reconstruction Error: Although the reconstruction error is much improved by the robust estimator, it is still potentially unbounded. The underlying assumption throughout this paper has been that a wavelet basis was appropriately chosen for the smoothness class of the signal of interest. Because of the compactness of wavelets, however, unbounded noise will still result in unbounded reconstruction error, a property that may be considered undesirable. This problem may be circumvented by making the assumption that the signal has bounded energy; in that case, at least two alternatives present themselves.

1) In many, and perhaps most, cases of practical interest, the signal is known to be bounded, and prior knowledge of the physical properties of the signal may be used to determine the $\|\cdot\|_\infty$ of the sequence of signal wavelet coefficients $\{C_i^s\}$. This information may be used to truncate observed coefficients $\{C_i^x\}$ not only below, as discussed earlier, but also above.
2) In the absence of such prior knowledge, it may still be possible to bound the reconstruction error through an adaptive supremum–secondary thresholding scheme based upon some representation criterion, e.g., entropy.

The first of these approaches is illustrated in Fig. 5, which uses the following modified thresholding: let $\alpha > 0$ be an upper bound on the magnitude of the signal coefficients; then

$$\tilde{C}_i^s = \begin{cases} 0, & \text{if } |C_i^x| \le \dfrac{a}{2} + \dfrac{\sigma^2}{a}\log N \\ \hat{C}_i^s = C_i^x, & \text{if } \dfrac{a}{2} + \dfrac{\sigma^2}{a}\log K \le |C_i^x| \le \alpha \\ \alpha\,\mathrm{sgn}\,(C_i^x), & \text{if } \alpha \le |C_i^x| \end{cases}$$

provided that $\log N > a^2/2\sigma^2$ and $\alpha > (a/2) + (\sigma^2/a)\log N$. The graph shows that although the robust estimator's reconstruction error initially grows more slowly than that of the Gaussian estimator, the two errors soon converge as the variance of the outliers grows. The reconstruction error for the bounded-error estimator, however, levels off after a while.

Sensitivity Analysis for the Fraction of Contamination: Although such crucial assumptions as the normality of the noise or exact knowledge of its variance σ^2 usually go unremarked, Huber's minimax approach to robustness is sometimes taken to task because it assumes that the fraction of contamination ε is known. To allay such concerns, we now show that the sensitivity of the robust estimator to changes in the assumed value of ε is far lower than that of the Gaussian estimator to changes in the true fraction of contamination.

Fig. 6 shows the total reconstruction error as a function of variation in the true fraction of contamination ε. In other words, an abscissa of 0 corresponds to an assumed fraction of contamination equal to the true fraction; larger abscissas correspond to outliers of larger magnitude than assumed by the robust estimator, and vice versa. Obviously, the Gaussian estimator assumes zero contamination throughout. The figure shows that the reconstruction error for the Gaussian estimator grows very rapidly as the true fraction of contamination increases, whereas that of the robust estimator is nearly flat over a broad range. This should not come as a surprise: outliers are, by definition, rare events, and for a localized procedure such as wavelet expansion, the precise frequency of outliers is much less important than their existence at all.

B. Example 2—Robust Best Basis Search

The example above assumed a fixed wavelet basis. As discussed in Section V, however, highly nonstationary signals can be represented most efficiently by using an adaptive basis that automatically chooses resolutions as the behavior of the signal varies over time. In this example, we turn our attention to the robust best basis algorithm introduced in Section V.

The example we analyze is once again the broken ramp signal, corrupted by noise distributed according to the Gaussian mixture discussed in Example 1. The original and noisy signals appear in Fig. 7, together with the reconstructions and reconstruction errors obtained by the Gaussian and robust estimators, respectively. Although the discontinuity in the signal is captured quite well by both estimators, the outliers in the noise cause the Gaussian best basis search to retain

