A New Information-Theoretic Approach to Signal Denoising and Best Basis Selection

Soosan Beheshti, *Member, IEEE,* and Munther A. Dahleh, *Fellow, IEEE*

***Abstract*—The problem of signal denoising with an orthogonal basis is considered. The existing approaches convert the considered problem into one of finding a threshold for estimates of basis coefficients. In this paper, a new solution to the denoising problem is proposed. The method is based on the description length of the noiseless data in subspaces of the bases. For each subspace, we estimate the desired description length and suggest choosing the subspace for which this quantity is minimized. We provide a method of probabilistically estimating the reconstruction error. This estimate is used for probabilistic validation of the desired description length.**

In existing thresholding methods, the optimum threshold is obtained as a function of the additive noise variance. In practical problems, where the noise variance is unknown, the first step is to estimate the noise variance. The estimated noise variance is then used in calculating the optimum threshold. Unlike such approaches, in the proposed method, the noise variance estimation and the signal denoising are done simultaneously.

***Index Terms*—Best basis, Shannon code, signal denoising, thresholding.**

I. Introduction

THE problem of estimating an unknown signal embedded in Gaussian noise has received a great deal of attention in numerous studies. The denoising process is to separate an observed data sequence into a "meaningful" signal and a remaining noise. One important approach is based on data projection on an orthogonal family of bases. The best representation of the noiseless signal is then chosen based on a proper choice of the associated nonzero bases. The choice of the denoising criterion depends on the properties of the additive noise, smoothness of the class of the underlying signal, and the selected signal estimator.

The pioneer method of wavelet denoising was first formalized by Donoho and Johnstone [5]. This wavelet thresholding method removes the additive noise by eliminating the basis coefficients with a small absolute value, which tend to be attributed to the noise. The method assumes a prior knowledge of the variance of the additive white Gaussian noise. Hard and soft thresholds are obtained by solving a minmax problem in the estimation of the expected value of the reconstruction error [4]. The suggested optimal threshold for the basis coefficient is $\sigma_w\sqrt{2\log(N)}$, where σ_w^2 is the additive noise variance, and N is the data length. The method is well adapted to signals that are approximately piecewise-smooth. The argument, however, fails for families of signals that are not smooth, i.e., the families of signals for which the noiseless coefficients might be nonzero, very small, and comparable with the noise effects, for a large number of basis functions.

The approach to the denoising problem in [7] proposes a thresholding method for any family of basis functions. Here, the challenge is to calculate the mean-square reconstruction error of the signal as a function of any given threshold. It provides estimates of such error for different families of basis functions such as wavelet and local cosine bases. The choice of the optimum threshold is given experimentally. The best basis search is based on the mean-square reconstruction error estimate. It demonstrates that for some class of signals, $\sigma_w\sqrt{2\log(N)}$ may not necessarily be the optimal wavelet threshold.

A different denoising approach is recommended in [8]. In each subspace of the basis functions, the normalized maximum likelihood (NML) of the noisy data is defined as the description length of the data. The minimum description length (MDL) denoising method suggests choosing the subspace that minimizes this description length. Here, noise is defined to be a part of the data that cannot be compressed with the considered basis functions, whereas the meaningful information-bearing signal need not be smooth. The method provides a threshold that is almost $1/\sqrt{2}$ of the suggested wavelet threshold in [5]: $\sigma_w\sqrt{\log(N)}$.

A new method of denoising is presented in this paper. The method is based on using a new information theoretic criterion that is the description length of the "noiseless" data. The description length of data is probabilistically validated for different subspaces of the bases. The method suggests choosing the subspace with minimum noiseless description length (MNDL). Since one of the main goals of this approach is to extract the most information from the noisy data, the method does not provide a threshold before estimating all the basis coefficients. Therefore, the subspace comparison can provide a threshold that is a function of the noisy data, the noise variance, and the basis coefficients. One advantage of the method is that the noiseless signal need not be smooth, i.e., need not be represented by a few elements of the basis family. It is important to mention that in the process of estimating the new description length, the reconstruction error is estimated probabilistically.

The existing thresholding methods provide thresholds that are functions of the additive noise variance. However, in most practical problems, the variance of the noise is unknown. The noise

Manuscript received February 23, 2004; revised November 10, 2004. This work was supported by MURI 0205-G-CB222. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fulvio Gini.

S. Beheshti is with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3 Canada (e-mail: soosan@ee.ryerson.ca).

M. A. Dahleh is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dahleh@mit.edu).

Digital Object Identifier 10.1109/TSP.2005.855075

variance is estimated by using some of the fine coefficients that are considered to be the noise coefficients [4], [8]. The variance estimate is then applied to provide the optimum threshold.

In the new approach, the data description length in each subspace is a function of the given data and the noise variance. To estimate the noise variance, we suggest choosing the noise variance and the subspace for which the description length of the data is minimum. Therefore, unlike the existing denoising methods, the estimation of the noise variance and the signal denoising are not two separate procedures.

In this paper, we consider additive Gaussian noise and orthogonal basis. However, unlike the existing thresholding methods, the new approach can be expanded and used for additive non-Gaussian noise and a complete nonorthogonal basis [6], [2].

The paper is organized as follows. Section II describes the considered denoising problem. Section III briefly describes the existing thresholding methods. Section IV introduces the new information theoretic approach. Sections V–VII focus on the reconstruction error and provide a new method of estimation of this error using the observed data. Sections VIII and IX finalize the discussions on subspace comparison, unknown noise variance, and best basis search. Section X provides the simulation results, and Section XI is the conclusion.

II. Denoising Problem and Subspace Selection

The unknown noiseless data $\bar{y}$ is corrupted by an additive white Gaussian noise w. The noisy data y of length N is available

$$y(n) = \bar{y}(n) + w(n) \tag{1}$$

for $i = 1, 2, \ldots, N$.

The additive noise $w(n)$ is a sample of the zero mean random variable $W(n)$ with variance σ_w^2. Data denoising is achieved by using an orthogonal basis that approximates the noiseless data with fewer nonzero coefficients than the length of the data.

Assume that the noiseless data belongs to the space $S_N, \bar{y}^N \in S_N$ (for example, if the elements of $\bar{y}$ are real, one choice is $S_N = R^N$). The orthonormal basis vectors $s_1, s_2, \ldots, s_N$ span the space S_N and are such that

$$\langle s_i, s_j \rangle = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \tag{2}$$

where $\langle a, b \rangle$ is the inner product of vectors a and b. The noiseless data is represented by this basis as follows:

$$\bar{y}^N = \sum_{i=1}^{N} \theta^*(i) s_i \tag{3}$$

where $\bar{y}^N = [\bar{y}(1), \bar{y}(2), \ldots, \bar{y}(N)]^T$, and $\theta^*(i)$ is the ith coefficient of the noiseless data.

The least square error estimate of the ith basis coefficient using the observed data (in (1)) is

$$\lambda(i) = \langle s_i, y^N \rangle \tag{4}$$

$$= \theta^*(i) + \langle s_i, w^N \rangle \tag{5}$$

where $y^N = [y(1), y(2), \ldots, y(N)]^T$, and $w^N = [w(1), w(2), \ldots, w(N)]^T$. If the noiseless signal has very few nonzero coefficients, there exists a large number of basis vectors for which $\theta^*(i) = 0$. For these bases, the projection of the noisy signal $\lambda(i) = \langle s_i, w^N \rangle$ is a sample of a zero mean Gaussian random variable with variance σ_w^2. Therefore, if the variance is small, this projection is usually very small with a high probability.

The main challenge of signal denoising is how to decide which of the estimated coefficients $\lambda(i)$s should be ignored (set to zero) and which of them should be used to represent the noiseless data. We restate the denoising problem by using a subspace selection method: Consider S_m, which is a subspace of S_N that is spanned by m elements of the basis. The estimate of noiseless data in this subspace is

$$\hat{y}_{S_m}^N = \sum_{i=1}^{N} \hat{\theta}_{S_m}(i) s_i \tag{6}$$

where for the estimated coefficient $\hat{\theta}_{S_m}(i)$, we have

$$\hat{\theta}_{S_m}(i) = \begin{cases} \lambda(i), & \text{if } s_i \in S_m \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The denoising question is which subspace S_m(and, therefore, which $\hat{y}_{S_m}^N \in S_m$) should be chosen to best represent the noiseless data.

Two important elements in analyzing the denoising problem are the following errors:

$$\textbf{Data error}: x_{S_m} = \frac{1}{N} \| y^N - \hat{y}_{S_m}^N \|_2^2 \tag{8}$$

$$\textbf{Reconstruction error}: z_{S_m} = \frac{1}{N} \| \bar{y}^N - \hat{y}_{S_m}^N \|_2^2. \tag{9}$$

The (noisy) data error x_{S_m} is the distance between the noisy observed data and its projection on subspace S_m. This error is available for each subspace. However, the noiseless data error z_{S_m} (reconstruction error) is not available since it is a function of the unknown noiseless data. This error is a desired criterion in both the existing denoising methods and the new method, which is proposed in this paper. In Section VI, we provide a novel method for estimating this error. The objective is to use the knowledge about the additive noise and the observed data to provide bounds on this error. In the next section, the importance of the reconstruction error in existing thresholding methods is reviewed.

