ON DENOISING AND SIGNAL REPRESENTATION

S. Beheshti, M.A. Dahleh
soosan@mit.edu, dahleh@mit.edu

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA 02139, U.S. A.

Keywords: Best basis, denoising, thresholding, MDL.

Abstract

The problem of signal denoising using an orthogonal basis is considered. The framework of previous solutions converts the denoising 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 calculation of the coefficient estimation error in each subspace of the basis. For each subspace, we estimate such criterion and suggest to choose the subspace for which this quantity is minimized. An information theoretic interpretation of the proposed approach introduces a new minimum description length (MDL) method of denoising. By comparison of the MDL of families of bases we can find the basis which minimizes this criterion. This offers a new method of best basis search for representation of the noisy data.

1 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. 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 [3]. The wavelet thresholding method removes the additive noise by eliminating the basis coefficients with 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 or soft thresholds are obtained by solving a minmax problem in estimation of the expected value of the reconstruction error [2]. The suggested optimal hard threshold for the basis coefficient is of order $\sqrt{2\log N/(N)}$. The method is well adapted to approximate piecewise-smooth signals. The argument however fails for the family of signals which are not smooth, i.e., the family 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 [4] proposes a thresholding method for any family of basis functions. Here the attempt is to calculate the mean-square reconstruction error of the signal as a function of any given threshold. It provides heuristic 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. For the best basis search the suggestion is to compare the error estimates for different families of bases and choose the one which minimizes such criterion.

A different denoising approach is recommended by Rissanen in [5]. In each subspace of the basis functions the normalized maximum likelihood (NML) of the noisy data is considered as the description length of the data in that subspace. The Minimum description length (MDL) denois-

ing method suggests to choose the subspace which minimizes this description length. Here noise is defined to be a part of the data that can not be compressed with the considered basis functions, while the meaningful information-bearing signal need not to be smooth. The method provides a threshold which is almost half of the suggested wavelet threshold in [3].

The new method of denoising in this paper proposes to estimate the coefficients error for each subspace of the basis functions. Such error in each subspace is the same as the reconstruction error of the noiseless signal. We introduce a method to use the noisy data error and probabilistically validate the reconstruction error. Our focus is not on setting a threshold for the coefficients beforehand, but to find the estimation error in each subspace separately and choose the subspace for which the error is minimized. Similar to MDL denoising no prior assumption on the smoothness of the noiseless part of the data is needed. We introduce a new minimum description length method of denoising and demonstrate how calculation of such description length is equivalent to finding the subspace estimation error.

2 Problem Formulation

Consider noisy data y of length N,

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

where $\bar{y}$ is the noiseless data and $w(n)$ is the additive white Gaussian noise with zero mean and variance σ_w^2. Data denoising is achieved by choosing an orthogonal basis which approximates the data with fewer nonzero coefficients than the length of data. Consider the orthogonal basis of order N, S_N. The basis vectors $s_1, s_2, \cdots, s_N$ are such that $||s_i||_2^2 = N$. Any vector of length N can be represented with such basis, therefore there exists h_is such that $\bar{y}(n) = \sum_{i=1}^{N} s_i(n)h_i$. As a result the noisy data is

$$y(n) = \sum_{i=1}^{N} s_i(n)h_i + w(n). \quad (2)$$

The least square estimate of each basis coefficient is

$$\hat{h}_i = \frac{1}{N} s_i^T y^N = h_i + \frac{1}{N} s_i^T w \quad (3)$$

where $y^N = [y(1), y(2), \cdots, y(N)]$, the observed noisy data, is a sample of random variable Y^N. The benefit of using a proper basis is that $\frac{1}{N} s_i^T w$ is almost zero as N is assumed to be large enough and we hope that there exist large number of basis vectors for which $h_i = 0$. Therefore the estimation of the noisy signal on this basis has the advantage of noise elimination. For such reason conventional basis denoising methods suggest choosing a threshold, τ, for the coefficient estimates $\hat{h}_i$'s. The denoising process is to ignore the coefficient estimates smaller than the threshold

$$\begin{aligned} \hat{h}_i &= \frac{1}{N} s_i^T y^N, \text{ if } |\frac{1}{N} s_i^T y^N| \geq \tau \\ \hat{h}_i &= 0, \quad \text{if } |\frac{1}{N} s_i^T y^N| < \tau \end{aligned} \quad (4)$$