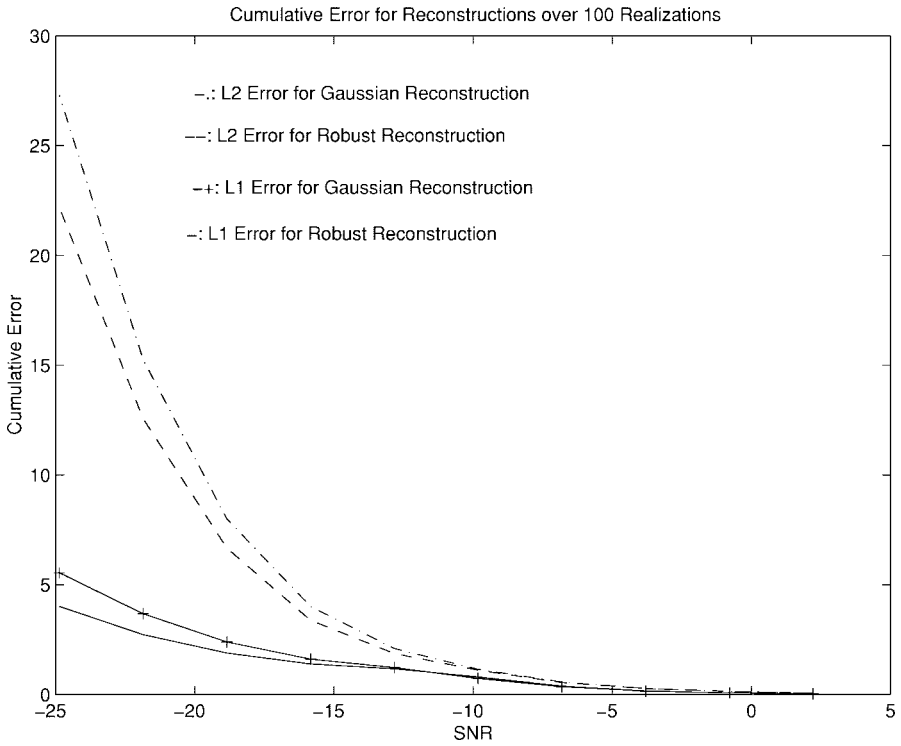

Fig. 4. L_1 and L_2 error performance versus SNR, for the Gaussian and robust estimators.

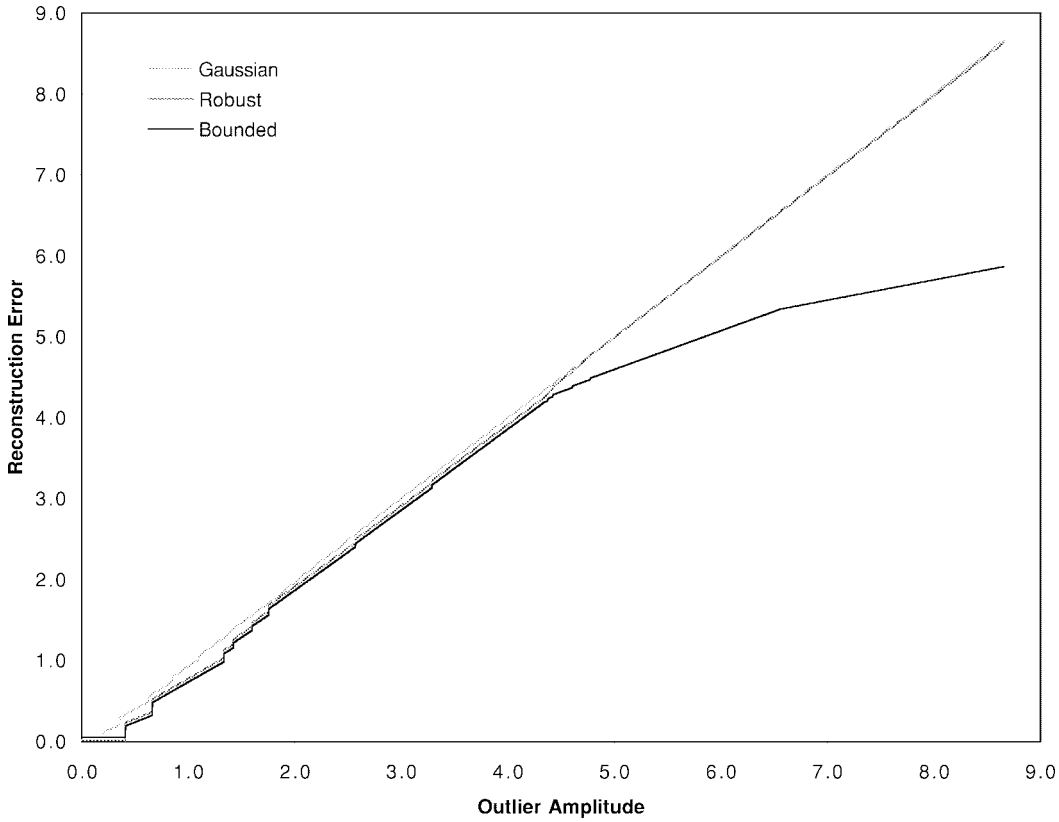

Fig. 5. Absolute reconstruction error versus outlier amplitude for three thresholding schemes.

too many high-resolution coefficients, resulting in a significant amount of background noise. By contrast, the robust estimator recognizes the outliers for what they are, and is thus able to capture the smoothness of the ramps without sacrificing fidelity to the sharp discontinuity. In other words, it chooses coarse basis functions in the regions corresponding to the linear segments of the original signal, and fine basis functions in the neighborhood of the discontinuity itself.