III. Existing Thresholding Methods

In the hard thresholding method, a threshold τ is used for the coefficient estimates $\lambda(i)$s.[1] The coefficient estimates with absolute value less than the threshold are ignored

$$\hat{\theta}^{\mathrm{H}}(i) = \begin{cases} \lambda(i), & \text{if } |\lambda(i)| \geq \tau \\ 0, & \text{if } |\lambda(i)| < \tau. \end{cases} \tag{10}$$

This thresholding method is equivalent to the following subspace selection approach: Given a threshold τ, a subspace $S_{m(\tau)}$

[1] In soft thresholding, the coefficients estimates are

$$\hat{\theta}^{\mathrm{S}}(i) = \begin{cases} \mathrm{sgn}(\lambda(i))|\lambda(i)| - \tau, & \text{if } |\lambda(i)| \geq \tau \\ 0, & \text{if } |\lambda(i)| < \tau. \end{cases}$$

This method is not discussed here.

is selected, where $m(\tau)$ is the number of nonzero coefficients in vector $\hat{\theta}^{\mathrm{H}}$, and $S_{m(\tau)}$ is a subspace of S_N for which

$$s_i \in S_{m(\tau)}, \quad \text{iff } \hat{\theta}^{\mathrm{H}}(i) \neq 0. \tag{11}$$

Therefore, for any chosen τ, a subspace is selected

$$\tau \longrightarrow S_{m(\tau)}. \tag{12}$$

The main concern in thresholding is how to choose a proper threshold and, therefore, the optimum subspace. One important factor in the thresholding methods has been the mean square reconstruction error ([5], [7])

$$\mathrm{E}(Z_{S_m}) = \frac{1}{N} E\left(\|\bar{y}^N - \hat{Y}^N_{S_m}\|_2^2\right). \tag{13}$$

In the existing approaches, it is desired to estimate this error for different thresholds and choose the threshold that minimizes this desired criterion [10]. The development in [5] estimates an upper bound for this error and provides an optimum threshold based on analysis of this upper bound. In this paper, we provide probabilistic upper and lower bounds on this error.

IV. New Information Theoretic Denoising Method

Inspired by the Kolmogorov complexity and the notion of minimal description length for a data string, we intend to search for the subspace S_m that can provide the minimum description length (DL) of the data.

The probability density function of the noisy data in (1) is

$$f(y^N; \bar{y}^N) = \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\|y^N - \bar{y}^N\|_2^2/2\sigma_w^2} \tag{14}$$

where y^N is a sample of random variable Y^N. For g^N, which is any sample of random variable Y^N, the Shannon code is used. Therefore, the codelength of the binary prefix code is

$$\mathrm{DL}(g^N; \bar{y}^N) = -\frac{1}{N}\log_2(f(g^N; \bar{y}^N)) \tag{15}$$

$$= \log_2\sqrt{2\pi\sigma_w^2} + \frac{\|g^N - \bar{y}^N\|_2^2}{2\sigma_w^2 N}\log_2 e. \tag{16}$$

This denotes the DL of g^N when it is described by the noiseless data $\bar{y}^N$.

In each subspace S_m, the best representative of the noiseless data is $\hat{y}^N_{S_m}$ in (6). For the random variable generated by this representative of $\bar{y}^N$, Shannon code is used. Therefore, the codelength of g^N, using this representation, is

$$\mathrm{DL}(g^N; \hat{y}^N_{S_m}) = \log_2\sqrt{2\pi\sigma_w^2} + \frac{\|g^N - \hat{y}^N_{S_m}\|_2^2}{2\sigma_w^2 N}\log_2 e. \tag{17}$$

In existing information theoretic approaches such as normalized MDL (NMDL), the goal is to compare the codelength of the noisy data by using the estimate $\hat{y}^N_{S_m}$ in different subspaces [3], [8]

$$\mathrm{DL}(y^N; \hat{y}^N_{S_m}) = \log_2\sqrt{2\pi\sigma_w^2} + \frac{\log_2 e}{2\sigma_w^2} x_{S_m}. \tag{18}$$

Note that since the noisy data is observed, the data error x_{S_m} in (8) is also available. For nested subspaces of different order, the data error x_{S_m} is a decreasing function of order m and is zero in S_N. Therefore, comparison and minimization of this codelength for a set of nested subspaces always leads to choosing the subspace with highest order S_N.[2]

We believe comparison of this error fails since the noisy data is used to provide the estimate $\hat{y}^N_{S_m}$, and then, the estimate is used to describe the same noisy data in (18). However, it is reasonable to use the noisy data once to provide the estimate $\hat{y}^N_{S_m}$ and then use this estimate to describe the "noiseless data." The DL of the noiseless data in subspace S_m, using $\hat{y}^N_{S_m}$, is

$$\mathrm{DL}(\bar{y}^N; \hat{y}^N_{S_m}) = \log_2\sqrt{2\pi\sigma_w^2} + \frac{\log_2 e}{2\sigma_w^2} z_{S_m}. \tag{19}$$

The MNDL is obtained for S_{m^*} when the following holds:

$$S_{m^*} = \arg\min_{S_m} \mathrm{DL}(\bar{y}^N; \hat{y}^N_{S_m}). \tag{20}$$

In order to compare the new DLs, the noise variance and the reconstruction errors z_{S_m} in (9) are needed. If the noise variance is known, it is enough to estimate the reconstruction error. In the following section, we present a method for estimating the reconstruction error. We discuss the denoising problem with unknown variance in Section VIII-B.

V. Reconstruction Error

The reconstruction error z_{S_m} in (9) is a sample of random variable Z_{S_m}. As was mentioned previously, the expected value of this random variable in (13) is an important element in the existing thresholding methods.

Here, we provide probabilistic bounds on both errors z_{S_m} and $\mathrm{E}(Z_{S_m})$. Note that based on Parseval's theorem, the coefficient error $(1/N)\|\theta^* - \hat{\theta}_{S_m}\|_2^2$ is the same as the reconstruction error

$$z_{S_m} = \frac{1}{N}\|\bar{y}^N - \hat{y}^N_{S_m}\|_2^2 = \frac{1}{N}\|\theta^* - \hat{\theta}_{S_m}\|_2^2. \tag{21}$$

We now study the characteristics of this random variable closely.

A. Reconstruction Error z_{S_m} in Subspace S_m

For simplicity and without loss of generality, we assume that S_m corresponds to only the first m bases ($s_i \in S_m$, $i = 1, \ldots, m$). For subspace S_m, (1) can be rewritten as follows:

$$y^N = [\, A_{S_m} \quad B_{S_m} \,] \begin{bmatrix} \theta^*_{S_m} \\ \\ \Delta_{S_m} \end{bmatrix} + w^N \tag{22}$$

where the columns of matrix A_{S_m} are $s_i \in S_m$, $1 \leq i \leq N$, and the columns of matrix B_{S_m} are basis vectors that are not in

[2] In two-stage MDL, the data description length is defined as

$$\mathrm{DL}(y^N; S_m) \triangleq \frac{m}{2N}\log_2(N) + \mathrm{DL}(y^N; \hat{y}^N_{S_m}).$$

The extra term $m\log_2(\sqrt{N})$ is the codelength of elements of S_m. Unlike the second term, this term is an increasing function of m, and therefore, the minimum of this DL occurs for some S_m, $m \leq N$. The codelength is the result of partitioning each dimension of the θ_{S_m} coefficients with grids of width $\log_2\sqrt{N}$. However, the partitioning can be done with any other discretization per dimension. The codelength for all of the θ_{S_m} coefficients is $m\log_2(A)$ when each dimension has $\log_2(A)$ elements [1].

S_m, $s_i \in \bar{S}_m$. The vector $\theta^*_{S_m}$ contains the coefficients of the noiseless data in S_m

$$\theta^* = \begin{bmatrix} \theta^*_{S_m} \\ \\ \Delta_{S_m} \end{bmatrix} \tag{23}$$

where Δ_{S_m} is a vector of length $N-m$, corresponding to the coefficients of bases that are not in S_m. Therefore, the vector of coefficient estimates, with the ith element $\hat{\theta}_{S_m}(i)$ in (7), is

$$\hat{\theta}_{S_m} = \begin{bmatrix} A^{\mathrm{H}}_{S_m} y^N \\ \\ 0_{(N-m\times 1)} \end{bmatrix} = \begin{bmatrix} \theta^*_{S_m} + A^{\mathrm{H}}_{S_m} w^N \\ \\ 0_{(N-m\times 1)} \end{bmatrix}. \tag{24}$$

The subspace coefficient error z_{S_m} can be expressed as a function of the basis vectors, additive noise, and the noiseless coefficients:

$$z_{S_m} = \frac{1}{N}\|\theta^* - \hat{\theta}_{S_m}\|_2^2 = \frac{1}{N}\|A^{\mathrm{H}}_{S_m} w^N\|_2^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 \tag{25}$$

where $\|\Delta_{S_m}\|_2$ is the l_2-norm of the vector of discarded coefficients in subspace S_m.

