On Denoising and Best Signal Representation

Hamid Krim, *Senior Member, IEEE*, Dewey Tucker, Stéphane Mallat, *Member, IEEE*, and David Donoho

***Abstract*— We propose a best basis algorithm for signal enhancement in white Gaussian noise. The best basis search is performed in families of orthonormal bases constructed with wavelet packets or local cosine bases. We base our search for the "best" basis on a criterion of minimal reconstruction error of the underlying signal. This approach is intuitively appealing because the enhanced or estimated signal has an associated measure of performance, namely, the resulting mean-square error. Previous approaches in this framework have focused on obtaining the most "compact" signal representations, which consequently contribute to effective denoising. These approaches, however, do not possess the inherent measure of performance which our algorithm provides.**

We first propose an estimator of the mean-square error, based on a heuristic argument and subsequently compare the reconstruction performance based upon it to that based on the Stein unbiased risk estimator. We compare the two proposed estimators by providing both qualitative and quantitative analyses of the bias term. Having two estimators of the mean-square error, we incorporate these cost functions into the search for the "best" basis, and subsequently provide a substantiating example to demonstrate their performance.

***Index Terms*— Best basis, denoising, Stein risk, thresholding, wavelet, wavelet packet.**

I. Introduction

THE quintessential goal of statistical estimation is to elicit useful information about a signal underlying an observed random process. This information, which could either completely characterize the signal or at least consist of signal parameters crucial to the problem at hand (e.g., delay estimation), is generally obtained by using some side information about the process itself. The reconstruction of an unknown (or minimally known) signal embedded in noise, for example, would generally make use of some prior information about the contaminating noise. An estimation problem also entails the specification of an error objective criterion to be optimized in some chosen functional space. The choice of the criterion results in estimates with different reconstruction performances which depend on

a) the additive noise,
b) the smoothness class of the underlying signal,
c) the selected estimator.

One can easily verify that the performance of a Wiener filter[1] loses its optimality to a nonlinear filter for signal inputs from a class of piecewise-smooth signals, confirming statement b) above.

The recent resurgence of interest in the nonparametric estimation problem may primarily be attributed to the emergence of wavelet bases as not only unconditional orthogonal bases for a large class of smoothness spaces [6], [10], but as an efficient framework for function expansion as well. The problem of estimating an unknown signal embedded in Gaussian noise has received a great deal of attention in numerous studies, and will also be of interest in this sequel. For such a problem, one is generally led to invoke the least squares error criterion in evaluating a given signal reconstruction/estimation procedure. Different estimation rules could subsequently be compared on the basis of their resulting *mean-square error* (MSE) (henceforth referred to as the risk).

Stein [16] has under quite general conditions, derived an *unbiased estimator* of such a risk for a Gaussian estimator. The weak differentiability he assumed for an adopted estimation rule allows one to theoretically evaluate a wide class of estimators, including those which are nonlinear, as discussed below. This resulting risk estimator thus provides one with a theoretical means to predict performance, which in turn is key to not only selecting an acceptable signal estimation procedure, but to also obviating costly and time-consuming simulations in its assessment.

Donoho and Johnstone [6] were first to formalize the wavelet coefficient thresholding for removal of additive noise from deterministic signals. The discrimination between signal and noise is achieved by choosing an orthogonal basis which efficiently approximates the signal (with few nonzero coefficients). A signal enhancement can thus be obtained by discarding components below a predetermined threshold. Wavelet orthonormal bases have been shown to be particularly well-adapted to approximate piecewise-smooth functions. The nonzero wavelet coefficients are typically located in the neighborhood of sharp signal transitions, and thresholding any coefficient at a specific level was shown to provide a quasi-

Manuscript received December 1, 1997; revised June 1, 1999. The work of H. Krim was supported in part by AFOSR under Grant F49620-98-1-0190 and by ONR-MURI under Grant JHU 8906–48182. The work of D. Tucker was supported in part by the Army Research Office under Contract DAAL-03-92-G-115 and the AFOSR under Contract F49620-92-J-2002. The work of S. Mallat was supported by AFOSR under Grant F49620-93-1-0102 and ONR under Grant N00014-91-J-1967. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, MIT, Cambridge, MA, August 16–21, 1998 and at ICASSP 1995, Detroit, MI, May 1995.

H. Krim is with the Electrical and Computer Engineering Department, North Carolina State University, Raleigh, NC 27695 USA (e-mail: ahk@eos.ncsu.edu).

D. Tucker is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dtucker@mit.edu),

S. Mallat is with the Centre de Mathematiques Appliquées, Ecole Polytechnique, Paris, France.

D. Donoho is with the Department of Statistics, Stanford University, Stanford, CA 94205 USA.

Communicated by C. Herley, Associate Editor for Estimation.

Publisher Item Identifier S 0018-9448(99)08207-3.

[1] This can be interpreted in terms of an optimal Karhunen–Loève expansion of a signal.

optimal *min-max estimator* of a noisy piecewise-smooth signal in a wavelet basis [6]. In spite of its nonlinearity, such a wavelet-based estimator can be theoretically evaluated with no need for experimentation by way of its predicted risk, thus affording one the ability to appropriately select an analysis wavelet.

As briefly alluded to earlier, a given wavelet function may not necessarily be best adapted to an underlying signal of an observed process; furthermore, the reconstruction performance is dependent upon the noise realization. This indicates that a universal wavelet basis is more than one could hope for, and that further optimization is required. When a signal includes more complex structures and in particular high-frequency oscillations, it becomes necessary to adaptively select an appropriate *best basis* which provides the best signal estimate upon discarding (thresholding) the noisy coefficients. Note that the entropy-based adapted/best basis search proposed in [2], [12], and [17] does not account for the statistical properties of the noise and, as a result, is fraught with highly variable performance, particularly in noisy scenarios. To address this inherent variability,[2] a new class of algorithms have recently been studied in [5] and also in [9]. An approach was first proposed in [5] and consisted of performing a best basis search in families of orthonormal bases constructed with wavelet packets or local cosine bases. This is achieved by capitalizing on a representation mismatch of the underlying signal and of the additive noise in such a basis. This, as a result, affords one the ability to discriminate against the noise and optimally retrieve the signal by minimizing a risk estimate similar to that described for wavelet coefficient thresholding. Estimating this risk in a given basis will be the first focus of this paper. By specializing the derivation to a white Gaussian noise setting, we are able to analyze this estimate and prove it to be biased by calling upon the Stein unbiased risk estimator of a mean of a multivariate normal distribution.

To stay within the intended scope of this paper, we assume throughout that the statistical properties of the noise are known, namely, Gaussian with zero-mean and known variance, and that the signal of interest is unknown. In the next section, we briefly discuss the issues associated with noise removal by thresholding. In Section III, we derive an unbiased risk estimate of a wavelet-based signal estimator, which we compare to a heuristically derived risk. In Section IV, we extend the application of the risk estimate to select a "best" basis which leads to an enhanced signal reconstruction. We give some concluding remarks in Section V.

II. Noise Removal by Thresholding

As briefly alluded to earlier, any prior knowledge (quantitative or qualitative) about an undesired noise contaminating a signal can and should be used in estimating the latter. In addition, implementing an estimator in an orthogonal basis is intuitively appealing on account of the distribution of the noise energy in such a basis. This indeed provides important information for discriminating between the signal and noise, which to a great extent contributes to obtaining a good approximation of the signal. To approximate a signal in a given smoothness class $\mathcal{S}$, which includes piecewise-smooth polynomial signals, an adapted wavelet basis offers, as noted earlier, more flexibility than the classical Karhunen–Loève (K-L) basis. This synergy between an adapted signal representation and the noise removal problem is of central importance to our proposed best basis search technique.

We can succinctly state the problem as one of retrieving an unknown deterministic signal $\{s(t)\}$ after observing a process $\{x(t)\}$ sampled over an interval of length N. We henceforth assume that the observed samples $\{x[m]\}$ are those of an underlying unknown signal $\{s[m]\}$ and of white noise $\{n[m]\}$, where

$$x[m] = s[m] + n[m] \tag{1}$$

for $m = 1, 2, \cdots, N$.

Let $\mathcal{B}^p \in \mathcal{D} = \{\mathcal{B}^p | p \in \mathcal{P}\}$ where $p \in \mathcal{P}$ is some partition of the unit interval $[0, 1]$ and $\mathcal{B}^p = \{\boldsymbol{W}^p_{x_i}\}_{1 \le i \le N}$ is a set of vectors forming a basis of our observation space. Our goal is to guard against the *worst case noise* coefficients (i.e., exclude the components which are potentially only noise) by using the supremum value of a Gaussian random variable. Towards that end, we call upon a statistical theory which stipulates that the extreme values assumed by variables from a given distribution enjoy a corresponding limit distribution which represents a domain of attraction [15]. This limit distribution may provide *suprema* values *in probability*, from which a thresholding procedure naturally follows. It consists of discarding all inner products $\{\langle \boldsymbol{x}, \boldsymbol{W}^p_{x_i} \rangle\}$ below a threshold T, in order to reconstruct an estimate $\{\hat{s}[m]\}$ of $\{s[m]\}$. We denote the vector of observed samples $\{x[m]\}$ by $\boldsymbol{x}$ and the ith-basis vector by $\boldsymbol{W}^p_{x_i}$. Let $K = \text{Card}\,\{|\langle \boldsymbol{x}, \boldsymbol{W}^p_{x_i} \rangle| > T\}$ and suppose that the coefficients $\{\langle \boldsymbol{x}, \boldsymbol{W}^p_{x_i} \rangle\}$ are sorted in decreasing magnitude for $1 \le i \le N$. We then have

$$\hat{\boldsymbol{s}} = \sum_{i=1}^{K} \boldsymbol{W}^p_{x_i} \langle \boldsymbol{x}, \boldsymbol{W}^p_{x_i} \rangle. \tag{2}$$

The threshold T will clearly vary with the noise statistics and is ideally chosen so that $\sup_i |\langle \boldsymbol{n}, \boldsymbol{W}^p_{x_i} \rangle| \to T$ *almost surely (a.s.)* [15]. For Gaussian white noise of variance σ^2, the coefficients $\{\langle \boldsymbol{n}, \boldsymbol{W}^p_{x_i} \rangle\}_{1 \le i \le N}$ are N independent Gaussian random variables with the same variance. Under some general conditions, the value assumed by the supremum of $\{|\langle \boldsymbol{n}, \boldsymbol{W}^p_{x_i} \rangle|^2\}_{1 \le i \le N}$ is then "$2\sigma^2 \log N$"[3] in probability (i.p.) [7]. To guarantee that the thresholded coefficients always include some signal information, one chooses $T = \sqrt{2\sigma^2 \log N}$, which was shown to be an optimal threshold from a number of perspectives [6], [8]. The vectors $\boldsymbol{W}^p_{x_i}$ for $i \le K$ will generally have weights that correspond to the nonzero signal coefficients (i.e., $|\langle \boldsymbol{s}, \boldsymbol{W}^p_{x_i} \rangle| \ne 0$). Wavelet bases are known to concentrate the energy of piecewise-smooth signals into a few high-energy coefficients [4]. If the energy of $\{s[m]\}$ is concentrated into a few high-amplitude coefficients, such a representation can provide an accurate

[2] The basis search is very sensitive to noise realization.