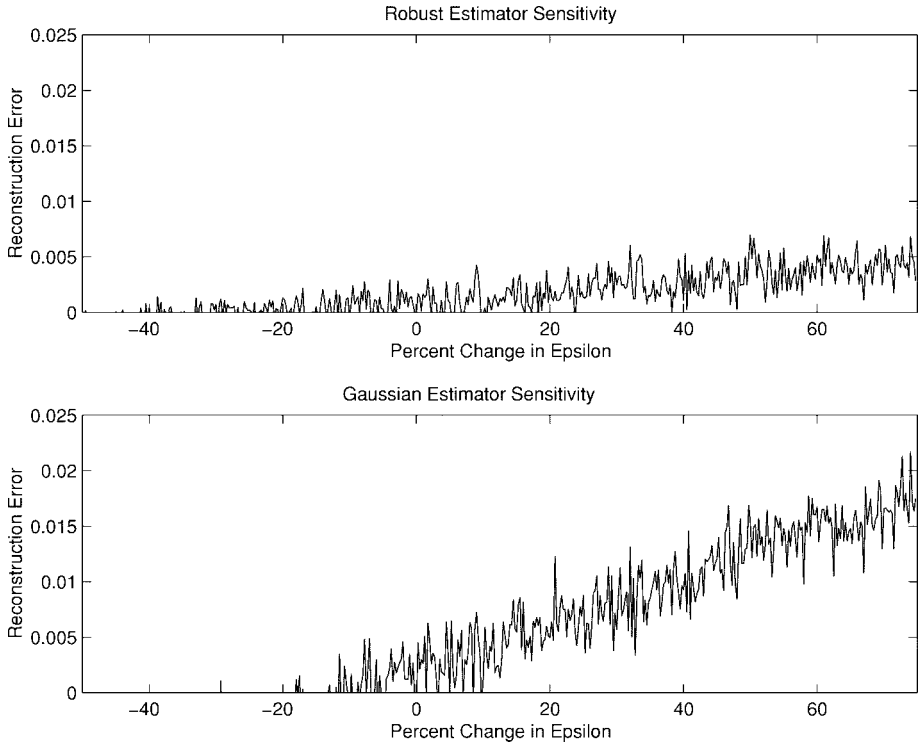

Fig. 6. Error stability versus variation in the mixture parameter ε.

C. Example 3—Speech Signal

Examples 1 and 2 use synthetic data; in this final example, we analyze speech, a type of real-world signal well known for its nonstationarity. In Fig. 8 we show the original signal, a noisy observation corrupted by additive Gaussian noise $(\sigma^2 = 0.3)$ contaminated by outliers $(\varepsilon = 0.1)$, along with the reconstructions and reconstruction errors obtained respectively by the Gaussian and robust estimators. Although the signal changes significantly over time, the BB reconstructions are able to capture both its high- and low-frequency components. Once again, the robust reconstruction does better than the reconstruction that assumes Gaussian noise.

VII. Conclusion

We have proposed the 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 to be Huber's distribution, which we used to derive a robust thresholding scheme that is resistant to outliers. We also applied this scheme to a best basis (BB) search, and demonstrated by example that it performs well also when the signal is highly nonstationary. Finally, we further assumed that the true signal has bounded amplitude and derived a two-sided thresholding scheme that results in bounded estimation error. In all cases, the robust estimator based on this minimax thresholding technique outperforms the estimator based on a Gaussian assumption when the noise contains outliers, but reduces to the Gaussian estimator when the noise is purely normal.

Appendix A
Proof of Proposition 1

To prove Proposition 1, we shall need some preliminary results. We follow the information-theory literature in preferring to deal with the additive inverse of entropy, or "negentropy." The following are presented without proof:

Lemma 1—[28]: The negentropy $\int f \log f \, dx$ is a convex function of f.

Lemma 2—[29]: $\mathcal{P}_\varepsilon$ is convex.