Lemma 1: The coefficient error z_{S_m}, in (25), is a sample of a random variable $Z_{S_m} = (1/N)\|\theta^* - \hat{\Theta}_{S_m}\|_2^2$. For this random variable, we have

$$\frac{N}{\sigma_w^2}\left(Z_{S_m} - \frac{1}{N}\|\Delta_{S_m}\|_2^2\right) \sim \chi^2_m \tag{26}$$

where χ^2_m is a Chi-square random variable of order m. Therefore, Z_{S_m} has the following expected value and variance:

$$\mathrm{E}(Z_{S_m}) = \frac{m}{N}\sigma_w^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 \tag{27}$$

$$\mathrm{var}(Z_{S_m}) = \frac{2m}{N^2}(\sigma_w^2)^2. \tag{28}$$

Proof: In (25), the noise-dependent component of z_{S_m} is a function of

$$\frac{1}{\sqrt{N}} A^{\mathrm{H}}_{S_m} w^N = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix} \tag{29}$$

where the u_is are independent, zero mean, white Gaussian noises with variance σ_w^2/N. Hence, z_{S_m} in (25), which is

$$z_{S_m} = \frac{1}{N}\|\Delta_{S_m}\|_2^2 + \sum_{i=1}^{m} |u_i|^2 \tag{30}$$

has the expected value and variance provided in this lemma. $\diamondsuit$

B. The Mean Square Reconstruction Error $\mathrm{E}(Z_{S_m})$, if $\|\Delta_{S_m}\|_2^2$ Is Available

If $\|\Delta_{S_m}\|_2^2$ is known, then $\mathrm{E}(Z_{S_m})$ is provided by (27). This expected value has two components: one caused by the noise and the other by the ignored vector coefficients. The tradeoff between these two parts minimizes $\mathrm{E}(Z_{S_m})$ for some m. This is called the bias-variance trade-off approach.

C. Probabilistic Bounds on z_{S_m} if $\|\Delta_{S_m}\|_2^2$ Is Available

The random variable z_{S_m} is near its mean with probability p_1 as follows:

$$\Pr\{|Z_{S_m} - E(Z_{S_m})| \le D_{S_m}\} = p_1 \tag{31}$$

where D_{S_m} is a function of p_1 and the structure of random variable Z_{S_m}. Since Z_{S_m} has the structure of a Chi-square random variable, D_{S_m} can be found using the table of Chi-square random variables and is only a function of the mean and the variance of Z_{S_m} in (27) and (28). Therefore, D_{S_m} is a function of $\|\Delta_{S_m}\|_2^2$, σ_w, m, and p_1. With probability p_1, the reconstruction error is bounded as follows:

$$\underline{z_{S_m}(p_1)} \le z_{S_m} \le \overline{z_{S_m}(p_1)} \tag{32}$$

where by using (27) and (31),

$$\underline{z_{S_m}(p_1)} = \frac{m}{N}\sigma_w^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 - D_{S_m}(p_1, \sigma_w, m, \|\Delta_{S_m}\|_2^2). \tag{33}$$

$$\overline{z_{S_m}(p_1)} = \frac{m}{N}\sigma_w^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 + D_{S_m}(p_1, \sigma_w, m, \|\Delta_{S_m}\|_2^2). \tag{34}$$

These bounds provide bounds for the desired description length in (19). The bounds can be used for the comparison of subspaces. For this comparison, it is important to compare the *events* with the same probability in the competing subspaces. For each subspace, the event is described by (31), which happens with a chosen probability p_1.

VI. Bounds on z_{S_m} and $\mathrm{E}(Z_{S_m})$ Using the Observed Noisy Data y^N

In the previous section, $\mathrm{E}(z_{S_m})$ and probabilistic bounds on z_{S_m} were provided when $\|\Delta_{S_m}\|_2^2$ was known. However, in the denoising problem, the only available information is the noisy data y^N. In this section, we provide a method that probabilistically validates $\|\Delta_{S_m}\|_2^2$ by using the noisy data. These bounds can then be used in providing probabilistic bounds on $\mathrm{E}(Z_{S_m})$ and z_{S_m}.

A. Probabilistic Bounds on $\|\Delta_{S_m}\|_2^2$ Using the Observed Data y^N

Using the observed data, we show that with validation probability p_2

$$L_{S_m}(y^N, p_2) \le \frac{1}{N}\|\Delta_{S_m}\|_2^2 \le U_{S_m}(y^N, p_2). \tag{35}$$

The validation is done through the use of the data error x_{S_m} in (8). First, we discuss the properties of this error.

B. Data Error x_{S_m} in Subspace S_m

The estimate of data in subspace S_m in (6) is

$$\hat{y}_{S_m} = [\, A_{S_m} \quad B_{S_m} \,]\hat{\theta}_{S_m} \tag{36}$$

and the data representation error is

$$x_{S_m} = \frac{1}{N}\|y^N - \hat{y}^N_{S_m}\|_2^2 = \frac{1}{N}\|B_{S_m}\Delta_{S_m} + G_{S_m} w^N\|_2^2 \tag{37}$$

where

$$G_{S_m} = I - A_{S_m} A_{S_m}^{\mathrm{H}} = B_{S_m} B_{S_m}^{\mathrm{H}} \tag{38}$$

is a projection matrix.

Lemma 2: The observed error in (37) is a sample of a Chi-square random variable $X_{S_m} = (1/N)\|y^N - \hat{Y}_{S_m}^N\|_2^2$. For this random variable, we have

$$\frac{N}{\sigma_w^2} X_{S_m} \sim \chi_{N-m}^2 \tag{39}$$

where χ_{N-m}^2 is a Chi-square random variable of order $N-m$. Therefore, X_{S_m} has the following expected value and variance:

$$E(X_{S_m}) = \left(1 - \frac{m}{N}\right)\sigma_w^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 \tag{40}$$

$$\mathrm{var}(X_{S_m}) = \frac{2}{N}\left(1 - \frac{m}{N}\right)(\sigma_w^2)^2 + \frac{4\sigma_w^2}{N^2}\|\Delta_{S_m}\|_2^2. \tag{41}$$

Proof: In Appendix A. ♢

C. Probabilistic Bounds on $\|\Delta_{S_m}\|_2^2$ Using x_{S_m}

Given the noisy data x_{S_m}, one sample of the random variable X_{S_m} is available. The variance of this random variable is of order $1/N$ of its expected value. Therefore, if the length of data is long enough, the variance of this random variable is close to zero. In this case, one method of estimating $\|\Delta_{S_m}\|_2^2$ is to assume that the available sample x_{S_m} is a good estimate of its expected value in (40)

$$\frac{1}{N}\|\hat{\Delta}_{S_m}\|_2^2 \approx x_{S_m} - \left(1 - \frac{m}{N}\right)\sigma_w^2. \tag{42}$$

However, if this estimate is used to compare the different subspaces, the variance of X_{S_m} is ignored. In each S_m, as shown in (41), X_{S_m} has a variance that is a function of m, N, and $\|\Delta_{S_m}\|_2^2$. Therefore, the confidence of the estimate in (42) for different subspaces is different. Similar to the argument for estimation of the reconstruction error, we suggest a probabilistic validation method based on comparison of *events* that occur with the same probability as follows:

The Chi-square probability distribution of the data error is a function of $\|\Delta_{S_m}\|_2^2$ and the noise variance. We suggest validating $\|\Delta_{S_m}\|_2^2$ such that x_{S_m} is in the neighborhood of its mean with probability p_2. For each value of $\|\Delta_{S_m}\|_2^2$, we have

$$\Pr\{|X_{S_m} - E(X_{S_m})| \le J_{S_m}\} = p_2. \tag{43}$$

The bound J_{S_m} is a function of $\|\Delta_{S_m}\|_2^2$, σ_w^2, m, and p_2. The value of J_{S_m} is calculated based on these numbers and the Chi-square table.[3]

[3]One choice for J_{S_m} in (43) is $J_{S_m} = \gamma\sqrt{\mathrm{var}X_{S_m}}$. In this case, by using the Chebychev inequality, we have

$$p_2 \ge 1 - \frac{1}{\gamma^2} \quad \text{or} \quad \gamma \le \sqrt{\frac{1}{1-p_2}}$$

which shows how γ and p_2 are related. The closeness of γ to $\sqrt{1/(1-p_2)}$ depends on the distribution of the error in each subspace.

Therefore, given x_{S_m} and with validation probability p_2, we validate $\|\Delta_{S_m}\|_2^2$s for which

$$|x_{S_m} - E(X_{S_m})| \le J_{S_m}(p_2, \sigma_w^2, m, \|\Delta_{S_m}\|_2^2). \tag{44}$$

This validation provides U_{S_m} and L_{S_m} upper and lower bounds on $(1/N)\|\Delta_{S_m}\|_2^2$ as a function of y^N and p_2 in (35).

Note that setting p_2 in (43) to zero is the same as ignoring the variance of X_{S_m}. In this case, $J_{S_m} = 0$ for all the subspaces, and instead of probabilistic bounds on $\|\Delta_{S_m}\|_2^2$, we have the estimate in (42).