[3] Unless otherwise indicated, "log" indicates the natural logarithm $\log_e$.

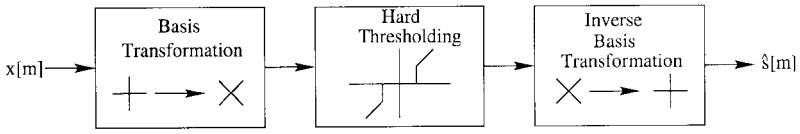


Fig. 1. Procedure for reconstructing a noisy signal.

estimate of $\{s[m]\}$. Thus the advantage of expressing $\{x[m]\}$ in an orthogonal wavelet basis is two-fold.

a) If the contaminating noise samples are independent and identically distributed (i.i.d.) Gaussian, so are the coefficients, and their statistical independence is preserved.
b) Intrinsic properties of the signal are preserved in a wavelet basis.

When a signal possesses more complex features, one has to search for the basis which would result in its best signal estimation. Before introducing this idea, we first discuss a method for estimating the mean-square error associated with thresholding wavelet coefficients at a given level T. Given a signal $\{s[m]\}$ in some basis representation, we will threshold the coefficients and estimate the resulting error, and this error will then be used in the search for the best basis, as discussed in Section IV.

III. Risk of a Wavelet-Based Estimator

In this section, we propose a mean-square error estimator and proceed to derive its bias. The mean-square error, or more formally the *risk,* is given by

$$\mathcal{R}(\boldsymbol{s}, T) = \mathrm{E}\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\} \tag{3}$$

where $\hat{\boldsymbol{s}}$ is the vector representation of the reconstructed signal. As shown in Fig. 1, a signal reconstruction is obtained by thresholding a set of coefficients in a given basis and then applying an inverse transformation. This is the general procedure that we use throughout the paper. For clarity, the thresholding procedure will be strictly limited to a hard thresholding rule.

A. Proposed Risk Estimator

It is often desirable to theoretically assess the quality of an estimator and predict its limitations in a variety of scenarios. We first follow a simple approach to derive an estimator of the risk as defined in (3). While this approach is certainly applicable to any noise scenario with a corresponding threshold T, we restrict our study to Gaussian noise for the clarity of exposition. Moreover, we prove the existence of an unbiased risk estimator for this case by deriving it.

To proceed, let $\gamma_T(y_x)$ for any given $y_x \in \mathbb{R}$ denote a thresholding rule which eliminates basis coefficients at or below a level T. As we have argued earlier, the choice of the threshold is based upon ensuring that any isolated noise coefficient (i.e., noise only and devoid of signal contribution) will be discarded

$$\gamma_T(y_x) = \begin{cases} y_x, & \text{if } |y_x| > T \\ 0, & \text{if } |y_x| \le T. \end{cases} \tag{4}$$

In recovering signal coefficients in additive noise, this decision rule has an associated quadratic loss which depends on T and the signal coefficient y_s, or

$$\mathcal{L}\{\gamma_T(y_x), y_s, T\} = (y_s - \gamma_T(y_x))^2. \tag{5}$$

Note that when applied to the wavelet coefficients in a particular basis $\mathcal{B}^p = \{\boldsymbol{W}^p_{x_i}\}$, the mean value of the loss is the estimation error or risk in $\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|$, or

$$\begin{aligned} \mathrm{E}\{\mathcal{L}(\gamma_T(\boldsymbol{W}^p_x), \boldsymbol{s}, T)\} &= \mathcal{R}(\boldsymbol{s}, T) \\ &= \mathrm{E}\{\|\boldsymbol{s} - \boldsymbol{W}^p_x \gamma_T(\boldsymbol{W}^p_x)\|^2\}. \end{aligned} \tag{6}$$

For compactness, $\boldsymbol{W}^p_x$ represents the vector with components $\mathcal{W}^p_{x_i} = \langle \boldsymbol{x}, \boldsymbol{W}^p_{x_i} \rangle$, and $\boldsymbol{W}^p_x$ represents the corresponding matrix of basis functions.

Since we only consider orthogonal bases here, the risk can be expressed in terms of the basis coefficients, or

$$\begin{aligned} \mathcal{R}(\boldsymbol{s}, T) &= \mathrm{E}\{\|\boldsymbol{W}^p_x \boldsymbol{W}^p_s - \boldsymbol{W}^p_x \gamma_T(\boldsymbol{W}^p_x)\|^2\} \\ &= \sum_{i=1}^{N} \mathrm{E}\{|\mathcal{W}^p_{s_i} - \gamma_T(\mathcal{W}^p_{x_i})|^2\}. \end{aligned} \tag{7}$$

There are two cases that must be analyzed in order to define an estimator.

Case 1: If $|\mathcal{W}^p_{x_i}|^2 \le T^2$ with the hard thresholding strategy, this coefficient is set to zero. This contributes the value of $|\mathcal{W}^p_{s_i}|^2$ to the total risk. Since

$$\mathrm{E}\{|\mathcal{W}^p_{x_i}|^2\} = |\mathcal{W}^p_{s_i}|^2 + \sigma^2 \tag{8}$$

$|\mathcal{W}^p_{s_i}|^2$ is evaluated as $|\mathcal{W}^p_{x_i}|^2 - \sigma^2$.

Case 2: If $|\mathcal{W}^p_{x_i}|^2 > T^2$, this coefficient is left unchanged, yielding a mean-square error that is on average equal to the noise variance σ^2.

The total approximation error can thus be estimated by

$$\mathcal{R}_B(\boldsymbol{s}, T) = \sum_{i=1}^{N} \Phi(|\mathcal{W}^p_{x_i}|^2) \tag{9}$$

where

$$\Phi(u) = \begin{cases} u - \sigma^2, & \text{if } u \le T^2 \\ \sigma^2, & \text{if } u > T^2. \end{cases} \tag{10}$$

We use the symbol $\mathcal{R}_B(\boldsymbol{s}, T)$ to denote that this estimator is biased, a fact which will be shown below. In the following theorem, we compute the true risk $\mathcal{R}(\boldsymbol{s}, T) = \mathrm{E}\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|\}^2$ and derive the bias using the Stein unbiased risk estimator [16].

Theorem 1: Let $\{\boldsymbol{W}_{x_i}^p\}_{1\le i\le N}$ be an orthonormal basis of the observation space. If the coefficients $\{n[m]\}$ are zero-mean, uncorrelated Gaussian random variables with variance σ^2, the bias of the estimator $\mathcal{R}_B(\boldsymbol{s}, T)$ with respect to $\mathcal{R}(\boldsymbol{s}, T)$ is

$$\mu = \mathcal{R}(\boldsymbol{s}, T) - \mathrm{E}\{\mathcal{R}_B(\boldsymbol{s}, T)\} = 2T\sigma^2 \sum_{i=1}^{N} [\phi(T - \langle \boldsymbol{s}, \boldsymbol{W}_{x_i}^p\rangle) + \phi(-T - \langle \boldsymbol{s}, \boldsymbol{W}_{x_i}^p\rangle)] \quad (11)$$

with

$$\phi(u) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(u^2/2\sigma^2)}.$$

Proof: Recall that the basis coefficients resulting from the orthogonal transform are denoted by $\{\mathcal{W}_{x_i}^p\}$ along with the corresponding vector of basis functions $\{\boldsymbol{W}_{x_i}^p\}$. In the proof below, we drop the superscript p for clarity. Define

$$\gamma_T(\mathcal{W}_{x_i}) = \mathcal{W}_{x_i}\mathcal{I}_{\{|\mathcal{W}_{x_i}|>T\}}$$
$$g_T(\mathcal{W}_{x_i}) = -\mathcal{W}_{x_i}\mathcal{I}_{\{|\mathcal{W}_{x_i}|\le T\}}$$

where $\mathcal{I}_{\{\cdot\}}$ is an indicator function constrained by its argument and where the noisy coefficient $\mathcal{W}_{x_i}$ has a normal distribution $\mathcal{W}_{x_i} \sim \mathcal{N}(\mathcal{W}_{s_i}, \sigma^2)$. We can then write

$$\gamma_T(\mathcal{W}_{x_i}) = \mathcal{W}_{x_i} + g_T(\mathcal{W}_{x_i})$$

to obtain the following:

$$\begin{aligned}\mathrm{E}\left\{\sum_{i=1}^{N}[\gamma_T(\mathcal{W}_{x_i}) - \mathcal{W}_{s_i}]^2\right\} &= \sum_{i=1}^{N}\mathrm{E}\{[(\mathcal{W}_{x_i} - \mathcal{W}_{s_i}) + g_T(\mathcal{W}_{x_i})]^2\} \\ &= \sum_{i=1}^{N}(\mathrm{E}\{(\mathcal{W}_{n_i})^2\} + 2\mathrm{E}\{\mathcal{W}_{n_i}g_T(\mathcal{W}_{x_i})\} + \mathrm{E}\{g_T^2(\mathcal{W}_{x_i})\}).\end{aligned} \quad (12)$$

Using the property described in [16] (i.e., differentiation of a distribution)

$$\begin{aligned}\mathrm{E}\{\mathcal{W}_{n_i}g_T(\mathcal{W}_{n_i})\} &= \int \mathcal{W}_{n_i}g_T(\mathcal{W}_{n_i} + \mathcal{W}_{s_i})\phi(\mathcal{W}_{n_i})\,d\mathcal{W}_{n_i} \\ &= -\sigma^2\int g_T(\mathcal{W}_{n_i} + \mathcal{W}_{s_i})\phi'(\mathcal{W}_{n_i})\,d\mathcal{W}_{n_i} \\ &= \sigma^2\int g_T'(\mathcal{W}_{n_i} + \mathcal{W}_{s_i})\phi(\mathcal{W}_{n_i})\,d\mathcal{W}_{n_i}\end{aligned}$$

where "$'$" denotes appropriate differentiation. Calling upon generalized derivatives, one can write

$$\frac{d}{d\mathcal{W}_{x_i}}\mathcal{I}_{\{|\mathcal{W}_{x_i}|\le T\}} = \delta(\mathcal{W}_{x_i} + T) - \delta(\mathcal{W}_{x_i} - T)$$

with $\delta(\cdot)$ denoting the Dirac impulse, and as a result, we obtain

$$\begin{aligned}\int g_T'(\mathcal{W}_{n_i} + \mathcal{W}_{s_i})\phi(\mathcal{W}_{n_i})\,d\mathcal{W}_{n_i} &= -\int \mathcal{I}_{\{|\mathcal{W}_{x_i}|\le T\}}\phi(\mathcal{W}_{n_i})\,d\mathcal{W}_{n_i} \\ &\quad + T(\phi(T - \mathcal{W}_{s_i}) + \phi(-T - \mathcal{W}_{s_i})).\end{aligned}$$