and the estimate of the noiseless signal is

$$\hat{y}^N(n) = \sum_{i=1}^{N} s_i(n)\hat{h}_i. \quad (5)$$

A very important factor in solving the denoising problem is the behavior of the mean square reconstruction error

$$\frac{1}{N} E(||\bar{y}^N - \hat{Y}^N||_2^2). \quad (6)$$

Donoho and Johnstone provide an upperbound for the mean square error, solving a minmax problem, in wavelet denoising. They show that the optimum threshold for wavelet denoising, of a piecewise smooth signal, asymptomatically is $\sigma_w \sqrt{\frac{2 \log N}{N}}$ [3]. In [4] an estimate of the mean square error as a function of a given threshold is provided heuristically. The estimate is for any class of bases. It demonstrates that, for a class of signals, $\sigma_w \sqrt{\frac{2 \log N}{N}}$ may not necessarily provide the optimal threshold.

Instead of focusing on finding a threshold one can compare the signal estimate in different subspaces of the basis. Choosing a subspace to estimate the

data is equivalent to setting the coefficients of the basis vectors out of that subspace to zero without thresholding. MDL denoising is the first method which approaches the denoising problem with this idea. In each subspace it calculates the defined description length of the data and suggests to pick the subspace which minimizes such criterion.

Here, similar to MDL denoising approach, we investigate on estimation of a criterion which is defined for the subspaces of the basis. For each subspace S_m, $\hat{h}_{S_m}$ denotes the estimate of the coefficients in that subspace. Our goal is to find an estimate of the coefficient estimation error in each subspace, $||h-\hat{h}_{S_m}||_2^2$. Note that, as a result of the Parseval's theorem, this error is the same as the *reconstruction error* for each subspace

$$||h-\hat{h}_{S_m}||_2^2 = \frac{1}{N}||\bar{y}^N - \hat{y}^N_{S_m}||_2^2. \quad (7)$$

Because of the additive noise the coefficients error is also a random variable. The objection is to compare the worst case behavior of this error in different subspaces probabilistically. The best representative of the signal is then the signal estimate of the subspace which minimizes such criterion. In the following section we describe the method in detail. The first step is to probabilistically validate the error caused by the elimination of the basis vectors out of the subspace. Next we estimate both the mean-square and the variance of the coefficients error. The approach is similar to the quality evaluation method for impulse response estimate of an LTI system in [1].

3 New Denoising Method

Consider a subspace of order m of the orthogonal basis, S_m. We want to estimate the error of coefficient estimation in this subspace, $||h-\hat{h}_{S_m}||_2^2$. Given the noisy data in (1), we suggest the following procedure to estimate the error: For the subspace S_m, matrix A_{S_m} separates the basis vectors as follows

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} A_{S_m} & B_{S_m} \end{bmatrix} \begin{bmatrix} h_{S_m} \\ \\ \Delta_{S_m} \end{bmatrix} + w \quad (8)$$

where columns of A_{S_m} are $s_i \in S_m$, columns of B_{S_m} are basis vectors which are not in S_m, $s_i \in \bar{S}_m$, and h_{S_m} is the coefficients of the noiseless data $\bar{y}^N = [\bar{y}(1), \cdots \bar{y}(N)]^T$ in S_m. The least square estimate of coefficients in each subspace using the noisy data is

$$\hat{h}_{S_m} = \frac{1}{N}A^T_{S_m}y^N = h_{S_m} + \frac{1}{N}A^T_{S_m}w. \quad (9)$$

Therefore for the subspace error we have

$$||\hat{h}_{S_m} - h_{S_m}||_2^2 = \frac{1}{N}||A^T_{S_m}w||^2 \quad (10)$$

$$||\hat{h}_{S_m} - h||_2^2 = ||\hat{h}_{S_m} - h_{S_m}||_2^2 + ||\Delta_{S_m}||_2^2. \quad (11)$$

The additive noise has a normal distribution of $N(0,\sigma_w^2)$. Therefore y^N is an element of a Gaussian random variable Y^N and $\hat{h}_{S_m}$ is also an element of a Gaussian random variable $\hat{H}_{S_m}$. Both errors in (10) and (11) are Chi-square random variables. Expected values and variance of coefficient error $Z_{S_m} = ||\hat{H}_{S_m} - h||_2^2$ is

$$\mathrm{E}(Z_{S_m}) = \mathrm{E}||\hat{H}_{S_m} - h||_2^2 = \frac{m}{N}\sigma_w^2 + ||\Delta_{S_m}||^2 \quad (12)$$

$$\mathrm{var}(Z_{S_m}) = \mathrm{var}||\hat{H}_{S_m} - h||_2^2 = \frac{2m}{N^2}(\sigma_w^2)^2. \quad (13)$$