D. Bounds on the Reconstruction Error

Using the bounds on $\|\Delta_{S_m}\|_2^2$ from the previous section, we can provide bounds on the reconstruction error. Note that when $\|\Delta_{S_m}\|_2^2$ is known, the bounds are only functions of the confidence probability p_1 in (32). Here, with validation probability p_2 and confidence probability p_1, the reconstruction error is bounded as follows:

$$\underline{z_{S_m}(p_1, y^N, p_2)} \le z_{S_m} \le \overline{z_{S_m}(p_1, y^N, p_2)} \tag{45}$$

where

$$\underline{z_{S_m}(p_1, y^N, p_2)} = \min\left\{0, \min_{(1/N)\|\Delta_{S_m}\|_2^2 \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) - D_{S_m}\}\right\} \tag{46}$$

$$\overline{z_{S_m}(p_1, y^N, p_2)} = \max_{(1/N)\|\Delta_{S_m}\|_2^2 \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) + D_{S_m}\}. \tag{47}$$

These upper and lower bounds are provided using the Chi-square table, σ_w^2, m, N, and the observed data y^N for a choice of confidence probability p_1 and validation probability p_2.

VII. Gaussian Distribution Estimation and Reconstruction Error

To provide bounds on z_{S_m}, in both probabilistic validation steps in the previous section, we used the Chi-square distribution table. In this section, we suggest using the Central Limit Theorem (CLT) to approximate the Chi-square distributions with Gaussian distributions. This gives us the advantage of finding mathematical expressions for the bounds on $\|\Delta_{S_m}\|_2^2$, without using any lookup table. In this case, the bounds on the reconstruction error in (46) and (47) are only functions of p_1, σ_w, m, p_2, and the observed noisy signal. The advantage of these closed forms is that their calculation is easy, and they provide some informative insights on the tightness of bounds on z_{S_m}.

A. Bounds on $\|\Delta_{S_m}\|_2^2$

In calculating bounds for $\|\Delta_{S_m}\|_2^2$, in (35), the observed data error x_{S_m} and the table of Chi-square random variables are used. The Chi-square distribution of X_{S_m} is a sum of $N-m$ independent squared Gaussian random variables. Since the noise variance is finite, the CLT enables us to estimate this Chi-square

distribution with a Gaussian distribution when $N-m$ is large enough. The following theorem allows us to provide closed-form bounds for $\|\Delta_{S_m}\|_2^2$, instead of using the Chi-square table, by employing the properties of a Gaussian distribution.

Theorem 1: If $N-m$ is large enough, the random variable X_{S_m} can be well estimated by a Gaussian distribution. In this case, the validation probability p_2 is considered in the form $p_2 = Q(\alpha)$, where $Q(\alpha) = \int_{-\alpha}^{\alpha}(1/\sqrt{2\pi})e^{-x^2/2}dx$. If p_2, or equivalently α, is chosen large enough such that

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1-\frac{m}{N}-\frac{x_{S_m}}{\sigma_w^2}\right) \tag{48}$$

then with validation probability $p_2 = Q(\alpha)$, we have

$$L_{S_m}(y,Q(\alpha)) \leq \frac{1}{N}\|\Delta_{S_m}\|_2^2 \leq U_{S_m}(y,Q(\alpha)) \tag{49}$$

where

$$U_{S_m}(y,Q(\alpha)) = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha) \tag{50}$$

and

$$K_{S_m}(\alpha) = 2\alpha\frac{\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N}+x_{S_m}-\frac{1}{2}m_w} \tag{51}$$

$$m_w = \left(1-\frac{m}{N}\right)\sigma_w^2. \tag{52}$$

The lower bound $L_{S_m}(y,Q(\alpha))$ is zero if

$$(m_w - \alpha\sqrt{v_{S_m}}) \leq x_{S_m} \leq (m_w + \alpha\sqrt{v_{S_m}}) \tag{53}$$

where

$$v_{S_m} = \frac{2}{N}\left(1-\frac{m}{N}\right)\sigma_w^4. \tag{54}$$

Otherwise, the lower bound is

$$L_{S_m}(y,Q(\alpha)) = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} - K_{S_m}(\alpha). \tag{55}$$

Proof: In Appendix B. ♢

B. Bounds on the Reconstruction Error

In the previous section, we showed how to simplify calculation of bounds on $\|\Delta_{S_m}\|_2^2$ for some subspaces. Another step for calculation of bounds on z_{S_m} is the calculation of D_{S_m} in (33) and (34). For this calculation, we use the table of Chi-square random variables after choosing the confidence probability p_1. This calculation can be simplified if m is large enough. The Chi-square distribution of the reconstruction error Z_{S_m} in (30) is the sum of m independent squared Gaussian random variables. Therefore, if m is large enough, we can estimate this Chi-square distribution with a Gaussian distribution. In this case, with $p_1 = Q(\beta)$, the probabilistic event (31) can be written in the form

$$Pr\{|Z_{S_m} - E(Z_{S_m})| \leq \beta\sqrt{\text{var}Z_{S_m}}\} = Q(\beta) \tag{56}$$

which implies that

$$D_{S_m} = \beta\sqrt{\text{var}Z_{S_m}} \tag{57}$$

$$= \frac{\beta\sqrt{2m}}{N}\sigma_w^2. \tag{58}$$

Therefore, the bounds on z_{S_m} (33) and (34), when $\|\Delta_{S_m}\|_2^2$ is known, are simply

$$\underline{z_{S_m}(Q(\beta))} = \frac{m}{N}\sigma_w^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 - \frac{\beta\sqrt{2m}}{N}\sigma_w^2 \tag{59}$$

$$\overline{z_{S_m}(Q(\beta))} = \frac{m}{N}\sigma_w^2 + \frac{1}{N}\|\Delta_{S_m}\|_2^2 + \frac{\beta\sqrt{2m}}{N}\sigma_w^2. \tag{60}$$

The last step is to use the observed data to provide bounds on these upper and lower bounds. The bounds on $(m/N)\sigma_w^2 + (1/N)\|\Delta_{S_m}\|_2^2$, which is the expected value of Z_{S_m} in (27), can be calculated by using the bounds on $(1/N)\|\Delta_{S_m}\|_2^2$, which was discussed in Section VI-C.

If both $N-m$ and m are large enough, we can use not only (58) for D_{S_m} but, in addition, Theorem 1 to provide bounds on $\|\Delta_{S_m}\|_2^2$. In this case, the lower bound (46) and upper bound (47) on z_{S_m}, with confidence probability $Q(\beta)$ and validation probability of $Q(\alpha)$, are

$$\underline{z_{S_m}(Q(\beta),y^N,Q(\alpha))} = \frac{m}{N}\sigma_w^2 + L_{S_m}(y^N,\alpha) - \beta\frac{\sqrt{2m}}{N}\sigma_w^2 \tag{61}$$

$$\overline{z_{S_m}(Q(\beta),y^N,Q(\alpha))} = \frac{m}{N}\sigma_w^2 + U_{S_m}(y^N,\alpha) + \beta\frac{\sqrt{2m}}{N}\sigma_w^2 \tag{62}$$

where $L_{S_m}(y^N,\alpha)$ and $U_{S_m}(y^N,\alpha)$ are calculated in Theorem 1.

C. Proper Choices of Validation Probability p_2 and Confidence Probability p_1

The bounds on z_{S_m} in previous sections are provided probabilistically. This means that with $p_1 = 0.3$ and $p_2 = 0.4$, the true z_{S_m} is between the provided bounds with confidence probability 0.3 and validation probability 0.4. The higher these probabilities are, the more confidence there is on the provided bounds. Therefore, it is important to choose the two probabilities close to one. The price for this choice is that the gap between the upper and lower bounds becomes larger as the probabilities approach one.

It is easier to observe the behavior of the bounds as a function of the two probabilities in (61) and (62). In this case, parameters α and β can be chosen large enough such that $p_2 = Q(\alpha)$ and $p_1 = Q(\beta)$ are close to one. However, to have tight bounds on z_{S_m}, β/N has to be chosen small enough such that $\pm\beta(\sqrt{2m}/N)\sigma_w^2$ are close. In addition, the upper and lower bounds on $(1/N)\|\Delta_{S_m}\|_2^2$ in (49) have to be close to each other. If α is chosen such that $\alpha/\sqrt{N}$ is small, then the bounds provided by Theorem 1 are tight. Therefore, if these conditions are

satisfied, the method provides tight bounds on the reconstruction error with confidence and validation probabilities close to one.

VIII. Subspace Comparison and Best Basis Search

If the noise variance is known, comparison of the noiseless data description length in (20) for different subspaces is the same as comparison of the reconstruction error z_{S_m}. For such a comparison, we suggest comparing the probabilistic worst-case reconstruction error $\overline{z_{S_m}}$ and choosing the optimum subspace based on this criterion

$$S_{m^*} = \arg\min_{S_m} \overline{z_{S_m}(p_1, y^N, p_2)}. \tag{63}$$

This decision is made with confidence probability p_1 and validation probability p_2.

A. Gaussian Distribution Estimation

For the subspaces in which the distributions of Z_{S_m} and X_{S_m} are estimated with Gaussian distributions, the bound on the reconstruction error in (62) can be used. Therefore, the optimum subspace, based on the worst-case reconstruction error, is

$$S_{m^*} = \arg\min_{S_m} \overline{z_{S_m}(Q(\beta), y^N, Q(\alpha))} \tag{64}$$

which is valid with confidence probability $Q(\beta)$ and validation probability $Q(\alpha)$.