Substituting the above expressions back into (12) yields

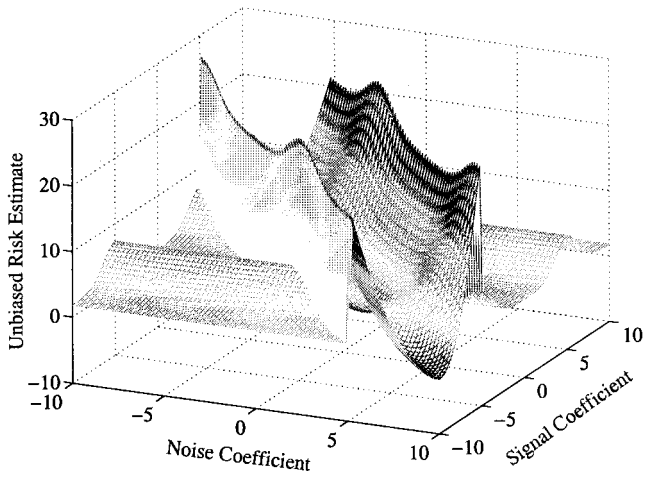
$$\begin{aligned}\mathrm{E}\left\{\sum_{i=1}^{N}[\gamma_T(\mathcal{W}_{x_i}) - \mathcal{W}_{s_i}]^2\right\} &= \mathrm{E}\{\mathcal{R}_B(\boldsymbol{s}, T)\} \\ &\quad + 2T\sigma^2\sum_{i=1}^{N}[\phi(T - \mathcal{W}_{s_i}) + \phi(-T - \mathcal{W}_{s_i})].\end{aligned}$$ □

We now define the risk estimator $\mathcal{R}_U(\boldsymbol{s}, T) = \mathcal{R}_B(\boldsymbol{s}, T) + \mu$, which Theorem 1 has shown to be unbiased. Henceforth, we typically refer to $\mathcal{R}_U(\boldsymbol{s}, T)$ as the unbiased risk for short. Theorem 1 also proves that the expected value of the suboptimal estimator $\mathcal{R}_B(\boldsymbol{s}, T)$ is a lower bound on the mean-square error. The estimator is biased because we have assumed that the magnitudes of the signal components are always above T in (9). Since we did not account for the errors due to an erroneous decision around T, we see that a coefficient composed of both signal and noise components may be present below the threshold T, when the signal contribution should have set it above T.

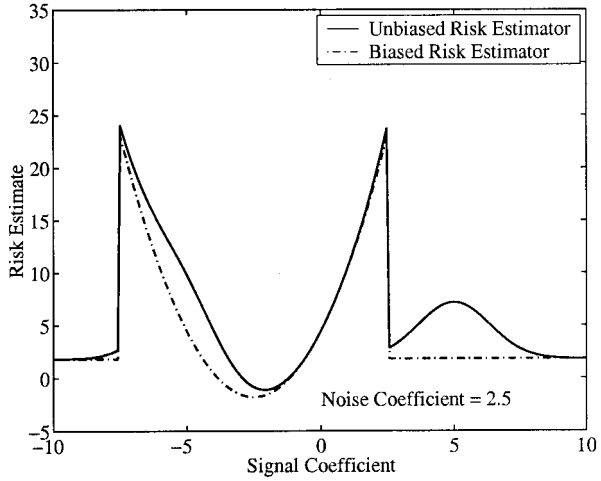
B. Interpretive Analysis of the Risk Bias

The biased risk $\mathcal{R}_B(\boldsymbol{s}, T)$ is clearly different from the optimal or unbiased risk, and the significance of this difference will be dependent upon T and $\{s[m]\}$. Heuristically, this difference is due to the naive and perhaps optimistic rule which attributes any coefficient below T to noise and any coefficient above T to the underlying signal. In short, a noisy signal coefficient can be less than or equal to T depending on its local energy and how it is modified by the noise, regardless of the noise-free coefficient. The nature of the underlying signal in the presence of noise at a level around the threshold T is therefore very relevant. Recall that T is solely determined by the noise variance and the length of the observation interval, N.

A first-order evaluation of the unbiased risk $\mathcal{R}_U(\boldsymbol{s}, T)$can be graphically performed by considering its variation with a single signal coefficient $\mathcal{W}_{s_i}^p$ and a single noise coefficient $\mathcal{W}_{n_i}^p$. Fig. 2(a) shows the resulting plot. The discontinuities in the risk estimate occur along the two 45° lines, $|\mathcal{W}_{n_i}^p + \mathcal{W}_{s_i}^p| = T$. For clarity, a cross section of this plot is shown in Fig. 2(b) for the case $\mathcal{W}_{n_i}^p = 2.5$, and the biased risk is included for comparison purposes. Note that both risks are asymptotically constant, since all of the errors up to T have been accounted for and since any component above T is considered to correspond to the underlying signal. As Fig. 2(b) demonstrates, the biased risk is a fairly good approximation to the unbiased risk in the regions where $|\mathcal{W}_{s_i}^p|$ is away from T (in this case $T = 5$). Note also that we are only plotting our estimate of the risk, so that for some noise realizations (e.g., a noise coefficient of 2.5) the risk estimate may be negative. In the remainder of this section and later in Section III-E, we will quantify the significance of the bias term μ.



(a)



(b)

Fig. 2. (a) Plot of the unbiased risk estimator as a function of a single signal coefficient and a single noise coefficient. (b) Cross section of the plot in (a) that shows the risk estimator as a function of a single signal coefficient and a noise coefficient equal to 2.5.

To better understand the effects of the bias term, we rewrite (11) as

$$\mathcal{R}(\boldsymbol{s}, T) - \mathrm{E}\{\mathcal{R}_B(\boldsymbol{s}, T)\} = 2T\sigma^2 \sum_{i=1}^{N} [\phi(T - \mathcal{W}_{s_i}^p) + \phi(-T - \mathcal{W}_{s_i}^p)]$$

$$\frac{\mathcal{R}(\boldsymbol{s}, T) - \mathrm{E}\{\mathcal{R}_B(\boldsymbol{s}, T)\}}{N\sigma^2} = 2T \frac{1}{N} \sum_{i=1}^{N} [\phi(T - \mathcal{W}_{s_i}^p) + \phi(-T - \mathcal{W}_{s_i}^p)] \quad (13)$$

$$= 2T \sum_{\mathcal{W}_{s_i}^p} [\phi(T - \mathcal{W}_{s_i}^p) + \phi(-T - \mathcal{W}_{s_i}^p)] h(\mathcal{W}_{s_i}^p). \quad (14)$$

In (14), $h(\cdot)$ represents the normalized histogram of the signal coefficients (i.e., $\sum_{\mathcal{W}_{s_i}^p} h(\mathcal{W}_{s_i}^p) = 1$). In this form, it is more apparent how the bias term is related to the underlying distribution of the signal coefficients. In addition, (14) has been

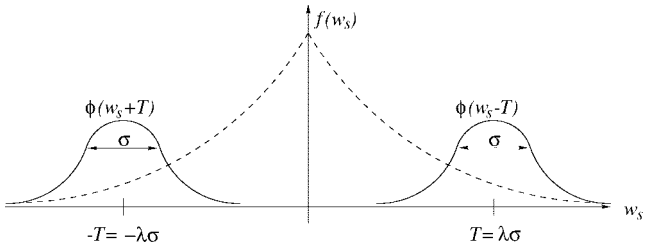


Fig. 3. Graphical illustration of the significance of the bias term as a function of σ and the underlying distribution of the signal coefficients.

normalized by the value $N\sigma^2$, which corresponds to the noise energy $\mathrm{E}\{\|\boldsymbol{n}\|^2\}$. To indicate the dependence of the bias on the noise variance, we let $T = \lambda\sigma$

$$\tilde{\mu}(\boldsymbol{s}, \sigma) = \frac{\mathcal{R}(\boldsymbol{s}, T) - \mathrm{E}\{\mathcal{R}_B(\boldsymbol{s}, T)\}}{N\sigma^2} = 2\lambda\sigma \sum_{\mathcal{W}_{s_i}^p} [\phi(\lambda\sigma - \mathcal{W}_{s_i}^p) + \phi(-\lambda\sigma - \mathcal{W}_{s_i}^p)] h(\mathcal{W}_{s_i}^p). \quad (15)$$

This parameterization is also useful, as we will later demonstrate that $\lambda = \sqrt{2 \log N}$ may not necessarily provide the optimal threshold, when an adaptive basis is used.

From (15), the value of $\tilde{\mu}(\boldsymbol{s}, \sigma)/2\lambda\sigma$ is composed of two shifted Gaussian functions weighted by the histogram of the signal coefficients. Fig. 3 graphically illustrates these two components (a continuous PDF $f(\mathcal{W}_{s_i}^p)$ is shown for graphical clarity). The illustration shows that the threshold T and the histogram of the signal coefficients will determine how well $\mathcal{R}_B(\boldsymbol{s}, T)$ approximates $\mathcal{R}(\boldsymbol{s}, T)$. The plot, however, does not provide insight about the bias term as a function of these parameters. A more formal and quantitative assessment of these factors will be provided in Section III-E.