If the norm of the discarded vector coefficients in each subspace, $||\Delta_{S_m}||^2$, was known, how do we choose the subspace which best represent the data? The suggestion is to compare $||\hat{h}_{S_m} - h||_2^2$ of different subspaces. If we compare subspaces with same order, m, the error random variable in each subspace has the same variance of $\frac{2m^2}{N^2}(\sigma_w^2)^2$. Therefore we can only compare the expected values of the error and pick the subspace which has the minimum $||\Delta_{S_m}||_2^2$. The expected value of the error has two components, one caused by the noise and other by the ignored vector coefficients. The tradeoff between the noise related and the ignored coefficients related parts minimizes the expected value of the error for some m. This is called the bias-variance tradeoff method. Here we argue that ignoring the variance of the random variable can be problematic. For example, what if we are comparing two subspaces with different orders? Instead of comparing only the expected values, lets compare an *event* happening in each

subspace with same probability. In this case both expected value and variance of the random variable might be involve in our decision. Assume that for a particular $\bar{m}$ the expected value of error is minimized. Therefore the expected value for the subspace of order $\bar{m}-1$, $E_{S_{\bar{m}-1}}$, is larger than $E_{S_{\bar{m}}}$. However the variance of the error in $S_{\bar{m}-1}$ is smaller than the variance in $S_{\bar{m}}$. Therefore when we are comparing two events in these two spaces, which occur with same probability, the worst case error might be smaller in space $S_{\bar{m}-1}$ than in $S_{\bar{m}}$.

The *event*, we consider in each subspace, is that the random variable $z_{S_m} = ||\hat{h}_{S_m} - h||_2^2$ is around its mean with a given probability $P1$

$$\Pr\{|Z_{S_m} - E(Z_{S_m})| \leq D_{S_m}\} = P1. \qquad (14)$$

Therefore D_{S_m} is a function of $||\Delta_{S_m}||$, σ_w, m and $P1$, and for each subspace S_m with probability $P1$ the error is between the following bounds

$$\frac{m}{N}\sigma_w^2 + ||\Delta_{S_m}||^2 \pm D_{S_m}(P1, \sigma_w, m, ||\Delta_{S_m}||). \quad (15)$$

To find the optimal subspace we suggest to choose the subspace which minimizes the worst case error with the same probability $P1$,

$$\begin{aligned} S_m^* &= \arg\min_{S_m}\{E(Z_{S_m}) + D_{S_m}\} \qquad (16) \\ &= \arg\min_{S_m}\{\frac{m}{N}\sigma_w^2 + ||\Delta_{S_m}||_2^2 + D_{S_m}\} \qquad (17) \end{aligned}$$

In the example we discussed previously, because the variance of error in $S_{\bar{m}-1}$ is lower than the variance in $S_{\bar{m}}$. Therefore $D_{S_{\bar{m}-1}}$, which depends on the variance, might be less than $D_{S_{\bar{m}}}$ and the worst case error in $S_{\bar{m}-1}$ might be less than that of $S_{\bar{m}}$. It is important to mention that since the variance of error is of order $\frac{1}{N^2}$, for large enough N we are able to pick $P1$ close to one and still have a bounded number for D_{S_m}. We will discuss this issue later in detail.

So far the argument was with the assumption that $||\Delta_{S_m}||$ is known. In our problem setting however $||\Delta_{S_m}||$ is unknown. To use a similar approach, we next suggest a method to probabilistically validate $||\Delta_{S_m}||$ using the observed noisy data.