B. Unknown Noise Variance

The existing thresholding methods are based on the knowledge of the additive noise variance. However, in most practical problems, the additive noise variance is not known, and only a range of possible noise variances is available. In [4] and for wavelet thresholding, it is suggested to estimate the variance with $\hat{\sigma}_w = \text{MAD}/0.6745$, where MAD is the median of absolute value of normalized fine scale wavelet coefficients. Another popular variance estimation is the maximum likelihood (ML) estimator. In these variance estimation methods and similar approaches, a primary threshold is first chosen to pick the fine-scale wavelet coefficients. The noise variance is estimated using these coefficients. The estimated noise variance is then used to provide the optimum threshold. The sensitivity of this approach to the choice of the first threshold, which estimates the noise variance, is not known. These methods of estimation are successful only if enough of the coefficients can be assumed to be noise dependent only. This requires the noiseless data to have a small number of nonzero coefficients, i.e., the noiseless data to be a smooth signal.

In the new proposed method, the noiseless DL in (20) $\text{DL}_{\sigma_w^2}(\overline{y}^N; \hat{y}^N_{S_m})$ is a function of the noise variance. To estimate the noise variance, we suggest the following approach: Find the optimum subspace S_{m^*} as a function of noise variances in the given range. For any noise variance σ_w^2, the bounds on the reconstruction error provides bounds on the description length

$$\underline{\text{DL}_{\sigma_w^2}(\overline{y}^N; \hat{y}^N_{S_m})} \le \text{DL}_{\sigma_w^2}(\overline{y}^N; \hat{y}^N_{S_m}) \le \overline{\text{DL}_{\sigma_w^2}(\overline{y}^N; \hat{y}^N_{S_m})} \tag{65}$$

where upper and lower bounds are calculated using (19) and the upper and lower bounds on z_{S_m} in (45). Therefore, the MNDL as a function of the noise variance and the observed data is

$$\text{MNDL}(y^N, \sigma_w^2) = \min_{S_m} \overline{\text{DL}_{\sigma_w^2}(\overline{y}^N; \hat{y}^N_{S_m})}. \tag{66}$$

To estimate the unknown variance, we suggest choosing the noise variance that minimizes the noiseless data DL. Therefore, the optimal noise variance is such that

$$\hat{\sigma}_w^2 = \arg\min_{\sigma_w^2} \text{MNDL}(y^N, \sigma_w^2). \tag{67}$$

C. Best Basis Search

One important advantage of the new method is that the estimated description length can be used to compare different basis families. For each basis family B_i, the MNDL is provided by

$$\text{MNDL}_{B_i}(y^N, \sigma_w^2) = \min_{S_m \in B_i} \overline{\text{DL}_{\sigma_w^2}(\overline{y}^N; \hat{y}^N_{S_m})}. \tag{68}$$

The comparison of basis results

$$\text{MNDL}(y^N) = \arg\min_{B_i} \text{MNDL}_{B_i}(y, \sigma_w^2). \tag{69}$$

If the noise variance is unknown, it can be estimated simultaneously by using (67). For each basis, the optimum noise variance is $\hat{\sigma}_w^2(B_i)$. The optimum Basis is the one for which the MDL is minimized

$$\text{MNDL}(y^N) = \arg\min_{B_i} \text{MNDL}_{B_i}(y^N, \hat{\sigma}_w^2(B_i)). \tag{70}$$

IX. New Thresholding Method

Search for the best basis among all 2^N subspaces of S_N may not be practically feasible. One alternative subspace comparison method is to compare nested subspaces of different order. After calculation of the least square estimates of coefficients $\lambda(i)$s in (5), they are sorted in a decreasing order based on their absolute value. This provides a nested set of subspaces. Subspace S_1 is the subspace spanned by s_j, where $\lambda(j) = \max_i |\lambda(i)|$, and subspace S_i is spanned by the first ith bases associated with the sorted set. In this case, the number of subspaces is at most N. Let S_{m^*} be the optimum subspace that minimizes the DL among these nested subspaces. This choice also provides the optimum threshold that is the maximum absolute value of coefficients associated with S_{m^*}

$$\tau^* = \max_i |\lambda(i)|, \quad i \in \{j |\ s_j \in S_{m^*}\}. \tag{71}$$

Therefore, a hard thresholding method results from the choice of S_{m^*}

$$S_{m^*} \longrightarrow \tau^*. \tag{72}$$

Comparing this approach with the thresholding method in (12), the new method provides an optimum threshold that is a function of the observed noisy data, validation, and confidence probabilities.

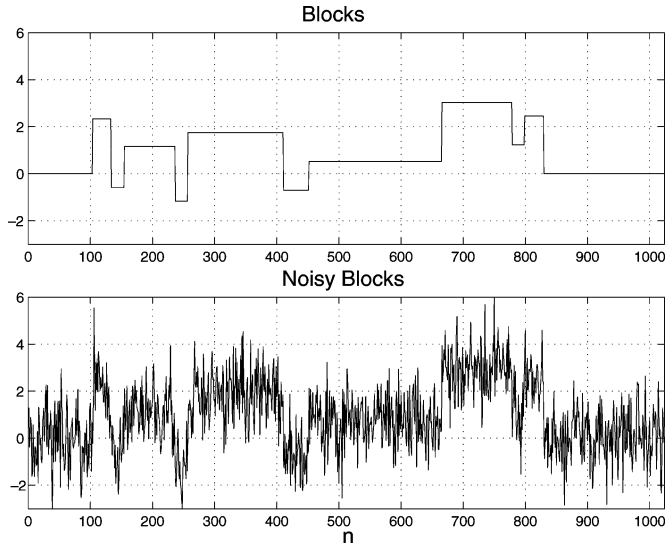

Fig. 1. Noiseless signal $\bar{y}[n]$ and noisy signal $y[n]$.

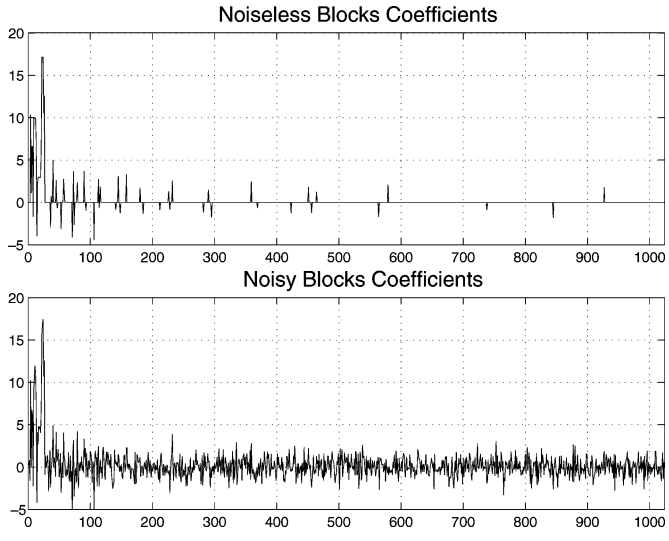

Fig. 2. Noiseless signal coefficients θ^* and noisy signal coefficients λ.

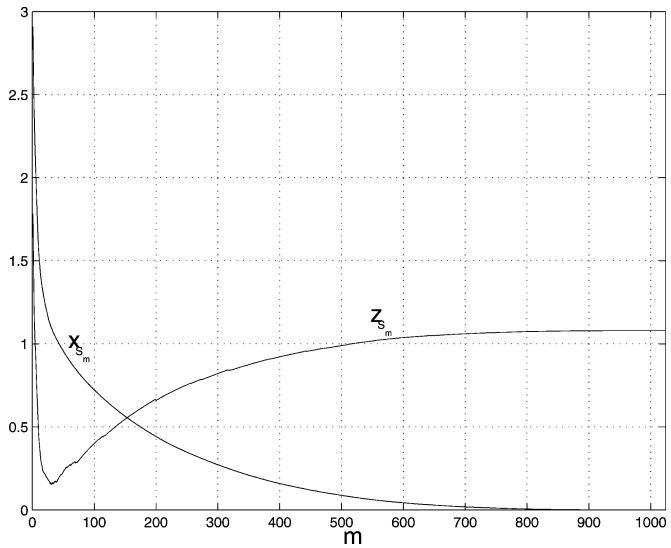

Fig. 3. Two important errors: the available data error x_{S_m} and the unavailable desired error z_{S_m}.