Lemma 3—[29]: For a convex function $f : [0, 1] \to \mathbb{R}$, $f(0) \le f(\lambda)$ for every $\lambda \in [0, 1]$ if and only if

$$0 \le \lim_{\lambda \downarrow 0} \frac{1}{\lambda} (f(\lambda) - f(0)). \tag{7}$$

With these facts in hand, we now proceed to show that Huber's least favorable distribution also maximizes entropy over the family of ε-contaminated normal distributions. (This development is strongly inspired by Huber's result for the Fisher information [19], [30].) For simplicity, we present the proof for $\sigma = 1$; the generalization is straightforward.

Proof of the Proposition:

By Lemma 1, the negentropy is convex in f, and by Lemma 2, $\mathcal{P}_\varepsilon$ is convex. Therefore, by Lemma 3, it is sufficient to prove that

$$\frac{\partial}{\partial \lambda} H(f_\lambda)\bigg|_{\lambda=0} \ge 0 \tag{8}$$

where H denotes the negentropy and $f_\lambda = (1 - \lambda) f_H + \lambda f$ for any $f \in \mathcal{P}_\varepsilon$.

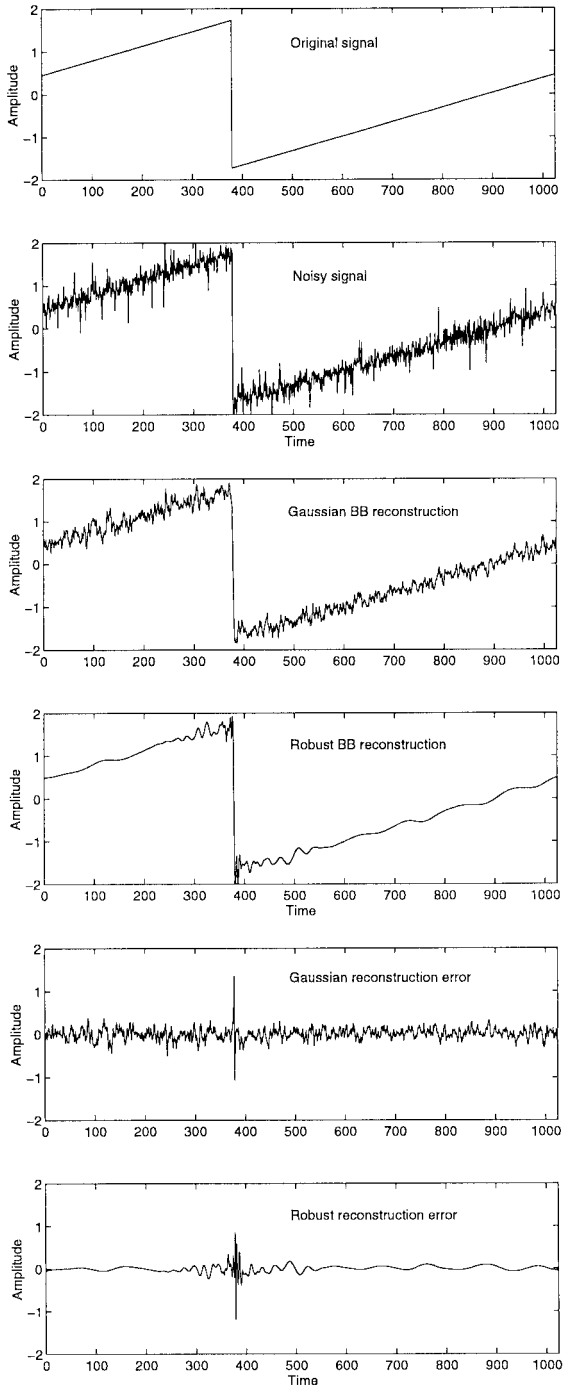

Fig. 7. Comparison of the reconstructions using the robust criterion and entropy-based criterion.