3.0.1 Estimation of $||\Delta_{S_m}||$

In each subspace the data representation error is

$$\begin{aligned} \frac{1}{N}||y^N - \hat{y}_{S_m}^N||_2^2 &= \frac{1}{N}||B_{S_m}\Delta_{S_m} + G_{S_m}w||^2 \\ &= ||(\Delta_{S_m} + v)||^2 \qquad (18) \end{aligned}$$

where $\hat{y}_{S_m}^N = A_{S_m}\hat{h}_{S_m}$ and $G_{S_m} = (I - \frac{1}{N}A_{S_m}A_{S_m}^T) = \frac{1}{N}B_{S_m}B_{S_m}^T$ is a projection matrix. Therefore, $G_{S_m}w = v$ where v_i's are independent Gaussian random variables. Note that using the Parseval's theorem we already know that

$$\frac{1}{N}||y^N - \hat{y}_{S_m}^N||_2^2 = ||\frac{1}{N}B_{S_m}^T y^N||_2^2 = \sum ||\hat{h}_{\bar{S}_m}||^2. \quad (19)$$

The data error $X_{S_m} = \frac{1}{N}||Y - \hat{Y}_{S_m}||_2^2$ is also a Chi-square random variable for which

$$\begin{aligned} E(X_{S_m}) &= (1 - \frac{m}{N})\sigma_w^2 + ||\Delta_{S_m}||^2 \qquad (20) \\ \text{var}(X_{S_m}) &= \frac{2\sigma_w^2}{N}((1 - \frac{m}{N})\sigma_w^2 + 2||\Delta_{S_m}||^2). \end{aligned}$$

Given the noisy data, one sample of this random variable is available. We call this observed error x_{S_m}. Note that the variance of the data error is of order $\frac{1}{N}$ of its expected value. Therefore one method of estimating $||\Delta_{S_m}||$ is to assume that this one sample is a good estimate of its expected value,

$$||\hat{\Delta}_{S_m}||_2^2 \approx x_{S_m} - (1 - \frac{m}{N})\sigma_w^2. \qquad (21)$$

This can be a convenient method of estimation of $||\Delta_{S_m}||$ when N is large enough. However, since we want to use the estimate to compare the different subspaces, we have to be more precise in the estimation process: Each X_{S_m} has a different variance and the confidence of the estimate is different for each of the subspaces even as N grows. So how "relatively" close we are to the estimate in each subspace is very important. As a result we suggest the following validation method for estimation and comparison of $||\Delta_{S_m}||_2^2$ in different subspaces.

The Chi-square probability distribution of the data error is a function of $||\Delta_{S_m}||$ and the noise

variance, $f_{X_{S_m}}(x_{S_m}; m, \sigma_w, ||\Delta_{S_m}||)$. We suggest validating $||\Delta_{S_m}||$ such that X_m is in the neighborhood of its mean with probability $P2$, i.e., validate $f_{X_{S_m}}(x_{S_m}; m, \sigma_w, ||\Delta_{S_m}||)$, and therefore $||\Delta_{S_m}||$, such that

$$\Pr(|X_{S_m} - E(X_{S_m})| \leq J_{S_m}) = P2. \quad (22)$$

The bound J_{S_m} is a function of $||\Delta_{S_m}||$, σ_w^2, m, and $P2$, $J_{S_m}(P2, \sigma_w^2, m, ||\Delta_{S_m}||)$. Therefore for each subspace S_m, with validation probability $P2$, we find U_{S_m} and L_{S_m}, the upper bound and lower bound on $||\Delta_{S_m}||$, $L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m}$.

3.0.2 Subspace Comparison

Using the estimate of $||\Delta_{S_m}||$ from the previous section, we can estimate the worst case error criterion in (16). The validation part finds bounds on $||\Delta_{S_m}||_2^2$. Therefore we suggest to pick m^* such that

$$\begin{aligned} S_m^* = & \arg \min_{S_m} \max_{||\Delta_{S_m}|| \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) \\ & + D_{S_m}(P1, \sigma_w, m, ||\Delta_{S_m}||)\}. \end{aligned} \quad (23)$$

The worst case estimate in each subspace is given with confidence probability $P1$ and validation probability $P2$. The confidence region of error here is between

$$\max_{||\Delta_{S_m}|| \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) + D_{S_m}\}, \quad (24)$$

and

$$\min\{0, \min_{||\Delta_{S_m}|| \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) - D_{S_m}\}\}. \quad (25)$$

Note that one choice for J_{S_m} in (22) is $J_{S_m} = \beta \text{var} X_{S_m}$. In this case using the Chebychev inequality we have

$$P2 \geq 1 - \frac{1}{\beta^2} \text{ or } \beta \leq \sqrt{\frac{1}{1-P2}} \quad (26)$$

which shows how β and $P2$ are related. How close β is to $\sqrt{\frac{1}{1-P2}}$ depends on the distribution of the error in each subspace.