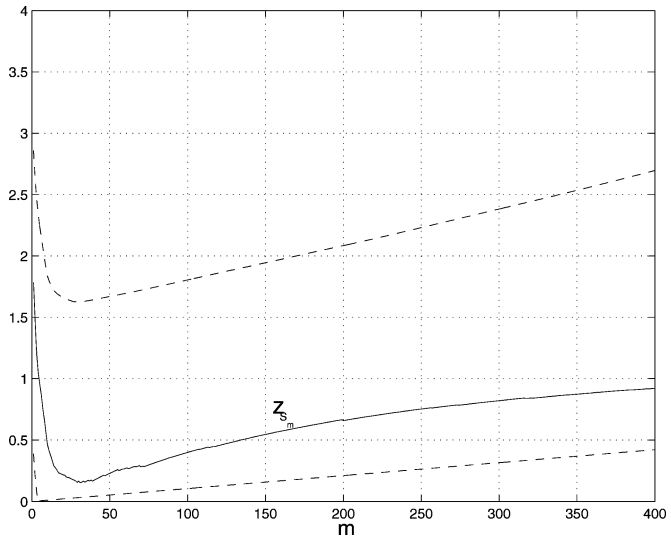

Fig. 4. Solid line is the reconstruction error for the nested S_ms. The dashed lines are the upper and lower bounds with validation probability $Q(15)$ and confidence probability $Q(\beta) = Q(70)$.

X. Simulation

In this section, we provide three denoising simulation examples. In the first example, the additive noise variance is known. In the second example, the same problem is considered when the noise variance is unknown. The third example highlights the performance of the new approach for an unsmooth signal.

A. Wavelet Denoising and Reconstruction Error Bounds

Fig. 1 shows the considered signal and its noisy version of length $N = 1024$. The noiseless signal is the standard blocks signal introduced in [5]. The signal-to-noise ratio of the noisy signal is 2.8 dB. Fig. 2 shows the 1024 wavelet coefficients of both signals in series (the first 32 numbers are approximation coefficients; the rest are detail coefficients of length 32, 64, 128, 256, and 512) with Haar wavelet filters at level five. While the noisy data has 1024 nonzero coefficients, the noiseless signal has only 66 nonzero coefficients. In most denoising problems, the goal is not only to estimate the noiseless data but to provide an optimum (minimum) number of nonzero coefficients for the purpose of compression as well.

To denoise the noisy observed signal, we consider the nested subsets of S_Ns, which are explained in Section IX. For example, S_1 corresponds to the subspace associated with the first element of the sorted coefficients. Fig. 3 shows the available data error x_{S_m} and the desired reconstruction error z_{S_m}. The goal is to use x_{S_m} in each subspace to provide bounds on z_{S_m}. Fig. 4 shows the probabilistic bounds on the reconstruction error z_{S_m} using the data error x_{S_m}. It shows the bounds for the first 400 subspaces. To choose the optimum subspace $\overline{z_{S_m}}$, the upper bound on the reconstruction error is minimized. As the figure shows, the minimum of upper bound is at $m^* = 30$.

Table I shows the results of using different thresholding methods. Methods (1) and (2) use the worse-case probabilistic bound on the reconstruction error $\overline{z_{S_m}}$ in (47). In the first method, the validation and confidence probabilities are

TABLE I
(1) NEW METHOD (MNDL) WITH $p_2 = Q(15.5)$ AND $p_1 = Q(40)$. (2) NEW METHOD WITH $p_2 = Q(15.5)$ AND $p_1 = Q(70)$. (3) DONOHO AND JOHNSTONE THRESHOLDING WITH $\tau^* = \sigma_w\sqrt{2\log(N)}$. (4) MDL THRESHOLDING WITH $\tau^* = \sigma_w\sqrt{\log(N)}$

	(1)	(2)	(3)	(4)
Number of nonzero Coeffs. m^*	66	30	27	38
Reconstruction error $z_{S_{m^*}}$	.2845	.1661	0.175	.175
Threshold τ^*	2.29	3.27	3.87	2.74

$p_2 = Q(\alpha)$, where $\alpha = 1.5\log(N) = 15$, and $p_1 = Q(\beta)$, where $\beta = 40$. In method (2), the probabilities are closer to one: $p_2 = Q(15)$ and $p_1 = Q(70)$. For these methods, the optimum subspace S_{m^*} is the subspace that minimizes the worse-case upper bound. Method (1) chooses $m^* = 66$, and method (2) chooses $m^* = 30$. As it was explained in Section VII-C, the choices of α and β partially depends on the observed data. For example, for the subspaces that the Chi-square random variables are estimated with the Gaussian distributions, α has to satisfy the inequality in (48). The α in this example is the minimum value for which this inequality holds. For p_1, β is chosen to be much larger than α and such that $\alpha/\sqrt{N}$ is comparable with β/N.

The table shows the reconstruction errors associated with the chosen subspaces. As it was discussed in Section IX, the choice of optimum subspace provides an optimum threshold τ^*, which is given in the table. Methods (3) and (4) are the existing thresholding methods. In method (3), the well-known wavelet threshold $\sigma_w\sqrt{2\log(N)}$ is used [5]. In method (4), the existing MDL threshold $\sigma_w\sqrt{\log(N)}$ is used [8]. As was discussed in Section III, the choice of threshold in these two methods leads to the choice of an optimum subspace with order m^*, which is shown in the table.

In this example, the reconstruction error in Fig. 4 is minimized for $m^* = 31$, and the minimum is 0.1534. Table I shows the associated reconstruction error for the different methods. Based on this error, the optimum approach among these four methods is method (2), in which the minimum reconstruction error is 0.1661. It is important to mention that the comparison of the bounds on the reconstruction error is equivalent to the comparison of the DL bounds. This is due to the fact that the DL is an affine function of the reconstruction error (67).

Note that in this simulation the reconstruction errors in methods (3) and (4) are the same with two different thresholds. In addition, it is important to stress that unlike methods (3) and (4), the new approach in methods (1) and (2) not only provide the thresholds but also provide the probabilistic bounds on the reconstruction error.[4]

B. Unknown Variance and Denoising

In the previous example, the noise variance (or SNR) was known. Here, we consider the same problem when the noise variance ($\sigma_w^2 = 1$) is unknown. Therefore, the first step is to estimate the DL in (18) as a function of noise variance. For this step, using the observed data, the upper and lower bounds on the reconstruction error as a function of noise variance are calculated. Similar to method (2), in the previous example, we provide the bounds with validation probability $p_2 = Q(15)$ and confidence probability $p_1 = Q(70)$. Consequently, these bounds provide bounds on the DL. Fig. 5 shows the upper bound on the description length for two different noise variances (or SNRs). For each noise variance, the optimum subspace is chosen by minimizing the provided upper bound on the description length. As Fig. 5 shows the MDL (and optimum subspace) for SNR $= 4$ dB is 2.51 (and $m^* = 33$). The minimum for the true SNR $= 2.8$ dB is 2.46 with $m^* = 30$.

Fig. 6 shows the provided MDL for SNRs between 1 and 5 dB (variances between 1.64 and 0.65). The last step is to compare these MDLs to estimate the noise variance (67). As the figure shows, the minimum is for SNR $= 2.2$ dB, or equivalently,

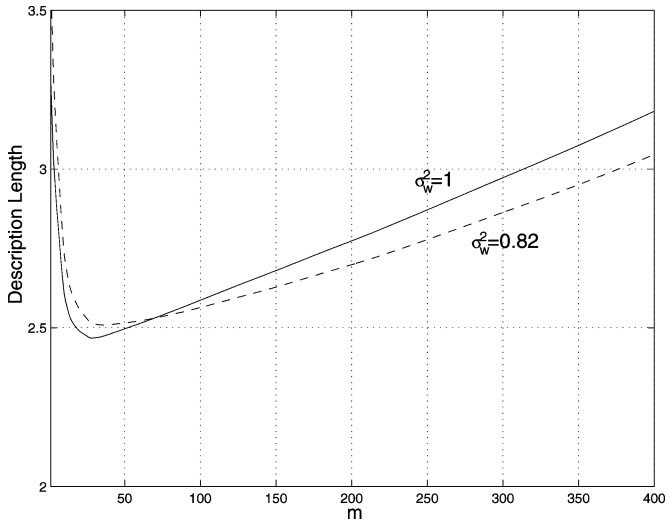

Fig. 5. DL upper bound for two noise variance assumptions: $\sigma_w^2 = 1$ (SNR $= 2.8$ dB) and $\sigma_w^2 = 0.8$ (SNR $= 4$ dB).

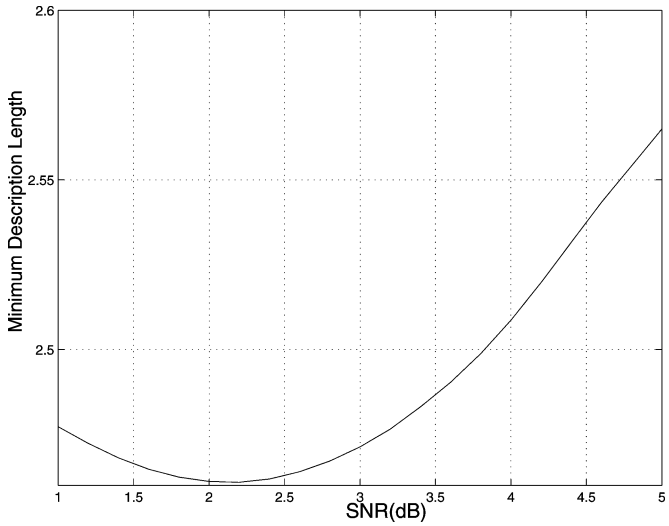

Fig. 6. New minimum description length for SNRs between 1 and 5 dB.