Some insight can nevertheless be obtained by finding a bound on the bias and by analyzing its asymptotic properties. For a crude approximation, note that $\phi(\cdot) \leq 1/\sqrt{2\pi\sigma^2}$, and, consequently,

$$\tilde{\mu}(\boldsymbol{s}, \sigma) \leq 2\lambda\sigma \frac{2}{\sqrt{2\pi}} \frac{1}{\sigma} \sum_{\mathcal{W}_{s_i}^p} h(\mathcal{W}_{s_i}^p) = \frac{2\sqrt{2}}{\sqrt{\pi}} \lambda. \quad (16)$$

For the case $\lambda = \sqrt{2 \log N}$, an upper bound is then given by

$$\boxed{\tilde{\mu}(\boldsymbol{s}, \sigma) \leq \frac{4\sqrt{\log N}}{\sqrt{\pi}}} \quad (17)$$

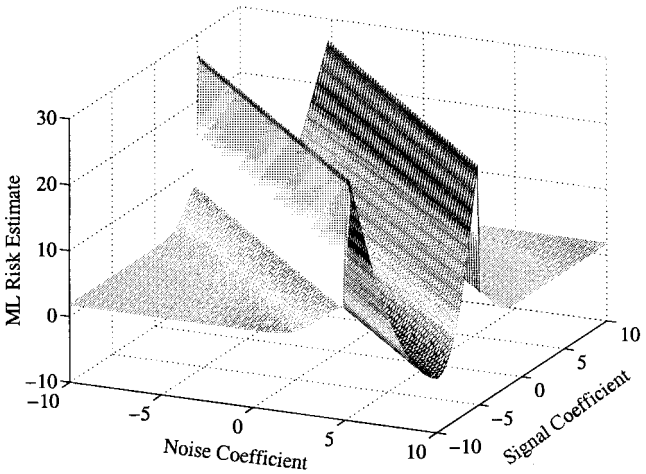
which is an increasing function of the signal length. Starting with the expression given in (15), we also evaluate the following asymptotic cases:

a) letting σ approach 0

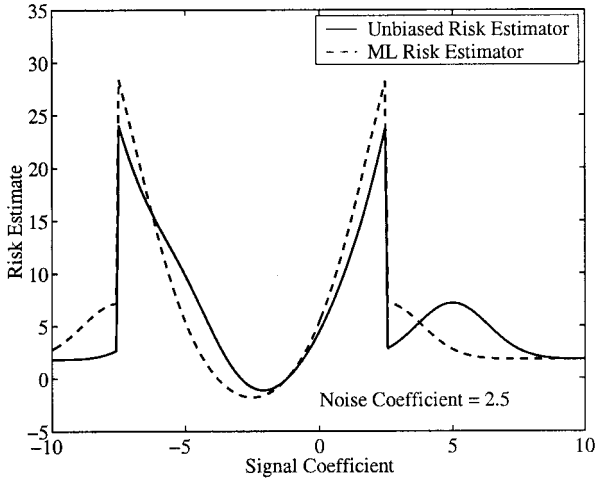
$$\tilde{\mu}_0(\boldsymbol{s}, \sigma) = \lim_{\sigma \to 0} \tilde{\mu}(\boldsymbol{s}, \sigma) = 0; \quad (18)$$

b) while letting σ approach ∞ results in

$$\tilde{\mu}_\infty(\boldsymbol{s}, \sigma) = \lim_{\sigma \to \infty} \tilde{\mu}(\boldsymbol{s}, \sigma) = \frac{2\sqrt{2}\,\lambda}{\sqrt{\pi}} e^{-\lambda^2/2}. \quad (19)$$



(a)



(b)

Fig. 4. (a) Plot of the MLE risk estimator as a function of a single signal coefficient and a single noise coefficient. (b) Cross section of the plot in (a) that shows the MLE risk estimator as a function of a single signal coefficient and a noise coefficient equal to 2.5.

For $\lambda = \sqrt{2 \log N}$, the asymptotic value then becomes

$$\boxed{\tilde{\mu}_{\infty}(\boldsymbol{s}, \sigma) = \frac{4\sqrt{\log N}}{N\sqrt{\pi}}.} \tag{20}$$

This approximation and qualitative analysis shows that independent of the distribution of the signal coefficients, a crude upper bound and asymptotic values of the risk bias may be obtained. The examples provided in Section III-D will show that the underlying distribution determines the "shape" of the risk bias.

C. Maximum-Likelihood Estimate of the Risk Bias

Note that the bias term in (11) assumes prior knowledge of the signal coefficients, and as a result, no true unbiased estimator can be achieved in practice. This difficulty, however, can be partially lifted by picking the maximum-likelihood estimate (MLE) (in this case, the MLE of the signal coefficient is the noisy coefficient [1]) to obtain an upper bound on the bias. The MLE of the bias term is then given by

$$\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma) = 2T\sigma^2 \sum_{i=1}^{N} [\phi(T - \mathcal{W}_{x_i}^p) + \phi(-T - \mathcal{W}_{x_i}^p)]. \tag{21}$$

For notational convenience, we let $\mathcal{R}_{\mathrm{ML}}(\boldsymbol{s}, T)$ denote the risk $\mathcal{R}_B(\boldsymbol{s}, T) + \mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)$, even though $\mathcal{R}_{\mathrm{ML}}(\boldsymbol{s}, T)$ is not the MLE of the true risk. To illustrate the function given in (21), we plot this risk estomatorfor the scalar case. Fig. 4(a) and (b) shows plots similar to those previously shown in Fig. 2(a) and (b). Fig. 4(a) shows that the risk estimator $\mathcal{R}_{\mathrm{ML}}(\boldsymbol{s}, T))$ is symmetric about the 45° axis in the x–y plane, and Fig. 4(b) compares $\mathcal{R}_{\mathrm{ML}}(\boldsymbol{s}, T)$ to $\mathcal{R}_U(\boldsymbol{s}, T)$.

One problem with using the MLE to estimate $\mu(\boldsymbol{s}, \sigma)$ is that it is a biased estimator. To determine the significance of this bias, we compute the expected value of $\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)$

$$\begin{aligned}\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\} = 2T\sigma^2 \sum_{i=1}^{N} [&\mathrm{E}\{\phi((T - \mathcal{W}_{s_i}^p) - \mathcal{W}_{n_i}^p)\} \\ &+ \mathrm{E}\{\phi((-T - \mathcal{W}_{s_i}^p) - \mathcal{W}_{n_i}^p)\}].\end{aligned} \tag{22}$$

To proceed, we must evaluate $\mathrm{E}\{\phi(y - \mathcal{W}_{n_i}^p)\}$

$$\begin{aligned}\mathrm{E}\{\phi(y - \mathcal{W}_{n_i}^p)\} &= \int_{-\infty}^{\infty} \phi(y - \mathcal{W}_{n_i}^p)\phi(\mathcal{W}_{n_i}^p)\, d\mathcal{W}_{n_i}^p \\ &= \frac{1}{2\pi\sigma^2} \\ &\quad \cdot \int_{-\infty}^{\infty} e^{-(1/2\sigma^2)[(y - \mathcal{W}_{n_i}^p)^2 + (\mathcal{W}_{n_i}^p)^2]}\, d\mathcal{W}_{n_i}^p.\end{aligned}$$

Letting $x = \mathcal{W}_{n_i}^p - \frac{1}{2}y$, we obtain

$$\begin{aligned}\mathrm{E}\{\phi(y - \mathcal{W}_{n_i}^p)\} &= \frac{1}{\sqrt{4\pi\sigma^2}} e^{-(y^2/4\sigma^2)} \int_{-\infty}^{\infty} \phi_{\sigma^2/2}(x)\, dx \\ &= \phi_{2\sigma^2}(y).\end{aligned} \tag{23}$$

In this case, $\phi_{2\sigma^2}(y)$ corresponds to a Gaussian function with variance $2\sigma^2$ evaluated at y. The final result can then be written as

$$\begin{aligned}&\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\} \\ &\quad = 2T\sigma^2 \sum_{i=1}^{N} [\phi_{2\sigma^2}(T - \mathcal{W}_{s_i}^p) + \phi_{2\sigma^2}(-T - \mathcal{W}_{s_i}^p)] \qquad (24) \\ &\quad = 2T\sigma^2 \sum_{i=1}^{N} \left[\frac{1}{\sqrt{2}}\, e^{(T - \mathcal{W}_{s_i}^p)^2/4\sigma^2} \phi_{\sigma^2}(T - \mathcal{W}_{s_i}^p) \right. \\ &\qquad \left. + \frac{1}{\sqrt{2}}\, e^{(-T - \mathcal{W}_{s_i}^p)^2/4\sigma^2} \phi_{\sigma^2}(-T - \mathcal{W}_{s_i}^p)\right]. \qquad (25)\end{aligned}$$

The final equation shows that $\sqrt{2}\,\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\}$ is an upper bound for the true bias. Equation (24) is useful because it shows that the only distinction between $\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\}$ and $\mu(\boldsymbol{s}, \sigma)$ is the variance of the Gaussian functions. As a result, the insight obtained from examining the value of $\mu(\boldsymbol{s}, \sigma)$ is directly applicable to understanding $\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\}$.

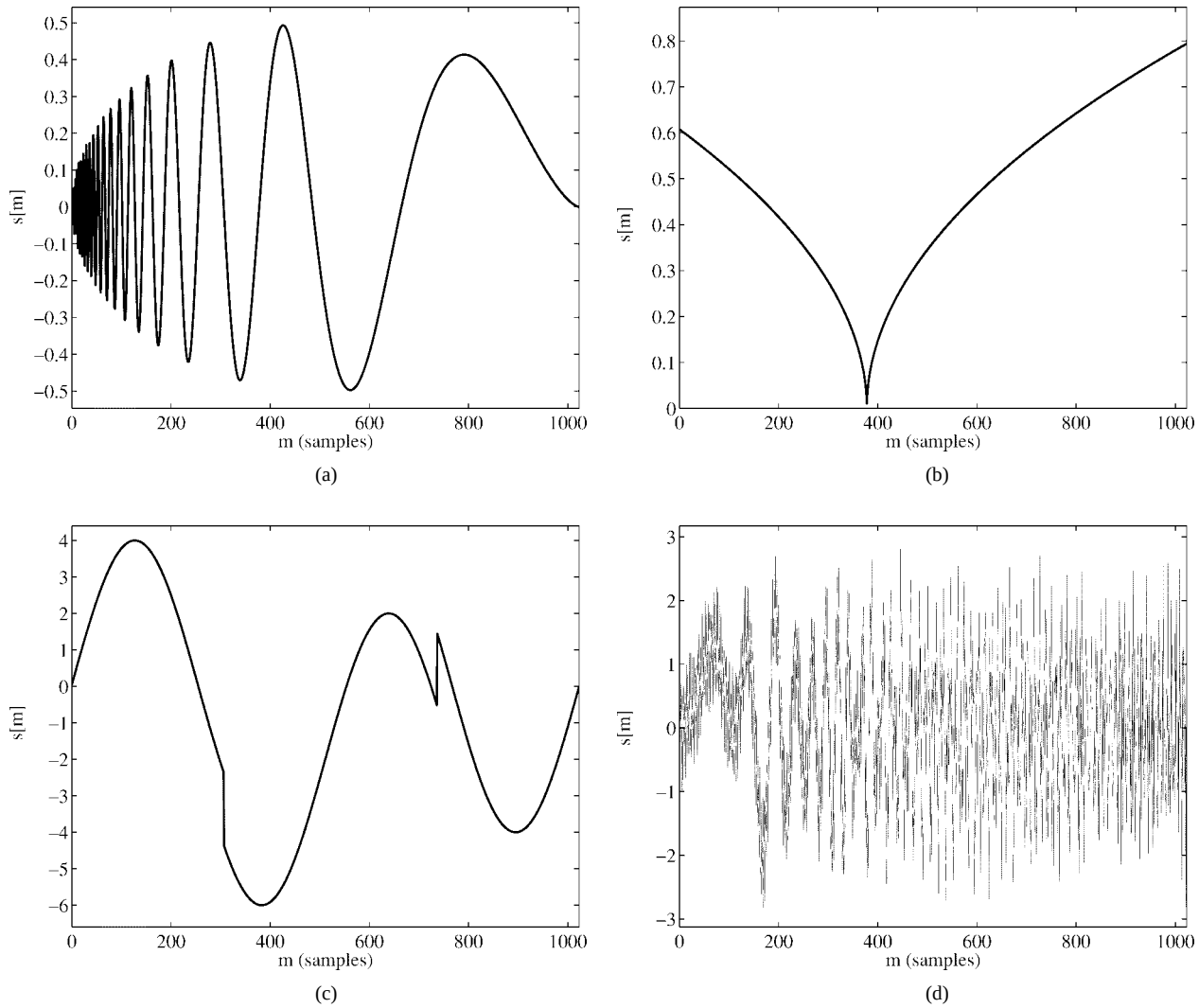


Fig. 5. The synthetic signals considered in this numerical experiment. (a) Doppler. (b) Cusp. (c) HeaviSine. (d) MishMash.