3.1 Gaussian Estimation

In both the probabilistic and validation part we use the table of Chi-square distribution. However, in this setting we can use the central limit theorem(CLT) to approximate the Chi-square distributions with Gaussian distributions. This gives us the advantage of finding a mathematical expression for the error bounds and worst case error (23),(24), (25) as a function of $P1, \sigma_w, m, P2$ and the observed noisy signal.

3.1.1 Data Error and Estimation of $||\Delta_{S_m}||$

The data error (18) is of form

$$\begin{aligned} \frac{1}{N}||y^N - \hat{y}_{S_m}^N||_2^2 &= ||\Delta_{S_m} + v||^2 \\ &= \sum_{i=1}^{N-m} (\delta_i + v_i)^2 \end{aligned} \quad (27)$$

where v_i's are zero mean white Gaussian random variables with variance $\frac{\sigma_w^2}{N}$. If $N - m$ is large enough we can estimate the Chi-square distribution of the data error with a Gaussian distribution. For a Gaussian random variable X with mean m_X and variance σ_X^2 we have

$$\Pr(m_X - \alpha\sigma_X < X < m_X + \alpha\sigma_X) = Q(\alpha), \quad (28)$$

where $Q(\alpha) = \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. For the data error, $X_{S_m} = \frac{1}{N}||Y - \hat{Y}_{S_m}||_2^2$,

$$\mathrm{E}(X_{S_m}) = m_w + m_\delta, \quad (29)$$

$$\text{var}(X_{S_m}) = \frac{2\sigma_w^2}{N}(m_w + 2m_\delta), \quad (30)$$

where $m_w = (1 - \frac{m}{N})\sigma_w^2$ and $m_\delta = ||\Delta_{S_m}||_2^2$. Using the one observed data error given the noisy data, x_{S_m}, with probability $Q(\alpha)$ we have

$$|x_{S_m} - (m_w + m_\delta)| \leq \alpha \sqrt{\frac{4\sigma_w^2}{N} m_\delta + v_m}, \quad (31)$$

where $v_m = \frac{2}{N}(1 - \frac{m}{N})\sigma_w^4$.

Lemma 1 Validation of (31) for feasible $||\Delta_{S_m}||$s provides the following upper and lower bound for $||\Delta_{S_m}||_2^2$

$$L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m}, \quad (32)$$

where

- If $x_{S_m} \leq (m_w - \alpha\sqrt{v_m})$, there is no valid $||\Delta_{S_m}||$ given the data.
- If $(m_w - \alpha\sqrt{v_m}) \leq x_{S_m} \leq (m_w + \alpha\sqrt{v_m})$,

$$L_{S_m} = 0 \quad (33)$$

$$U_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha). \quad (34)$$

where

$$K_{S_m}(\alpha) = \quad 2 \quad \alpha\frac{\sigma_w}{\sqrt{N}} \times \sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}. \quad (35)$$

- If $(m_w + \alpha\sqrt{v_m}) \leq x_{S_m}$,

$$L_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} - K_{S_m}(\alpha) \quad (36)$$

$$U_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha). \quad (37)$$

Proof In Appendix A. $\diamondsuit$

Note that to avoid the first case we have to increase α such that

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1 - \frac{m}{N} - \frac{x_{S_m}}{\sigma_w^2}\right). \quad (38)$$

3.1.2 Comparison of Subspaces

For the error in each subspace S_m, $Z_{S_m} = ||\hat{H}_{S_m} - h||^2$, in (11), we have

$$||\hat{h}_{S_m} - h||^2 = ||\Delta_{S_m}||_2^2 + \sum_{i=1}^{m} u_i^2, \quad (39)$$

where u_is are zero mean white Gaussian noises with variance $\frac{\sigma_w^2}{N}$. If m is large enough we can estimate the Chi-square distribution of the error with a Gaussian distribution. Then the probabilistic bounds on this error are provided as following. With probability $Q(\beta)$ we have

$$|Z_{S_m} - E(Z_{S_m})| \leq \beta\sqrt{\text{var}Z_{S_m}}. \quad (40)$$

The bounds on expected value and variance of Z_{S_m}, in (12) and (13), can be calculated by using the bounds from lemma one. Therefore, the worst case error bound in subspace S_m with probability $Q(\beta)$ and validation probability of $Q(\alpha)$ is

$$E(Z_{S_m}) + D_{S_m} = \frac{m}{N}\sigma_w^2 + U_{S_m} + \beta\frac{\sqrt{2m}}{N}\sigma_w^2. \quad (41)$$

For the choice of optimum subspace we suggest to choose m^* for which, from (23),

$$S_m^* = \arg\min_{S_m}\{\frac{m}{N}\sigma_w^2 + U_{S_m} + \beta\frac{\sqrt{2m}}{N}\sigma_w^2\}, \quad (42)$$

which is the bound valid with probability $Q(\beta)$ and validation probability of $Q(\alpha)$.