[4]The soft threshoding method with the Donoho and Johnstone threshold provides 27 nonzero coefficients, and the reconstruction error in this method is 0.43. The performance of this method is worse than its counterpart hard thresholding with the same threshold.

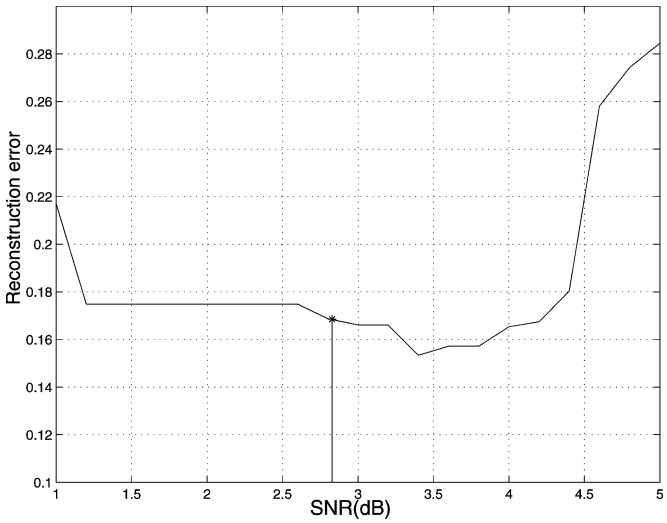

Fig. 7. Reconstruction error for SNRs between 1 and 5 dB.

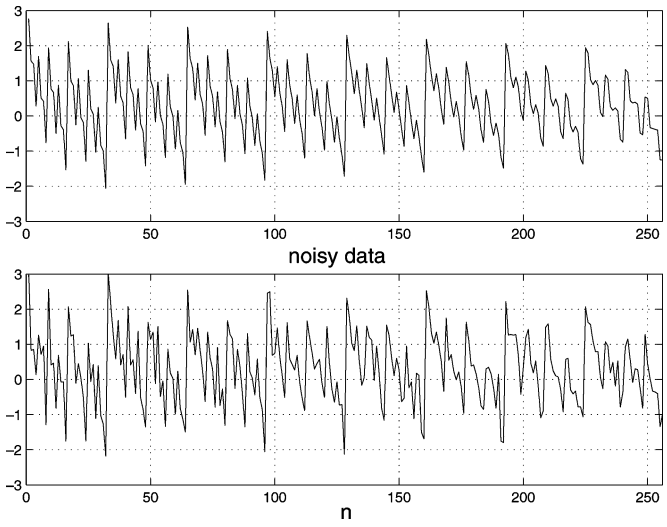

Fig. 8. Noiseless signal $\bar{y}[n]$ and noisy signal $y[n]$.

$\hat{\sigma}_w^2 = 1.25$. With this noise variance, the optimum subspace has order $m^* = 27$.

To compare this example with method (2) in the previous example, we can compare the corresponding reconstruction error associated with the optimum subspaces chosen for different noise variances. Fig. 7 shows the true reconstruction error for different SNR assumptions. As the figure shows, the reconstruction error when the noise variance is estimated to be 1.25 (SNR $= 2.2$) is 0.1750. Therefore, the reconstruction error here is $0.1750 - 0.1661$ worse than the case with the known variance.

C. Unsmooth Noiseless Signal

Here, we repeat the same simulation as in Section X-A for another signal. Fig. 8 shows the noiseless signal of length 256, for which all 256 wavelet coefficients are nonzero. For the noisy signal, the SNR is 10 dB. Fig. 9 shows the wavelet coefficients of both signals when the same wavelet as in the previous examples is used. Fig. 10 shows the desired reconstruction error and the probabilistic bounds, which are provided by using the data

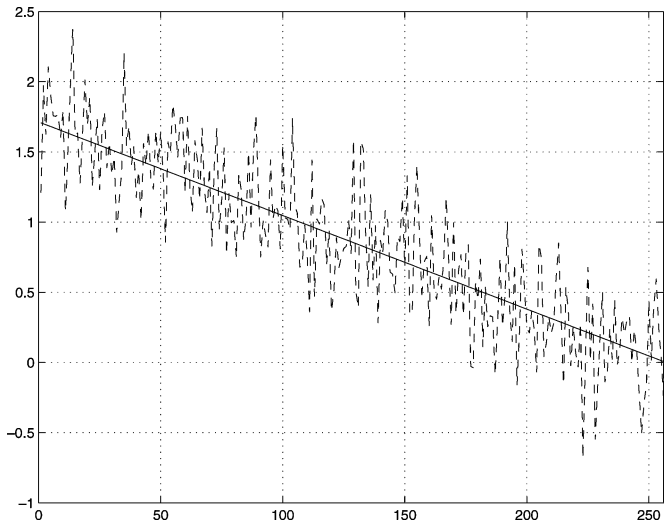

Fig. 9. Solid line: noiseless signal coefficients θ^*. Dashed line: noisy signal coefficients λ.

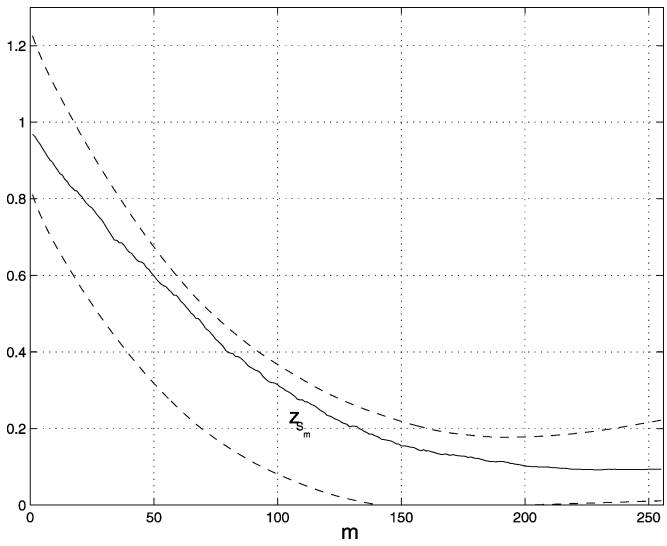

Fig. 10. Solid line is the reconstruction error for the nested S_ms. The dashed lines are the upper bound and lower bound with validation probability $Q(\alpha) = Q(5)$ and confidence probability $Q(10)$.

error. In this example, $\alpha = 5$, and $\beta = 10$. The desired reconstruction error is minimized for $m^* = 230$, and the minimum is 0.092, with threshold $\tau^* = 0.3$. Table II shows the number of nonzero coefficients, the threshold, and the associated reconstruction error for the four methods.[5]

As the table shows, methods (3) and (4) provide thresholds that are very far from the optimum threshold that minimizes the reconstruction error. Note that the l_2-norm of the noiseless signal is one. Therefore, the reconstruction errors in both methods (3) and (4) are almost 33% and 16% of the l_2-norm of the noiseless signal itself. The new methods choose smaller thresholds and more nonzero coefficients. The reconstruction error for these methods are much smaller than methods (3) and (4).

[5]The soft threshoding method with Donoho and Johnstone threshold provides 97 nonzero coefficients, and the reconstruction error in this method is 0.6. The performance of this method is worse than its counterpart hard thresholding with the same threshold.

TABLE II
(1) NEW METHOD (MNDL) WITH $p_2 = Q(5)$ AND $p_1 = Q(10)$. (2) NEW METHOD WITH $p_2 = Q(5)$ AND $p_1 = Q(25)$. (3) DONOHO AND JOHNSTONE THRESHOLDING WITH $\tau^* = \sigma_w\sqrt{2\log(N)}$. (4) MDL THRESHOLDING WITH $\tau^* = \sigma_w\sqrt{\log(N)}$

	(1)	(2)	(3)	(4)
Number of nonzero Coeffs.	192	190	97	149
Reconstruction error $z_{S_{m^*}}$	.1120	.1122	0.33	.16
Threshold τ^*	0.389	0.4	1	0.75

The new method provides the optimum threshold based on the probabilistic bounds on the reconstruction error. The bounds not only provide an information about the behavior of the reconstruction error but also provide a quality evaluation method for any given threshold. For example, consider a scenario where for the purpose of data compression, a reconstruction error of order 0.3 is acceptable. In this case, method (1) accepts any threshold smaller than or equal to 0.9, and the minimum number of nonzero coefficients is 120. This decision is based on the behavior of the probabilistic upper bound, which is shown in Fig. 10. In this case, Method (2) accepts any threshold smaller than or equal to 125. However, neither method (3) nor (4) can provide any other threshold for this case. This example illustrates the potential application of the new proposed method not only for thresholding but for data compression as well [9].

In this example, if the noise variance is unknown, the new method efficiently estimates the noise variance. However, the existing methods of noise estimation, such as ML and MAD, cannot be used since the approach is very sensitive to the choice of the fine wavelet coefficients.