For completeness, we include the asymptotic values of $\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\}/N\sigma^2$

$$\boxed{\lim_{\sigma \to 0} \frac{\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\}}{N\sigma^2} = 0} \tag{26}$$

$$\boxed{\lim_{\sigma \to \infty} \frac{\mathrm{E}\{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)\}}{N\sigma^2} = \frac{2\lambda}{\sqrt{\pi}}\, e^{-\lambda^2/4}.} \tag{27}$$

D. Numerical Experiment

In this example, we continue the analysis of the risk bias by considering some specific numerical examples. The four synthetic signals considered here are shown in Fig. 5. The signals shown in Fig. 5(a)–(c) are well-represented in a wavelet basis, and consequently, a histogram of the wavelet coefficients for each of the signals is highly concentrated around zero. On the other hand, the more complex signal shown in Fig. 5(d) has wavelet coefficient values which are less concentrated around zero.

To illustrate the "shape" of the risk bias, the normalized biases $\tilde{\mu}(\boldsymbol{s}, \sigma)$ and $\tilde{\mu}_{\mathrm{ML}}(\boldsymbol{s}, \sigma)$ were computed for different values of σ and $\lambda = \sqrt{2 \log N}$, where

$$\tilde{\mu}(\boldsymbol{s}, \sigma) = \frac{\mu(\boldsymbol{s}, \sigma)}{N\sigma^2}$$
$$\tilde{\mu}_{\mathrm{ML}}(\boldsymbol{s}, \sigma) = \frac{\mu_{\mathrm{ML}}(\boldsymbol{s}, \sigma)}{N\sigma^2}.$$

The results are shown in Fig. 6 for the four signals of interest. We note that the asymptotic values are equal for all four signals because the value of λ is constant in this example, and since N is large, the asymptotic value $\tilde{\mu}_{\infty}(\boldsymbol{s}, \sigma)$ is quite small. One must remember, though, that Fig. 6 shows the normalized bias; therefore, the actual risk bias grows quadratically as a function of σ.

Comparing the plots shown in Fig. 6, we note that the major differences are in the "shape" of the different bias terms. The shape will, in fact, be dependent on the histogram of the underlying signal coefficients, since the locations of the local minima and maxima are functions of the coefficient values. These interesting features, however, only occur for very small values of σ, since the risk bias approaches its asymptotic value very rapidly. The intuitive reason for this is that as σ increases, the signal coefficients contribute less and less to the differences

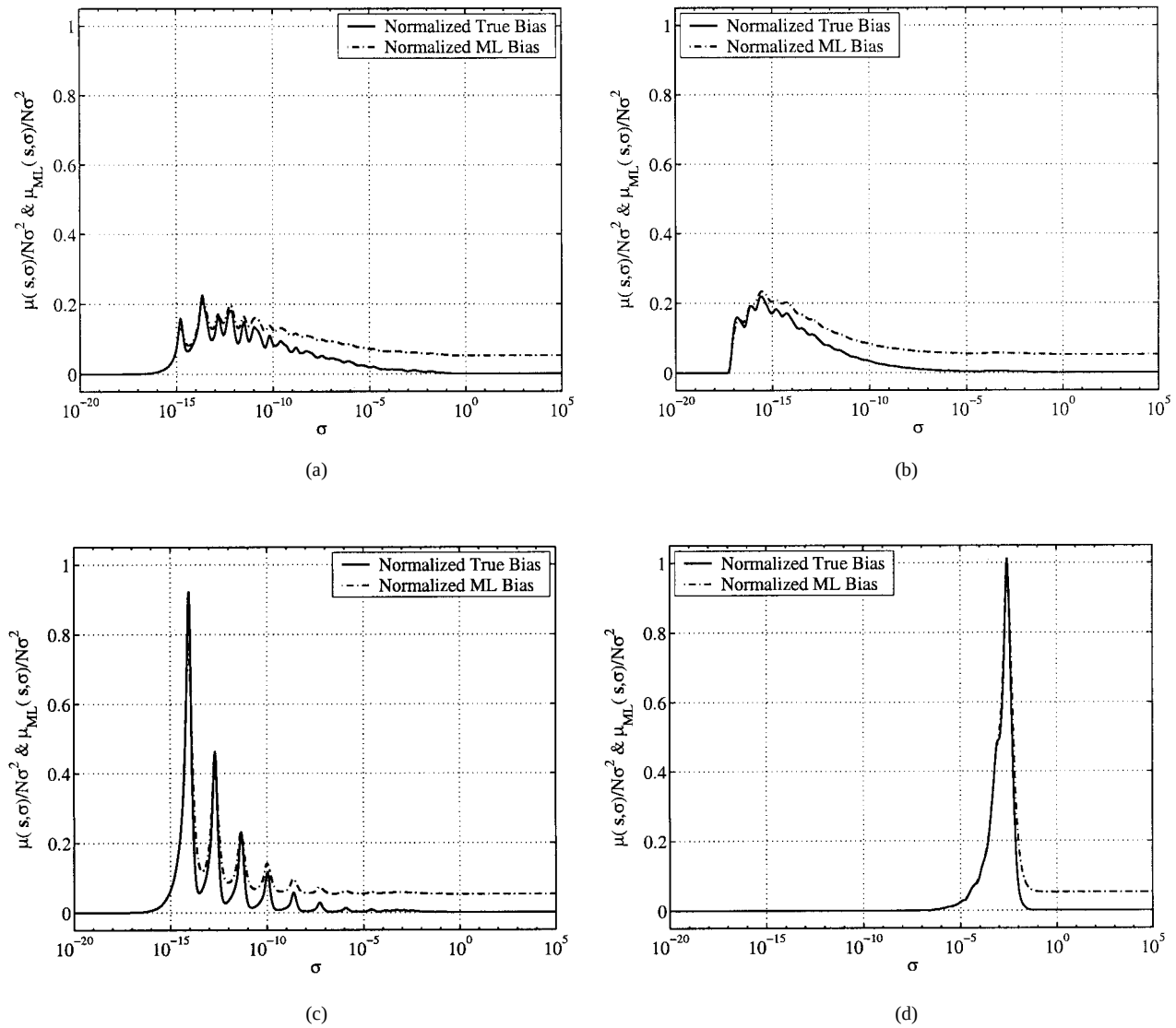


Fig. 6. Comparison of the risk bias and the maximum-likelihood estimate of the risk bias for different values of σ, the standard deviation of the contaminating noise. The signals shown in Fig. 5 were examined. (a) Doppler. (b) Cusp. (c) HeaviSine. (d) MishMash.

$(\lambda\sigma - \mathcal{W}_{s_i}^p)$ and $(-\lambda\sigma - \mathcal{W}_{s_i}^p)$ in (15). As a result, for large values of σ, the normalized histogram of the signal coefficients can be approximated by

$$h(\mathcal{W}_{s_i}^p) = \begin{cases} 1, & \mathcal{W}_{s_i}^p = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Using this histogram in (15) yields

$$\tilde{\mu}(\boldsymbol{s}, \sigma) = 4\lambda\sigma\phi(\lambda\sigma) = \frac{2\sqrt{2}\lambda}{\sqrt{\pi}}\, e^{-\lambda^2/2} \tag{29}$$

which is the asymptotic value $\tilde{\mu}_\infty(\boldsymbol{s}, \sigma)$ given in (19). This shows that the risk bias approaches its asymptotic value quickly as the signal coefficients become insignificant when compared to $\lambda\sigma$. As a consequence, the risk bias should exhibit this property when the histogram $h(\mathcal{W}_{s_i}^p)$ drops off rapidly as the magnitude of $\mathcal{W}_{s_i}^p$ increases. Comparing the results shown in Fig. 6, we note that the more complex MishMash signal does not approach its asymptotic value as rapidly as the other signals because its histogram has a slower rate of decay.

Fig. 6 also compares the maximum-likelihood estimate of the risk bias with the true risk bias. For this example, the asymptotic value $\lim_{\sigma\to\infty} \tilde{\mu}_{\mathrm{ML}}(\boldsymbol{s}, \sigma)$ is in fact larger than $\tilde{\mu}_\infty(\boldsymbol{s}, \sigma)$. We also note that the MLE of the risk bias is very close to the true bias for extremely small values of σ. This is understandable since $\mathcal{W}_{x_i}^p$ is a good approximation of $\mathcal{W}_{s_i}^p$ if $\mathcal{W}_{n_i}^p$ is almost zero. As σ increases, however, the two curves deviate and $\tilde{\mu}_{\mathrm{ML}}(\boldsymbol{s}, \sigma)$ approaches its asymptotic value. The examples included in Section IV-C will show how these differences in the risk bias will affect the search for the basis which produces the minimal reconstruction error.

E. Risk Optimality Dependence on Signal Statistics

A more rigorous and systematic analysis of the bias may be performed and its behavior quantified in terms of the signal statistics, if these were available. This Bayesian-like approach lets us use this prior knowledge about $\{s[m]\}$ to evaluate the significance of the bias term and to fully characterize it. As demonstrated below, a prior probability density $f(\mathcal{W}_{s_i}^p)$ for

the signal coefficients is shown to have a strong influence on the bias and thus plays a key role in the search for an optimal threshold T.