3.1.3 Proper Choice of α and β

In order to have the probability of validation close to one, α and β should be as large as possible. Simultaneously, to have limited tight bounds, at both stages of finding bounds on $||\Delta_{S_m}||$ in lemma 1 and finding bounds for the subspace error in (40), we have to choose $\alpha/\sqrt{N}$ and β/N^2 small enough. Also as a result of the validation stage in lemma one, another necessary condition is that α satisfies the inequality in (38). Note that with this proper choice of α and β, the upper and lower bounds provided in (24) and (25) can be used to evaluate the quality of estimate of each subspace.

4 New MDL Denoising and Best Basis Search

In two-stage MDL the assumption is that the length of the code describing any element of subspace S_m is the same and is of order $m\frac{log(N)}{N}$. The probability distribution of the noisy data, in subspace S_m, is defined as $f_{S_m}(y_{S_m}; \hat{h}_{S_m})$. This is the probability distribution of y_{S_m} given it is generated by $\hat{h}_{S_m} \in S_m$, where $\hat{h}_{S_m}$ is the maximum likelihood(ML) estimate of h in S_m. The description length of the data in each subspace is the code which describes the noisy data using the coefficient estimates in that subspace. The famous Shannon coding method is to choose a code such that the length of the code length is proportional to the logarithm of the inverse of the probability distribution. Then the data code is of order

$\log \frac{1}{f_{S_m}(y;\hat{h}_{S_m})}$. Therefore the complete two-stage description length of the noisy data, y, in each subspace is

$$\mathrm{DL}(S_m) = m\frac{log(N)}{N} + \log \frac{1}{f_{S_m}(y;\hat{h}_{S_m})}. \quad (43)$$

The two-stage MDL method is to pick the subspace which minimizes this criterion. In [5], the MDL denoising method is derived with same basic idea. Here the criterion is the normalized maximum-likelihood (NML) density function. Calculation of $f_{S_m}(y;\hat{h}_{S_m})$ is a part of the calculation of this criterion as well.

One important fact is that the calculation of $f_{S_m}(y;\hat{h}_{S_m})$ is meaningful only if y has been generated with an element of S_m. The conditional probability distribution function in S_m is defined for elements which can be represented in form of

$$y_{S_m} = \sum_{s_i \in S_m} s_i h_i + w. \quad (44)$$

In MDL methods the ignored basis vectors effects, $\sum_{s_i \in \bar{S}_m} s_i h_i$, is considered as a part of the additive *zero* mean noise. If such effects are indeed nonzero, the new defined noise is not anymore zero mean and it contradicts the prior assumption on the noise to be zero mean. As a result of such approach the estimate of the noise variance in different subspaces are different, even though the ignored coefficient part only effects the mean and not the variance of estimates. This causes problem in the evaluation of the description length even if the true number of the basis vectors, which generated the noiseless data $\bar{y}$, is finite. Consider an example for which only h_1 and h_3 are nonzero and the prior knowledge is that only two basis vectors are enough to represent noiseless data. With the prior assumption that the additive noise is zero mean, the description length in (43) can be calculated. However, except for one subspace of order two, $\{s_1, s_3\}$, such assumption is not valid and the mean of the noise is the effects of $h_1 s_1$ and/or $h_3 s_3$.

Here we define a new subspace description length for which this prior assumption inconsistency is avoided. We use the true model h, which is unknown, to code not the noisy data but the estimate of the noisy data in each subspace. The prior assumption is that the noisy data is generated with h

$$y = \sum_{i=1}^{N} h_i s_i + w = \sum_{s_i \in S_m} s_i h_i + \sum_{s_i \in \bar{S}_m} s_i h_i + w. \quad (45)$$