XI. CONCLUSION

In this paper, a new approach to signal denoising is proposed. In this method, no particular assumption on the data length or the number of nonzero coefficients of the noiseless data is needed. The probabilistic bounds of the new description length is given by using the probabilistic bounds on the reconstruction error. The method provides a probabilistic confidence region for the desired reconstruction error. By using the noisy data, we validate the expected value and the variance of the reconstruction error. The bounds are provided with two important probabilities: confidence and validation probabilities. The paper also discusses the proper choices of the probabilities and the asymptotic behavior and tightness of the error bounds as a function of the two probabilities.

The new method also provides an estimate of the noise variance for signal denoising. The advantage of this method is that the denoising and the noise variance estimation are provided simultaneously.

The new approach not only provides an optimum threshold but also provides bounds on the desired criterion as a function of any threshold. Therefore, it can be used as a method of quality evaluation for any given threshold. The consistent theory behind the proposed method promises to overcome several practical problems with the existing noise variance estimation and thresholding methods.

APPENDIX A
PROOF OF LEMMA 2

Rewrite the data error in (37) using the projection matrix $G_{S_m} = B_{S_m}B_{S_m}^{\mathrm{H}}$,

$$\frac{1}{N}\|y^N - \hat{y}_{S_m}^N\|_2^2 = \frac{1}{N}\|B_{S_m}\Delta_{S_m} + G_{S_m}w^N\|_2^2$$
$$= \frac{1}{N}\|B_{S_m}(\Delta_{S_m} + B_{S_m}^{\mathrm{H}}w^N)\|_2^2 \quad (73)$$
$$= \frac{1}{N}\|\Delta_{S_m} + B_{S_m}^{\mathrm{H}}w^N\|_2^2 \quad (74)$$
$$= \left\|\frac{1}{\sqrt{N}}\Delta_{S_m} + v\right\|_2^2. \quad (75)$$

Equation (74) is obtained from (73) since

$$B_{S_m}^{\mathrm{H}}B_{S_m} = I_{N-m\times N-m}$$

and since $v = (1/\sqrt{N})B_{S_m}^{\mathrm{H}}w^N$ is a vector of length $N-m$ whose elements are white Gaussian random variables. Each element v_i has a zero mean with variance σ_w^2/N, and v_is are independent. Therefore, we have

$$\frac{1}{N}\|y^N - \hat{y}_{S_m}^N\|_2^2 = \sum_{i=1}^{N-m}(\delta_i + v_i)^2 \quad (76)$$

where the sum of means of the Chi-square random variable is

$$\sum_{i=1}^{N-m}\delta_i^2 = \frac{1}{N}\|\Delta_{S_m}\|_2^2. \quad (77)$$

Therefore, the data error is a Chi-square random variable that is the sum of $N-m$ independent Gaussian random variables. The expected value and variance of the data error are

$$E(X_{S_m}) = (N-m)\mathrm{var}(v_i) + \sum_{i=1}^{N-m}\delta_i^2 \quad (78)$$

$$\mathrm{var}(X_{S_m}) = 2\mathrm{var}(v_i)\left((N-m)\mathrm{var}(v_i) + 2\sum_{i=1}^{N-m}\delta_i^2\right). \quad (79)$$

By using (77) and $\mathrm{var}(v_i) = \sigma_w^2/N$, we have (40) and (41).

APPENDIX B
PROOF OF THEOREM 1

If m is large enough, using the CLT, the Chi-square distribution is estimated by a Gaussian distribution. Therefore, the probabilistic bounds on x_{S_m} are provided by choosing $J_{S_m} = \beta\sqrt{\mathrm{var}X_{S_m}}$ and $p_2 = Q(\alpha)$ in (43)

$$\Pr\{|X_{S_m} - \mathrm{E}(X_{S_m})| \le \alpha\sqrt{\mathrm{var}(X_{S_m})}\} = Q(\alpha). \quad (80)$$

Define $\bar{x}_{S_m} = x_{S_m} - (1-m/N)\sigma_w^2$. By using (40) and (41) in (80), we want to validate $m_\delta = (1/N)\|\Delta_{S_m}\|_2^2$ for which

$$m_\delta - \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_{S_m}} \le \bar{x}_{S_m} \le m_\delta + \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_{S_m}} \quad (81)$$

where $m_w = (1-m/N)\sigma_w^2$, and $v_{S_m} = (2/N)(1-m/N)\sigma_w^4$.

A. Lower Bound on m_δ

$$\bar{x}_{S_m} - m_\delta < \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_{S_m}}. \tag{82}$$

If $\bar{x}_{S_m} \leq \alpha\sqrt{v_{S_m}}$, the inequality holds for $m_\delta > 0$.

If $\bar{x}_{S_m} \geq \alpha\sqrt{v_{S_m}}$, the lower bound for m_δ is the smallest root of the following equation:

$$(\bar{x}_{S_m} - m_\delta)^2 = \alpha^2\left(\frac{4\sigma_w^2}{N}m_\delta + v_{S_m}\right) \tag{83}$$

which is

$$L_{S_m} = \bar{x}_{S_m} + \frac{2\sigma_w^2\alpha^2}{N} - \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}. \tag{84}$$

Note that $L_{S_m} \leq \bar{x}_{S_m}$.

B. Upper Bound on m_δ

$$m_\delta - \bar{x}_{S_m} > \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_{S_m}}. \tag{85}$$

If $\bar{x}_{S_m} \leq -\alpha\sqrt{v_{S_m}}$, then the inequality does not hold for any m_δ.

If $\bar{x}_{S_m} \geq -\alpha\sqrt{v_{S_m}}$, then the upper bound is the largest root of equation

$$(\bar{x}_{S_m} - m_\delta)^2 = \alpha^2\left(\frac{4\sigma_w^2}{N}m_\delta + v_{S_m}\right) \tag{86}$$

which is

$$U_{S_m} = \bar{x}_{S_m} + \frac{2\alpha^2\sigma_w^2}{N} + \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}. \tag{87}$$

To calculate the upper bound and to avoid the case in which no upperbound can be found, where $\bar{x}_{S_m} \leq -\alpha\sqrt{v_{S_m}}$, α has to be chosen large enough such that

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1 - \frac{m}{N} - \frac{x_{S_m}}{\sigma_w^2}\right). \tag{88}$$

REFERENCES

[1] S. Beheshti and M. A. Dahleh, "Minimum description complexity (part I): A new order selection method," *IEEE Trans. Inf. Theory*, 2005, submitted for publication.

[2] S. G. Chang, Y. Bin, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532–1546, Sep. 2000.

[3] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[4] D. Donoho, "Denoising by soft thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. , pp. 613–627, 1995.

[5] D. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, pp. 425–455, 1994.

[6] S. Efromovich, J. Lakey, M. C. Pereyra, and N. Tymes Jr., "Data-driven and optimal denoising of a signal and recovery of its derivative using multiwavelets," *IEEE Trans. Signal Process.*, vol. 52, no. 3, pp. 628–635, Mar. 2004.

[7] H. Krim, D. Tucker, S. Mallat, and D. Donoho, "On denoising and best signal representation," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2225–2238, Nov. 1999.

[8] J. Rissanen, "Minimum description length denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov. 2000.

[9] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[10] L. Zhang, P. Bao, and Q. Pan, "Threshold analysis in wavelet-based denoising," *Electron. Lett.*, vol. 37, pp. 1485–1486, 2001.

Soosan Beheshti (M'03) received the B.S. degree from Isfahan University of Technology, Isfahan, Iran, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996 and 2002 respectively, all in electrical engineering.

From September 2002 to June 2005, she was a Postdoctoral Associate and a Lecturer at MIT. Since July 2005, she has been with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada, where she is currently an Assistant Professor. Her research interests include data processing, system dynamics and modeling, and statistical learning theory and generalization.

Dr. Beheshti received the MIT EECS Carlton E. Tucker Award for teaching excellence in 1998.

Munther A. Dahleh (F'01) was born in 1962. He received the B.S. degree from Texas A&M University, College Station, in 1983, and the Ph.D. degree from Rice University, Houston, TX, in 1987, all in electrical engineering.

Since then, he has been with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, where he is now a Full Professor. He was a visiting Professor with the Department of Electrical Engineering, California Institute of Technology, Pasadena, for the Spring of 1993. He has held consulting positions with several companies in the US and abroad. His interests include robust control and identification, the development of computational methods for linear and nonlinear controller design, learning from complex data, and applications of feedback control in several disciplines including automotive systems and modeling of biological systems.

Dr. Dahleh received the Ralph Budd Award in 1987 for the best thesis at Rice University, the George Axelby outstanding paper award (paper coauthored with J. B. Pearson in 1987), a National Science Foundation Presidential Young Investigator award in 1991, the Finmeccanica career development chair in 1992, the Donald P. Eckman Award from the American Control Council in 1993, and the Graduate Students Council teaching award in 1995. He was a plenary speaker at the 1994 American Control Conference and at the Mediterranean Conference on Control and Automation in 2003. He was an Associate Editor for IEEE TRANSACTIONS ON AUTOMATIC CONTROL and for *Systems and Control Letters*. He is the co-author (with I. Diaz-Bobillo) of the book *Control of Uncertain Systems: A Linear Programming Approach* (Englewood Cliffs, NJ: Prentice-Hall, 1995) and the co-author (with N. Elia) of the book *Computational Methods for Controller Design* (London, U.K.: Springer, 1998).