Proposition 1: Assume a probability density $f(\mathcal{W}_{s_i}^p)$ of the form

$$f(\mathcal{W}_{s_i}^p) = \epsilon f_1(\mathcal{W}_{s_i}^p) + (1-\epsilon) f_2(\mathcal{W}_{s_i}^p)$$

where $f_1(\mathcal{W}_{s_i}^p)$ is analytic and $f_2(\mathcal{W}_{s_i}^p)$ has a finite or countably infinite number of singularities (i.e., $f_2(\mathcal{W}_{s_i}^p) = \sum_{k=0}^{\infty} p_k \delta(\mathcal{W}_{s_i}^p - \nu_k)$). The expected value of the bias term, $\mu(\boldsymbol{s}, \sigma)$, is then given by

$$\begin{aligned} \mathrm{E}_s\{\mu(\boldsymbol{s}, \sigma)\} = 2T\sigma^2 N \Bigg[& \epsilon \sum_{j=0}^{\infty} \frac{\sigma^{2j}}{(2j)!} 1 \cdot 3 \cdots (2j-1) \\ & \cdot [f_1^{(2j)}(T) + f_1^{(2j)}(-T)] + (1-\epsilon) \\ & \cdot \sum_{k=0}^{\infty} p_k [\phi(T-\nu_k) + \phi(-T-\nu_k)] \Bigg]. \end{aligned} \tag{30}$$

The proof of the above proposition is included in the Appendix. Equation (30) shows that the bias term of the suboptimal risk is strongly dependent on T. This implies that the overall minimum of the true risk will be dependent on the *a priori* probability density $f(\cdot)$ (if available). The mode of the $\mathrm{E}_s\{\mu(\boldsymbol{s}, \sigma)\}$ will indeed determine the extremal point, and when combined with $\mathcal{R}_B(\boldsymbol{s}, T)$ will result *a posteriori* in a minimum at a corresponding "optimal" threshold T.

Illustrative Example: For illustration purposes, we numerically analyze the two risks $\mathcal{R}_B(\boldsymbol{s}, T)$ and $\mathcal{R}(\boldsymbol{s}, T)$ by considering a class of signals that are well-approximated by K coefficients of the orthonormal basis $\{\boldsymbol{W}_{x_i}^p\}_{1 \le i \le N}$. We associate to the inner products $\{\langle \boldsymbol{s}, \boldsymbol{W}_{x_i}^p \rangle\}$ a distribution density given by

$$f(\theta) = \frac{N-K}{N} \delta(\theta) + \frac{K}{N} h(\theta). \tag{31}$$

Out of N coefficients, there are an average of $N-K$ zero coefficients and K nonzero coefficients whose values are specified by $h(\theta)$. As the proportion K/N becomes smaller, the performance of the noise removal algorithm improves. Fig. 7 shows the mean-square error $E\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\}$ as a function of the threshold, for different values of K/N. For this example, we adjusted the parameters of $h(\theta)$ so that the total signal energy was equal to the total noise energy (i.e., a signal-to-noise ratio (SNR) of 0 dB). The minimum expected value of the unbiased risk is obtained for a value of T which is close to $\sqrt{2\sigma^2 \log N}$ (in this case $\sqrt{2\sigma^2 \log N} \approx 2.9$). However, the value of this optimal T does not remain invariant and is a function of K/N.

Fig. 8 compares the risk $\mathcal{R}(\boldsymbol{s}, T)$ with the expected error $E\{\mathcal{R}_B(\boldsymbol{s}, T)\}$ computed with our estimator. The precision of this lower bound increases when the proportion of nonzero coefficients K/N decreases. For small values of T the bias is very large but is considerably reduced at $T = \sqrt{2\sigma^2 \log N}$ which corresponds to the threshold we choose in our practical algorithm. For this threshold, the suboptimal error estimator provides a reasonable estimate of the mean-square error.

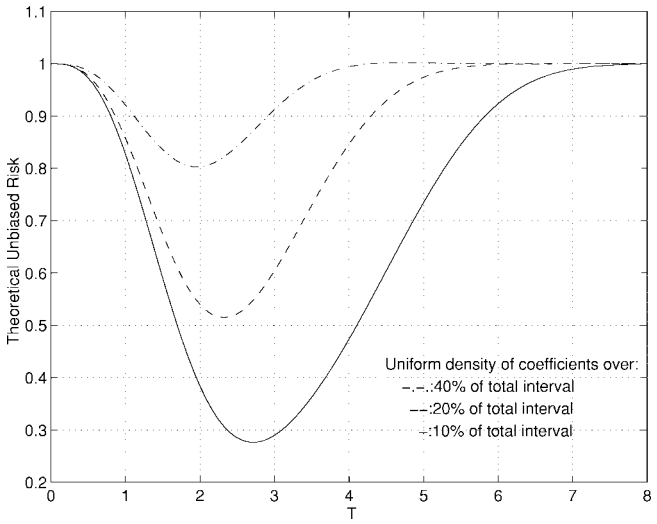


Fig. 7. Theoretical unbiased risk estimated for various cardinality ratios of signal/noise coefficients.

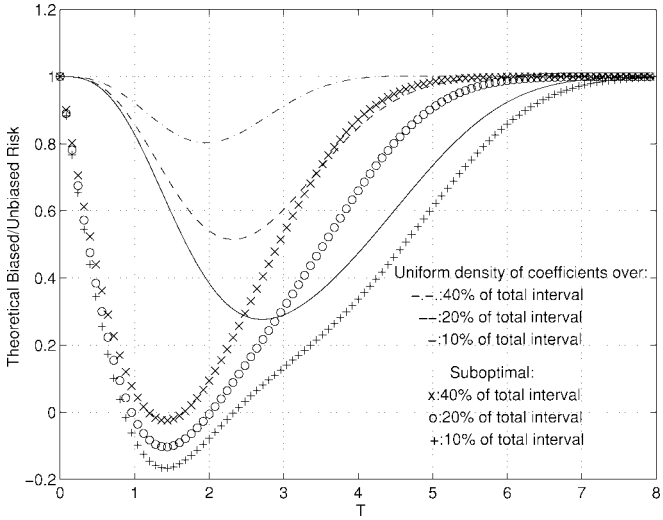


Fig. 8. Comparison of the biased and unbiased theoretical risks estimated for various cardinality ratios of signal/noise coefficients.

IV. Adaptive Signal Representation

A. Best Basis Search

When the signal possesses more complex features, one proceeds to search for the basis which would result in its most parsimonious representation. In searching for a wavelet packet or local cosine best basis, we typically have a dictionary $\mathcal{D}$ of possible bases, which for efficiency is endowed with a binary tree structure. Each node (j, j') (where $j \in \{0, \cdots, J\}$ represents the depth and $j' \in \{0, \cdots, 2^j - 1\}$ represents the branches on the jth level) of the tree then corresponds to a given orthonormal basis $\mathcal{B}_{j,j'}$ of a vector subspace of $\ell^2(\{1, \cdots, N\})$. Since a particular partition $p \in \mathcal{P}$ of $[0, 1]$ is composed of intervals $I_{j,j'} = [2^{-j}j', 2^{-j}(j'+1)]$, an orthonormal basis of $\ell^2(\{1, \cdots, N\})$ is given by

$$\mathcal{B}^p = \cup_{\{(j,j') | I_{j,j'} \in p\}} \mathcal{B}_{j,j'}.$$

By taking advantage of the property

$$\text{Span}\{\mathcal{B}_{j,\,j'}\} = \text{Span}\{\mathcal{B}_{j+1,\,2j'}\} \overset{\perp}{\oplus} \text{Span}\{\mathcal{B}_{j+1,\,2j'+1}\} \quad (32)$$

where $\oplus$ denotes a subspace direct sum, we associate to each node a cost $\mathcal{C}(\cdot)$. We can then perform a bottom-up comparison of children versus parent costs[4] and ultimately prune the tree.

Our goal is to choose the basis which leads to the best estimate $\{\hat{s}[m]\}$ among a collection of orthonormal bases $\{\mathcal{B}^p = \{\boldsymbol{W}^p_{x_i}\}_{1\le i\le N} | p \in \mathcal{P}\}$. In this analysis, we consider two particular classes of orthonormal bases. Trees of wavelet packet bases studied by Coifman and Wickerhauser [17] are constructed by quadrature mirror filter banks and comprise functions that are well-localized in time and frequency. This family of orthonormal bases divides the frequency axis into intervals of different sizes, with each set corresponding to a specific wavelet packet basis. Another family of orthonormal bases studied by Malvar [12], and Coifman and Meyer [2], can be constructed with a tree of windowed cosine functions, and correspond to a division of the time axis into intervals of dyadically varying sizes.

For a discrete signal of size N, one can show that a tree of wavelet packet bases or local cosine bases has $P = N(1 + \log_2 N)$ distinct vectors but includes more than $2^{N/2}$ different orthogonal bases. One can also show that the signal expansion in these bases is computed with algorithms that require $O(N \log_2 N)$ operations. Wickerhauser and Coifman [17] proposed that for any signal $\{f[m]\}$ and an appropriate functional $\mathcal{C}(\cdot)$, one finds the best basis $\mathcal{B}^{p_0}$ by minimizing an "additive" cost function

$$\text{Cost}\,(\boldsymbol{f},\, \mathcal{B}^p) = \sum_{i=1}^{N} \mathcal{C}(|\langle \boldsymbol{f},\, \boldsymbol{W}^p_{f_i}\rangle|^2) \quad (33)$$

over all bases. In this section, we select an expression for $\mathcal{C}(\cdot)$ so that $\text{Cost}\,(\boldsymbol{f},\, \mathcal{B}^p)$ approximates the mean-square error $\text{E}\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\}$ of the noise-removal algorithm. This expression corresponds to the estimator that was previously derived in Section III. As a result, the basis which results from minimizing this cost function corresponds to the "best" estimator of the underlying signal.

It was shown in (9) that $\text{E}\{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2\}$ can be estimated by

$$\text{Cost}\,(\boldsymbol{x},\, \mathcal{B}^p) = \sum_{i=1}^{N} \Phi(|\langle \boldsymbol{x},\, \boldsymbol{W}^p_{x_i}\rangle|^2). \quad (34)$$

This corresponds to an additive cost function and can therefore be efficiently minimized in a wavelet packet or local cosine dictionary. The best basis $\mathcal{B}^{p_0}$ for estimating $\{s[m]\}$ is then defined by

$$\text{Cost}\,(\boldsymbol{x},\, \mathcal{B}^{p_0}) = \min_{p\in\mathcal{P}} \text{Cost}\,(\boldsymbol{x},\, \mathcal{B}^p). \quad (35)$$

Some examples illustrating the performance of this estimator are given in Section IV-C.

[4]This in effect will eliminate the inadequate leaves of the tree.

B. Threshold Selection and Cost of Adaptivity

If we wish to adaptively choose a basis, we must use a higher threshold T than the threshold value $\sigma\sqrt{2\log N}$ used when the basis is set in advance. Indeed, an adaptive basis choice may also find vectors that better correlate the noise components. Let us consider the particular case $s[m] = 0$ for all m. To ensure that the estimated signal is close to zero, since $\boldsymbol{x} = \boldsymbol{n}$, we must choose a threshold T that has a high probability of being above all the inner products $\langle \boldsymbol{n},\, \boldsymbol{W}^p_{x_i}\rangle$ for all vectors in the dictionary $\mathcal{D}$. For a dictionary including P distinct vectors and P large, there is negligible probability for the noise coefficients to be above

$$T = \sigma\sqrt{2\log P}. \quad (36)$$

This threshold, however, is not optimal, and smaller values can improve the expected estimation error [11, p. 463].