The code length of any signal of length N, considering the distribution of y is of form

$$L(x) = \log \frac{1}{f_h(x;h)} \quad (46)$$

$$= \log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||x-\bar{y}||^2}{2\sigma_w^2}} \quad (47)$$

where $\bar{y} = \sum_{i=1}^{N} s_i h_i$ is the expected value of random variable Y. We define the description length of S_m as the code length of the estimate of data using the estimate of h in S_m,

$$\mathrm{DL}_h(S_m) = L(\hat{y}_{S_m}) = \log \frac{1}{f_h(\hat{y}_{S_m};h)} \quad (48)$$

$$= \log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\hat{y}_{S_m}-\bar{y}||^2}{2\sigma_w^2}} \quad (49)$$

Comparison of such description length for different subspaces leads to comparison of the reconstruction error

$$\frac{1}{N}||\hat{y}_{S_m} - \bar{y}||^2 = ||\hat{h}_{S_m} - h||^2 \quad (50)$$

which was the main goal in this paper and has been discussed in the previous sections.

For a given noisy data one might proceed to search for the basis which would best represent the data. We suggest to compare the new proposed MDL of different families of basis functions. The method leads to the choice of the basis which minimizes such criterion.

4.1 Thresholding Denoising Methods

In threshold methods, a threshold τ, is provided before finding the estimates of coefficients. It is not known that for this choice of threshold how

many coefficient estimates, which are less than τ, are due to the additive noise only. In cases where we know a priori that there are few nonzero coefficients to represent the noiseless part of the data, it might be intuitive to pick the threshold only as a function of the variance of the noise and the length of the data, as it is suggested in [3]. But as [4] shows without such prior assumption it is not trivial to decide on the optimum threshold beforehand.

What we showed in our method is that the critical possible thresholds are the absolute values of the coefficient estimates. Lets sort the basis vectors based on the absolute value of the coefficient estimates. Our method is computing the estimation error for any of those absolute values as the threshold. We find the optimal of those thresholds comparing the error estimation of such thresholding. Depending on the tradeoff between the eliminated coefficients and the noise effects there is a subspace for which the estimation error is minimized.

Similar to our approach, MDL denoising suggests a criterion to be calculated for different subspaces. However, as we discussed previously, in this method the effect of the eliminated coefficients in each subspace is considered as a part of the additive noise. The comparison of the NML criterion for different subspaces asymptotically provides a threshold which is only a function of the variance and the length of the data [5]. As we argued previously, in the proposed method in this paper, there is a distinction between the noise effects and the eliminated coefficients effects in different subspaces. Therefore even the asymptotic results can not provide a threshold which is only a function of the noise variance and the length of the data. The optimal threshold is sensitive to the coefficient estimates of all the basis vectors.

5 Simulation

The unit-power signal shown in Figure (1) used to illustrate the performance of the proposed method. Figure (2) shows the absolute value of the discrete Fourier transform of the signal. Figure (3) shows the subspace error of the noiseless data. The subspace of order m is the one among the subspaces of same order which minimizes the error. As we expect such error decreases as the subspace order increases. Figure (4) shows the subspace error in presence of additive noise with variance 0.25. It shows that the subspace error in this scenario is minimum for S_7 and our method also picks S_7.

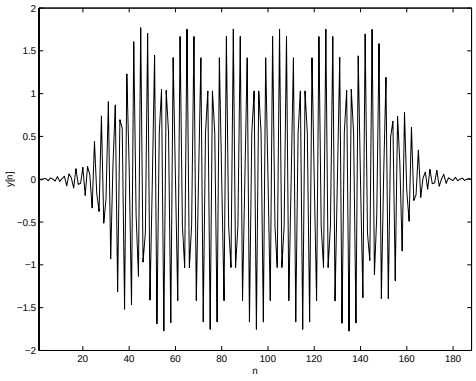

Figure 1: Noiseless unit-power signal of length 188.

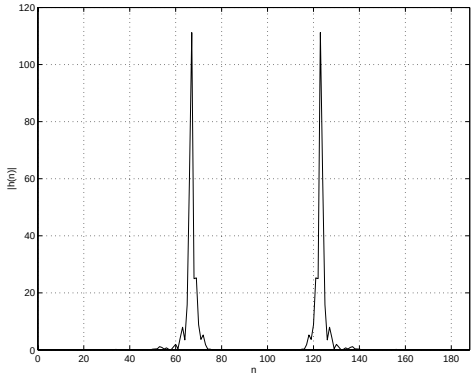

Figure 2: 188 points discrete Fourier transform of the signal.