Assuming sufficiently well-behaved f_λ

$$\begin{aligned}\frac{\partial}{\partial\lambda}H(f_\lambda) &= \int \frac{\partial}{\partial\lambda} f_\lambda \log f_\lambda \, dc \\ &= \int \frac{\partial}{\partial\lambda}((1-\lambda)f_H + \lambda f) \\ &\quad \cdot \log((1-\lambda)f_H + \lambda f)\, dc \\ &= \int (f - f_H)\log((1-\lambda)f_H + \lambda f)\, dc \\ &\quad + \int (f - f_H)\, dc\end{aligned}$$

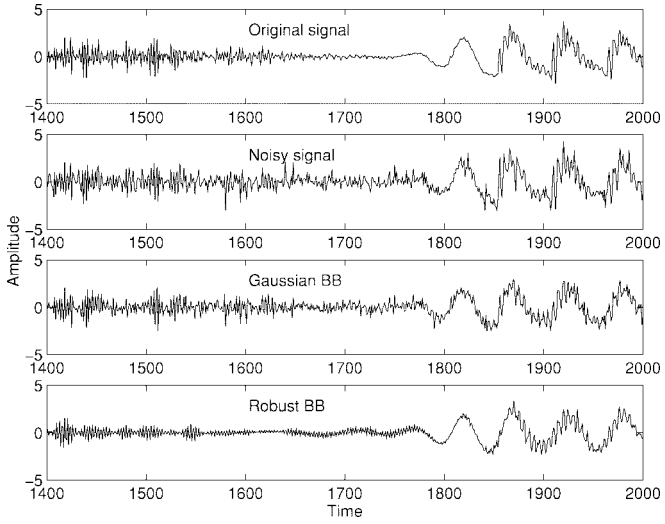

Fig. 8. Speech signal: comparison of the reconstructions using the robust criterion and entropy-based criterion.

where the last term vanishes since f and f_H are both distributions and, therefore, integrate to unity. It follows that

$$\frac{\partial}{\partial\lambda}H(f_\lambda)\bigg|_{\lambda=0} = \int (f - f_H)\log f_H \, dc$$

and we must show that this quantity is nonnegative for any $f \in \mathcal{P}_\varepsilon$.

From (2), we have

$$\log f_H(c) = \begin{cases} \log(1-\varepsilon)\phi(a) + (ac + a^2), & c \le -a \\ \log(1-\varepsilon)\phi(a) + \frac{1}{2}(-c^2 + a^2), & -a \le c \le a \\ \log(1-\varepsilon)\phi(a) + (-ac + a^2), & a \le c. \end{cases}$$

As before

$$\int (f - f_H)\log(1-\varepsilon)\phi(a)\, dc = 0$$

since f and f_H are both distributions and the rest of the integrand is a constant with respect to c. We are left with

$$\begin{aligned}\int (f - f_H)\log f_H \, dc = &\int_{-\infty}^{-a} (f - f_H)(ac + a^2)\, dc \\ &+ \int_{-a}^{a} (f - f_H)\,\tfrac{1}{2}(-c^2 + a^2)\, dc \\ &+ \int_{a}^{\infty} (f - f_H)(-ac + a^2)\, dc.\end{aligned} \tag{9}$$

Recall that $f = (1-\varepsilon)\phi + \varepsilon g$ for some $g \in \mathcal{F}$, and consider first the interval $-a \le c \le a$. Here, $f_H = (1-\varepsilon)\phi$ so that $f - f_H = \varepsilon g \ge 0$; furthermore, $-c^2 + a^2 \ge 0$, so that the middle term in (9) is nonnegative. Now, since $f - f_H \ge 0$ for $-a \le c \le a$ and the difference $f - f_H$ must integrate to zero, we have that

$$\int_{-\infty}^{-a} (f - f_H)\, dc + \int_{a}^{\infty} (f - f_H)\, dc \le 0.$$

Furthermore, $ac + a^2 \le 0$ when $c \le -a$, and $-ac + a^2 \le 0$ when $a \le c$. Thus

$$\int_{-\infty}^{-a} (f - f_H)(ac + a^2)\, dc + \int_{a}^{\infty} (f - f_H)(-ac + a^2)\, dc \ge 0$$

establishing (8).

Finally, (3) is obtained by setting $\int f_H\,dc = 1$. We have

$$\int_{-\infty}^{-a}(1-\varepsilon)\phi(a)e^{(ac+a^2)}\,dc = \int_a^{\infty}(1-\varepsilon)\phi(a)e^{(-ac+a^2)}\,dc = (1-\varepsilon)\frac{\phi(a)}{a}$$

and

$$\int_{-a}^{a}(1-\varepsilon)\phi_\sigma(c)\,dc = (1-\varepsilon)(\Phi(a)-\Phi(-a)) = (1-\varepsilon)(1-2\Phi(-a)).$$

Substitution concludes the proof. □

Appendix B
Proof of Proposition 2

We first present without proof a slight restatement of a theorem due to Verdú and Poor [31].