In choosing an adaptive basis, it is also important to consider the costs associated with this adaptivity. An approximation in a basis adaptively selected is necessarily more precise than an approximation in a basis chosen *a priori*. However, in the presence of noise, estimations by thresholding may not be improved by an adaptive basis choice. Indeed, using a dictionary of several orthonormal bases requires raising the threshold, because the larger number of dictionary vectors allows possibly better correlation with the noise. The higher threshold removes more signal components, unless it is compensated by the adaptivity, which can better concentrate the signal energy over few coefficients. The same issue appears in parameterized models, where increasing the number of parameters may fit the noise as well as the data.

For example, if the original signal is piecewise-smooth, then a best wavelet packet basis does not concentrate the signal energy much more efficiently than a wavelet basis. In the presence of noise, in regions where the noise dominates the signal, the best basis algorithm optimizes the basis to fit the noise. This is why the threshold value must be increased. Hence, the resulting best basis estimation is not as precise as a thresholding in a fixed-wavelet basis with a lower threshold. On the other hand, for oscillatory signals, such as those considered in the next section, a best local cosine basis concentrates the signal energy over much fewer coefficients than a wavelet basis, and thus provides a better estimator [11, p. 464].

C. Numerical Experiment

In this example, we further analyze the risk estimators $\mathcal{R}_B(\boldsymbol{s},\, T)$, $\mathcal{R}_U(\boldsymbol{s},\, T)$, and $\mathcal{R}_{\text{ML}}(\boldsymbol{s},\, T)$. For comparison purposes, we will use the entropy cost function described in [3] and defined as

$$\mathcal{C}_{\text{Entropy}}(\boldsymbol{x}) = -\sum_{i=1}^{N} \tilde{\mathcal{W}}^p_{x_i} \log(\tilde{\mathcal{W}}^p_{x_i})$$

where

$$\tilde{\mathcal{W}}^p_{x_i} = \frac{|\mathcal{W}^p_{x_i}|^2}{\sum_{m=1}^{N} |x[m]|^2}.$$

Selecting a best basis by minimizing this function leads to a compact representation, where most of the signal energy is concentrated in a few coefficients. The cost functions that we have presented, however, will not necessarily lead to the most compact representation. The advantage of our approach is that a given basis has an associated cost that directly relates to the reconstruction error.

In this analysis, white Gaussian noise with variance σ^2 is added to a known signal at a specified SNR level, where SNR is defined as

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{m=1}^{N} |s[m]|^2}{N\sigma^2} \right).$$

Using one of the three risk estimators under consideration or the entropy cost given above, a best basis is obtained for the noisy signal by minimizing this cost (risk) in a dictionary of possible bases. Due to the nature of the signals we consider in this example, we have chosen to use a local cosine dictionary. The thresholding rule defined in (4) (for $T = \sigma\sqrt{2 \log P}$ and $P = N(1 + \log_2 N)$) is then applied to the coefficients, and a reconstructed or estimated signal is obtained by applying the appropriate inverse transformation. In this example, we focus on the performance of two real signals shown in Fig. 9. The first signal, shown in Fig. 9(a), corresponds to the voiced fricative /S/ in the word *Greasy*, and the second signal, shown in Fig. 9(b), corresponds to a bird chirp. Both of these signals possess high-frequency components; therefore, an adaptive basis should generate lower reconstruction errors than a normal wavelet decomposition.

To compare the performances of the estimators, the risk was computed through an average of 600 different noise realizations for 100 different SNR levels. Specifically, we computed

$$\hat{\mathcal{R}}(\boldsymbol{s}, \lambda\sigma) = \frac{1}{M} \sum_{j=1}^{M} \left[\frac{||\boldsymbol{s} - \hat{\boldsymbol{s}}_j||^2}{||\boldsymbol{s}||^2} \right] \tag{37}$$

where j is the index of the realization number and M is the number of realizations. This average risk was computed for different values of SNR. Note that the risk is normalized by $||\boldsymbol{s}||^2$ to allow comparisons between the risks corresponding to signals with different energies. Fig. 10(a) and (c) shows the results for the two signals considered here. Fig. 10(b) and (d) emphasizes the differences between the estimators by subtracting the unbiased risk from the risks associated with the other three estimators.

The risks of all four cost functions are very similar, with the risk associated with the optimal estimator being slightly smaller than the others. For the two signals considered here, the entropy cost function has slightly worse performance than the estimators we have presented. We note that this cost function essentially measures the parsimony of a given signal representation. In fact, in the theory of inequalities, there are a variety of criteria comparing the sparseness of the components of two vectors, with the entropy criterion being one of them [13]. The results corresponding to the entropy cost function in Fig. 10 show that the most compressed

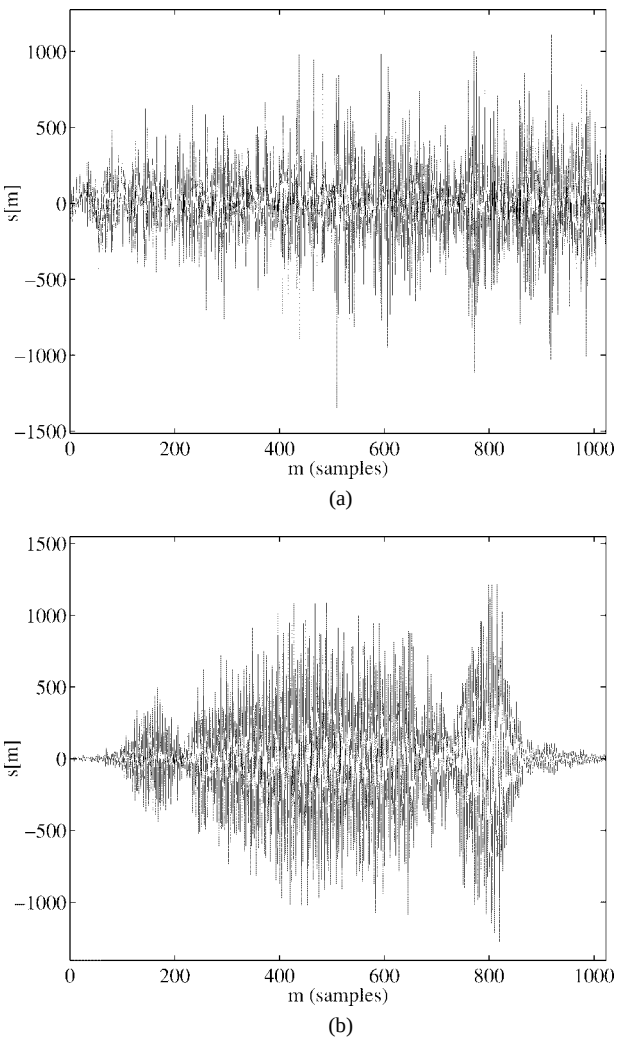


Fig. 9. Real signals used to illustrate the performances of the proposed cost functions. (a) Speech signal (/S/ in the word *Greasy*). (b) Bird chirp signal.

representation is certainly effective but does not guarantee that the reconstruction is minimal in the mean-square sense. We also note that the risk that uses the maximum-likelihood estimate of the bias gives slightly better performance than the biased risk for the *Chirp* signal. Trying to estimate the bias term with the MLE, in this case, appears to provide a more reliable estimate of the true risk than simply ignoring the bias term. The results, however, are exactly opposite for the *Greasy* signal. In this case, the biased estimator generates a lower risk than the ML estimator.

In this example, we have considered two real signals which possess high-frequency oscillations. To show that an adaptive basis is useful for these types of signals, we compare the previous results to those obtained by using a simple wavelet decomposition. Fig. 11(a) and (b) provides a comparison of the risks. The disparity in the risks demonstrates that, in this case, adaptivity is useful in reducing the mean-square error.

V. Conclusions

In this paper, we first used a simple-minded approach to propose a risk estimator, and subsequently showed this estimator to be biased. Comparing the biased and unbiased risks, we found that the risk bias was strongly dependent on

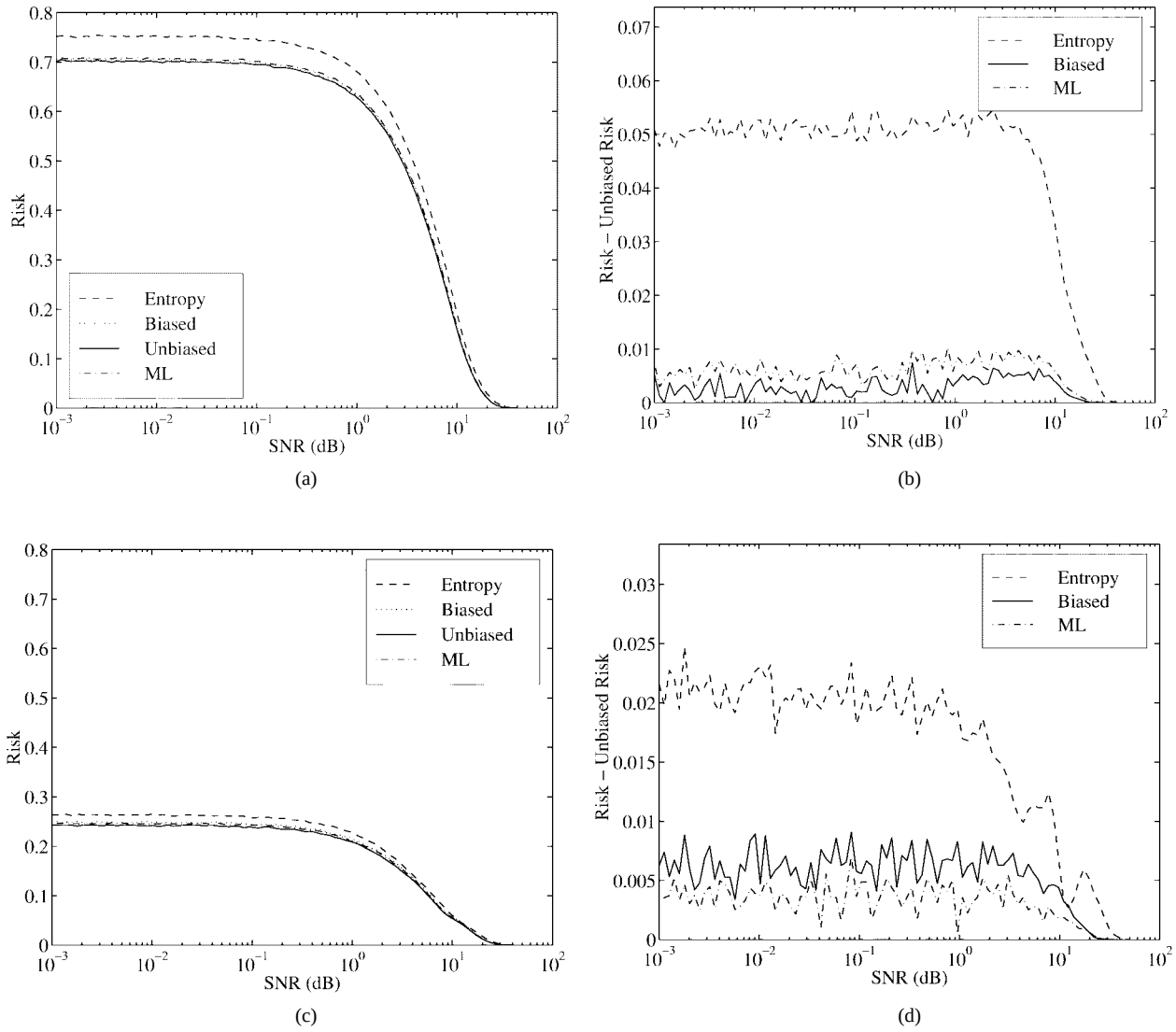


Fig. 10. Performances for the *Greasy* and *Chirp* signals as a function of SNR: (a) risk associated with the *Greasy* signal, (b) difference between the estimated risk and the unbiased risk for the *Greasy* signal, (c) risk associated with the *Chirp* signal, and (d) difference between the estimated risk and the unbiased risk for the *Chirp* signal.