Theorem 1: Let a minimax problem be defined by a set $\mathcal{S}$ of allowable estimators, a set $\mathcal{P}$ of distributions, and an objective function $J\colon \mathcal{S}\times\mathcal{P}\to\mathbb{R}$. Let $\mathcal{P}$ be convex, and let $J(\theta, P)$ be convex on $\mathcal{P}$ for every $\theta\in\mathcal{S}$. Let P_0 be the least favorable distribution in $\mathcal{P}$, and define

$$P_\lambda = (1-\lambda)P_0 + \lambda P$$

for any $P\in\mathcal{P}$ and $0\le\lambda\le 1$. Finally, let $\theta^*(P)\in\mathcal{S}$ denote the optimal estimator corresponding to the distribution $P\in\mathcal{P}$. If $\theta_0\in\mathcal{S}$ is an estimator satisfying the regularity condition

$$J(\theta^*(P_\lambda), P_\lambda) - J(\theta_0, P_\lambda) = o(\lambda)$$

for all $P\in\mathcal{P}$, then (θ_0, P_0) is a saddle point solution.

We use this theorem to establish that Huber's least favorable distribution f_H and the MLE based on it constitute the solution of our minimax problem.

Proof of the Proposition: By Lemma 1, the negentropy is convex in f, and by Lemma 2, $\mathcal{P}_\varepsilon$ is convex. Therefore, by Theorem 1, it is sufficient to prove that the regularity condition holds, i.e., that

$$\int f_\lambda(c, \hat\theta_\lambda(c))\log f_\lambda(c, \hat\theta_\lambda(c))\,dc - \int f_\lambda(c, \hat\theta_H(c))\log f_\lambda(c, \hat\theta_H(c))\,dc = o(\lambda) \quad (10)$$

where $\hat\theta_\lambda$ is the MLE based on f_λ.

Note first that by definition,

$$f_\lambda = (1-\lambda)f_H + \lambda f = f_H + \lambda(f-f_H) = f_H + \lambda\varepsilon(g-g_H)$$

since f, $f_H\in\mathcal{P}_\varepsilon$. Thus for fixed c

$$f'_\lambda = f'_H + \lambda\varepsilon(g'-g'_H) \quad (11)$$

where the prime denotes differentiation with respect to θ. In particular, since $\hat\theta_H(c)$ maximizes $f_H(c,\theta)$ for each c, $f'_H(c, \hat\theta_H(c))$ vanishes and

$$f'_\lambda(c, \hat\theta_H(c)) = \lambda\varepsilon\Big(g'(c, \hat\theta_H(c)) - g'_H(c, \hat\theta_H(c))\Big). \quad (12)$$

Next, for fixed c, we write $f'_\lambda(c,\theta)$ as a first-order Taylor expansion around the point $(c, \hat\theta_H(c))$

$$f'_\lambda(c,\theta) = f'_\lambda(c, \hat\theta_H(c)) + (\theta-\hat\theta_H(c))f''_\lambda(c, \hat\theta_H(c)) + o(\theta-\hat\theta_H(c)).$$

In particular, setting $\theta = \hat\theta_\lambda = \hat\theta_H + \Delta(\lambda)$ (for some suitable function Δ), and noting that $\hat\theta_\lambda(c)$ maximizes $f_\lambda(c,\theta)$ for each c, so that $f'_\lambda(c, \hat\theta_\lambda(c))$ vanishes, we get

$$0 = f'_\lambda(c, \hat\theta_H(c)) + \Delta(\lambda)f''_\lambda(c, \hat\theta_H(c)) + o(\Delta(\lambda))$$

which, combined with (12), yields

$$\lambda\varepsilon\Big(g'(c, \hat\theta_H(c)) - g'_H(c, \hat\theta_H(c))\Big) + \Delta(\lambda)f''_\lambda(c, \hat\theta_H(c)) + o(\Delta(\lambda)) = 0. \quad (13)$$

But differentiating (11)

$$f''_\lambda(c, \hat\theta_H(c)) = f''_H(c, \hat\theta_H(c)) + \lambda\varepsilon\Big(g''(c, \hat\theta_H(c)) - g''_H(c, \hat\theta_H(c))\Big)$$

and substituting this into (13) we obtain

$$\lambda\varepsilon(g'(c, \hat\theta_H(c)) - g'_H(c, \hat\theta_H(c))) + \Delta(\lambda)\Big(f''_H(c, \hat\theta_H(c)) + \lambda\varepsilon(g''(c, \hat\theta_H(c)) - g''_H(c, \hat\theta_H(c)))\Big) + o(\Delta(\lambda)) = 0$$

or

$$O(\lambda) + O(\Delta(\lambda)) + o(\Delta(\lambda)) = 0$$

whence it follows that

$$O(\Delta(\lambda)) = O(\lambda). \quad (14)$$

Again for fixed c, we write $f_\lambda(c, \hat\theta_\lambda)$ as a first-order Taylor expansion around the point $(c, \hat\theta_H(c))$, and obtain

$$f_\lambda(c, \hat\theta_\lambda) = f_\lambda(c, \hat\theta_H(c)) + \Delta(\lambda)f'_\lambda(c, \hat\theta_H(c)) + o(\Delta(\lambda)) = f_\lambda(c, \hat\theta_H(c)) + \Delta(\lambda)O(\lambda) + o(\Delta(\lambda)) = f_\lambda(c, \hat\theta_H(c)) + o(\lambda)$$

from (12) and (14). Substitution into (10), and noting that f_λ integrates to unity, conclude the proof.

Appendix C
Proof of Proposition 3

As in [3], our strategy for denoising is to interpret the problem as one of coding a data string. The underlying assumption is that the signal of interest has a very compact representation in the wavelet domain, and this assumption is key to the result. The efficiency of this coding, as first described by Rissanen [4] and used in [3], may be measured by the description length. Parsimony of representation in the wavelet domain is quantified by the MDL criterion, which is, in turn, a function of the joint density of the coefficients $\{C_i^x\}$.

Proof of the Proposition: The set $\{C_i^x\}$ is reindexed, if necessary, so that the magnitudes of the coefficients are in ascending order. For some $K\in\{1,\cdots,N\}$, suppose that

$\{C_1^x, \cdots, C_K^x\}$ contain signal, while $\{C_{K+1}^x, \cdots, C_N^x\}$ are pure noise. From (5), we have that

$$\mathcal{L}(\mathcal{C}^N; K) - \mathcal{L}(\mathcal{C}^N; K-1) = -\log f_H(0) + \log f_H(C_K^x) + \log N < 0$$

since K minimizes the MDL by hypothesis. It follows that

$$\log f_H(C_K^x) < \log f_H(0) - \log N = \log \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} - \log N. \quad (15)$$

Now, if $|C_K^x| < a$, then it follows from (2) that

$$\log f_H(C_K^x) = \log \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} - \frac{(C_K^x)^2}{2\sigma^2}$$

and substituting this into (15), we get

$$|C_K^x| > \sigma\sqrt{2 \log N}.$$

By constrast, if $|C_K^x| > a$, then (2) implies that

$$\log f_H(C_K^x) = \log \frac{1-\varepsilon}{\sqrt{2\pi}\sigma} + \frac{a^2}{2\sigma^2} - \frac{a|C_K^x|}{\sigma^2}$$

which, upon substitution into (15), yields

$$|C_K^x| > \frac{a}{2} + \frac{\sigma^2}{a} \log N.$$

Similarly, setting

$$\mathcal{L}(\mathcal{C}^N; K) - \mathcal{L}(\mathcal{C}^N; K+1) = \log f_H(0) - \log f_H(C_{K+1}^x) - \log N < 0$$

can be shown to yield

$$|C_{K+1}^x| < \sigma\sqrt{2 \log N}$$

when $|C_K^x| < a$, and

$$|C_{K+1}^x| < \frac{a}{2} + \frac{\sigma^2}{a} \log N$$

when $|C_K^x| > a$.

Finally, it is easy to see from Fig. 1 that an abscissa of a corresponds to an ordinate of $a^2/2\sigma^2$, so the two cases above, $|C_K^x| < a$ and $|C_K^x| > a$, respectively, correspond to $\log N < a^2/2\sigma^2$ and $\log N > a^2/2\sigma^2$. This concludes the proof. □

ACKNOWLEDGMENT

The authors are grateful to D. Tucker for help with some of the numerical examples.