the statistics of the underlying signal and the threshold T. We then used the proposed estimators to determine the wavelet basis which minimized the reconstruction error of a signal embedded in noise.

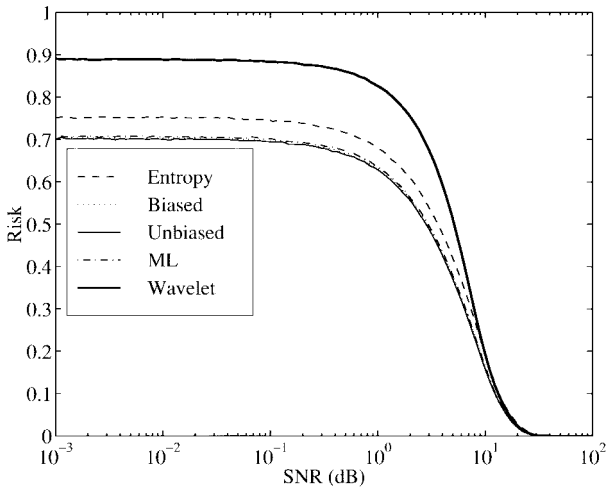
In this analysis, we adopted a thresholding strategy that removes coefficients which are purely or primarily noise. Previously, this thresholding strategy and the search for a "best" basis were unrelated. In our approach, the derived additive cost function accounts for the threshold T. By minimizing this cost, the proposed algorithm finds the best representation of the signal, so that discarding coefficients serves to improve signal quality.

The examples in Section IV-C were included to illustrate the performance of the proposed estimators. For real signals containing high-frequency oscillations, we argued that an adaptive signal representation, offered by wavelet packets or local cosine bases, provides more flexibility than a wavelet decomposition. This adaptivity allows "better" estimations to be made with respect to the risk criterion that we proposed.

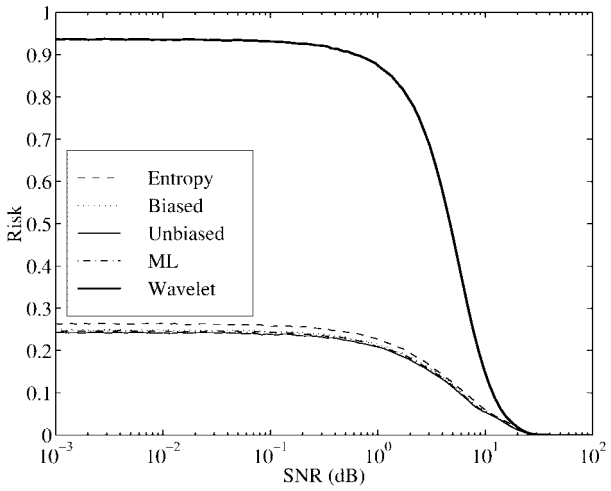
When an unbiased risk estimator is available for a given noise distribution, this analysis may be repeated using the established framework. This may be accomplished by finding an appropriate threshold level T and then using the resulting reconstruction error as a search criterion. Extending this approach to two-dimensional signals is not only interesting but challenging as well. For the one-dimensional case, we have assumed that the signal samples are independent. In images, however, the dependencies between neighboring pixels must be taken into account in order to produce quality reconstructions. Subsequent research will reveal how to properly extend this denoising procedure to two-dimensional signals.

APPENDIX
PROOF OF PROPOSITION 1

Proof: We assume that the wavelet coefficients of the underlying signal are identically distributed. The expected



(a)



(b)

Fig. 11. Comparison of the risks associated with an adaptive best basis search and a wavelet decomposition. (a) Risks associated with the *Greasy* signal. (b) Risks associated with the *Chirp* signal.

value of the bias term is then given by

$$\mathrm{E}_s\{\mu(\boldsymbol{s},\,\sigma)\} = 2T\sigma^2 \sum_{i=1}^{N} \int [\phi(T - \mathcal{W}_{s_i}^p) + \phi(-T - \mathcal{W}_{s_i}^p)] \cdot f(\mathcal{W}_{s_i}^p)\, d\mathcal{W}_{s_i}^p. \tag{38}$$

We will only consider densities of the following form, where $f(x)$ is the distribution for any $\mathcal{W}_{s_i}^p$

$$f(x) = \epsilon f_1(x) + (1-\epsilon) f_2(x).$$

In particular, $f_1(x)$ is infinitely differentiable, and $f_2(x)$ has a finite or countably infinite number of singularities. Since $f_1(x)$ is analytic, it can be represented by a Taylor series expansion, and $f_2(x)$ can be represented by

$$f_2(x) = \sum_{k=0}^{\infty} p_k \delta(x - \nu_k)$$

where

$$\sum_{k=0}^{\infty} p_k = 1.$$

As a result, $\mathrm{E}_s\{\mu(\boldsymbol{s},\,\sigma)\}$ can be separated into two expressions, one that is dependent on $f_1(x)$ and the other dependent on $f_2(x)$, or

$$\begin{aligned}\mathrm{E}_s\{\mu(\boldsymbol{s},\,\sigma)\} &= 2T\sigma^2 \sum_{i=1}^{N} \Bigg[\epsilon \int [\phi(T - \mathcal{W}_{s_i}^p) + \phi(-T - \mathcal{W}_{s_i}^p)] \\ &\qquad \cdot f_1(\mathcal{W}_{s_i}^p)\, d\mathcal{W}_{s_i}^p + (1-\epsilon) \int [\phi(T - \mathcal{W}_{s_i}^p) \\ &\qquad + \phi(-T - \mathcal{W}_{s_i}^p)] f_2(\mathcal{W}_{s_i}^p)\, d\mathcal{W}_{s_i}^p \Bigg].\end{aligned} \tag{39}$$

Given the similarity of the two terms $\phi(\cdot)$ in the first integral of (39), we only evaluate the first term. Letting $\tau_i = T - \mathcal{W}_{s_i}^p$, we obtain the Taylor series expansion of $f_1(T - \tau_i)$ around T

$$\int \phi(\tau_i) f_1(T - \tau_i)\, d\tau_i = \sum_{j=0}^{\infty} \int \frac{(-\tau_i)^j}{j!} f_1^{(j)}(T) \phi(\tau_i)\, d\tau_i. \tag{40}$$

This last expression is the sum of scaled moments of the Gaussian function, which are known to be [14]

$$m_j = \begin{cases} 1 \cdot 3 \cdots (j-1)\sigma^j, & j \text{ even} \\ 0, & j \text{ odd.} \end{cases} \tag{41}$$

The other term in the first integral of (39) leads to a similar expression. Evaluating the second integral for an arbitrary $x = \mathcal{W}_{s_i}^p$, gives

$$\begin{aligned}&\int [\phi(T-x) + \phi(-T-x)] f_2(x)\, dx \\ &= \int [\phi(T-x) + \phi(-T-x)] \left[\sum_{k=0}^{\infty} p_k \delta(x - \nu_k)\right] dx \\ &= \int \sum_{k=0}^{\infty} p_k [\phi(T-\nu_k)\delta(x-\nu_k) + \phi(-T-\nu_k)\delta(x-\nu_k)]\, dx \\ &= \sum_{k=0}^{\infty} p_k [\phi(T-\nu_k) + \phi(-T-\nu_k)].\end{aligned} \tag{42}$$

Combining the results of (39)–(42), we obtain an expression which proves the proposition. □

REFERENCES

[1] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics.* Oakland, CA: Holden-Day, 1977.
[2] R. Coifman and Y. Meyer, "Remarques sur l'analyse de Fourierá fenêtre," *C. R. Acad. Sci. Sér. I*, pp. 259–261, 1991.
[3] 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.
[4] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Com. Pure and Appl. Math.*, vol. XLI, pp. 909–996, 1988.
[5] D. Donoho and I. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases," *C. R. Acad. Sci. Paris*, Oct. 1994.
[6] ———, "Ideal spatial adaptation by wavelet shrinkage," Dept. Stat., Stanford Univ., June 1992, preprint.
[7] B. Gnedenko, "Sur la distribution limite du terme maximum d'une serie aleatoire," *Ann. Math.*, vol. 44, no. 3, pp. 423–453, July 1943.
[8] H. Krim, S. Mallat, D. Donoho, and A. Willsky, "Best basis algorithm for signal enhancement," in *IEEE, Int. Conf. Acoustics, Speech and Signal Processing ASSP'95* (Detroit, MI, May 1995).

[9] H. Krim and J.-C. Pesquet, "On the statistics of best bases criteria," in *Wavelets in Statistics of Lecture Notes in Statistics.* New York: Springer-Verlag, July 1995, pp. 193–207.
[10] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 674–693, July 1989.
[11] ———, *A Wavelet Tour of Signal Processing.* Boston, MA: Academic, 1998.
[12] H. Malvar, "Lapped transforms for efficient transform subband coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 969–978, June 1990.
[13] A. W. Marshall and I. Olkin, *Inequalities: Theory and Majorization and Its Applications.* New York: Academic, 1979.
[14] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984.
[15] S. I. Resnick, "Extreme values, regular variation and point processes," *Appl. Probab.*, 1987.
[16] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, 1981.
[17] M. V. Wickerhauser, "INRIA lectures on wavelet packet algorithms," *Ondelettes et Paquets D'ondelettes*, Roquencourt, France, June 17–21, 1991, pp. 31–99.