REFERENCES

[1] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inform. Theory*, vol. 38, pp. 617–643, 1992.
[2] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. ASA*, vol. 90, pp. 1200–1223, 1995.
[3] H. Krim and J.-C. Pesquet, "On the statistics of best bases criteria," in *Wavelets in Statistics of Lecture Notes in Statistics*, A. Antoniadis and G. Oppenheim, Eds. New York: Springer-Verlag, 1995, pp. 193–207.
[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. Math.*, vol. 44, pp. 423–453, 1943.
[6] M. Neumann and V. Spokoiny, "On the efficiency of wavelet estimators under arbitrary error distributions," *Math. Meth. Statist.*, vol. 4, no. 2, pp. 137–166, 1995.
[7] M. Neumann and R. von Sachs, "Wavelet thresholding: Beyond the Gaussian i.i.d. situation," in *Wavelets and Statistics, Lecture Notes in Statistics*, vol. 103. New York: Springer-Verlag, Nov. 1994, pp. 301–329; papers from the 15th Franco-Belgian Meeting of Statisticians, Villard de Lans, France, Nov. 16–18, 1994.
[8] A. Bruce, D. Donoho, H.-Y. Gao, and R. Martin, "Denoising and robust nonlinear wavelet analysis," in *Wavelet Applications, Proc. SPIE*, H. Szu, Ed. Orlando, FL, Soc. Photo-optical Instrum. Eng., Apr. 1994, pp. 325–336.
[9] P. Huber, "Robust estimation of a location parameter," *Ann. Math. Stat.*, vol. 35, pp. 1753–1758, 1964.
[10] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, pp. 713–718, Mar. 1992.
[11] D. Donoho and I. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases," *C. R. Acad. Sci. Paris*, vol. 319, pp. 1317–1322, Oct. 1994.
[12] H. Krim, S. Mallat, D. Donoho, and A. Willsky, "Best basis algorithm for signal enhancement," presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95), Detroit, MI, May 1995.
[13] H. Krim, D. Tucker, S. Mallat, and D. Donoho, "On denoising and best basis representation," *IEEE Trans. Inform. Theory*, to be published.
[14] B. Vidakovic, "On algorithmic complexity, universal priors and ockham's razor," ISDS, Duke Univ., Duke Univ., Durham, NC 27708-0251, Tech. Rep., 1996.
[15] H. Krim and D. Brooks, "Feature-based best basis segmentation of ecg signals," in *IEEE Symp. Time-Freq./Time Scale Anal.* (Paris, France, June 1996).
[16] D. Leporini, J.-C. Pesquet, and H. Krim, *Bayesian Approach to Best Basis Selection*, B. Vidakovic and P. Müller, Eds. New York: Springer Verlag, 1999, to be published.
[17] J.-C. Pesquet, H. Krim, and H. Carfantan, "Time invariant orthonormal wavelet representations," *IEEE Trans. Signal Processing*, vol. 44, pp. 1964–1970, Aug. 1996.
[18] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.
[19] P. Huber, "Théorie de l'inférence statistique robuste," Univ. de Montréal, Montreal, Que., Canada, Tech. Rep., 1969.
[20] B. Vidakovic, "Nonlinear wavelet shrinkage with bayes rules and bayes," *J. Amer. Statist. Assoc.*, vol. 93, pp. 173–179, 1998.
[21] G. M. Lachlan and T. Krishnan, "The EM algorithm and extensions," in *Probability and Statistics.* New York: Wiley, 1997.
[22] D. Donoho and P. Huber, "The notion of breakdown point," in *Ferstschrift for E. Lehman*. Belmont, CA: Wadsworth, 1984, pp. 158–184.
[23] D. Tucker, "Wavelet denoising techniques with applications to high resolution radar," Master's thesis, MIT, Cambridge, MA, June 1997.
[24] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distorsion sense," *IEEE Trans. Image Processing*, vol. 2, pp. 160–175, Feb. 1993.
[25] I. Schick and H. Krim, "Robust wavelet denoising," presented at the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97) (Munich, Germany, May 1997).
[26] R. Coifman and Y. Meyer, "Remarques sur l'analyze de Fourier à fenêtre," *C. R. Acad. Sci. Série I*, pp. 259–261, 1991.
[27] M. V. Wickerhauser, "INRIA lectures on wavelet packet algorithms," in *Ondelettes et Paquets d'Ondelettes*, Roquencourt, France, June 17–21, 1991, pp. 31–99.
[28] T. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.
[29] I. Schick, "Robust recursive estimation of the state of a discrete-time stochastic linear dynamic system in the presence of heavy-tailed observation noise," Tech. Rep. LIDS-TH-1975, MIT, Cambridge, MA, 02139, 1989.
[30] P. Huber, *Robust Statistics*, 1st ed. New York: Wiley, 1981.
[31] S. Verdú and H. V. Poor, "On minimax robustness: A general approach and applications," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 328–340, 1984.