6 Conclusion

A new approach to the denoising problem based on best basis solution was proposed. We showed how to use the one sample of the data error in each subspace of the basis and probabilistically validate the data estimate in each subspace. The criterion for comparison of the different subspaces

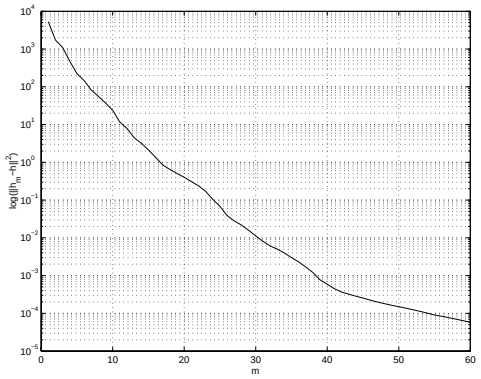

Figure 3: Subspace error for the noiseless signal for subspaces with order m. The subspace with order m is the one among all the subspaces with same order which minimizes the error.

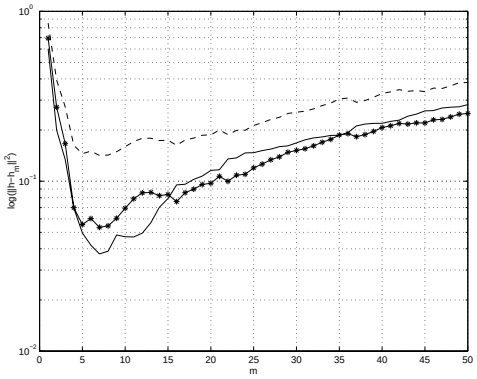

Figure 4: Subspace error for the noisy signal for subspaces with order m. The subspace with order m is the one among all the subspaces with same order which minimizes the error. Noise variance is $\sigma_w^2 = .25$. The solid line is the subspace error. The line with "*" is the estimate of the expected value of subspace error using the proposed method. The dashed line is the error upper bound, U_{S_m} in (41), with $\alpha = \log(N)/2$ and $\beta = \log(N)$.

is defined as the reconstruction error in each subspace. To extract the most information from the noisy data we were not able to provide a threshold before estimating all the basis vector coefficients. We calculated the proposed criterion asymptotically when the length of data is large enough. However, the important advantage of this method is that the criterion can be calculated for any finite length data and without any asymptotic approximations.

The proposed approach can be viewed from the information theoretical perspective as a new MDL method. We compared the new MDL approach with the existing MDL denoising method. We showed how the definition of noise in existing MDL denoising causes some inconsistency in the process of MDL calculation and demonstrated the advantages of the new proposed MDL.

Although here the additive noise is assumed to be Gaussian with known variance, the method can be generalized for any independent identically distributed additive noise, with even an unknown variance. The proposed denoising method is comparable with the current order estimation methods in blind channel identification.

A Proof of Lemma 1

Define $\bar{x}_{S_m} = x_{S_m} - (1 - \frac{m}{N})\sigma_w^2$, we want to validate $m_\delta = ||\Delta_{S_m}||^2$ for which

$$\begin{aligned} m_\delta - \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_m} \quad &\leq \qquad (51) \\ \bar{x}_{S_m} \leq m_\delta \quad &+ \quad \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_m} \end{aligned}$$

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

Lower Bound on m_δ

$$\bar{x}_{S_m} - m_\delta < \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_m} \qquad (52)$$

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

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

$$(\bar{x}_{S_m} - m_\delta)^2 = \alpha^2(\frac{4\sigma_w^2}{N}m_\delta + v_m) \qquad (53)$$

which is

$$L_{m_\delta} = \bar{x}_{S_m} + \frac{2\sigma_w^2\alpha^2}{N} - \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w} \quad (54)$$

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

Upper Bound on m_δ

$$m_\delta - \bar{x}_{S_m} > \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_m} \quad (55)$$

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

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

$$(\bar{x}_{S_m} - m_\delta)^2 = \alpha^2(\frac{4\sigma_w^2}{N}m_\delta + v_m) \quad (56)$$

which is

$$U_{m_\delta} = \bar{x}_{S_m} + \frac{2\alpha^2\sigma_w^2}{N} + \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}. \quad (57)$$

References

[1] S. Beheshti and M.A. Dahleh. "On model quality evaluation of stable LTI systems", *Proceedings of the 39th IEEE Conference on Decision and Control*, vol.3, pp. 2716 -2721, (2000).

[2] D. Donoho. "De-noising by Soft Thresholding", *IEEE Trans. on Information Theory,* vol.41, pp. 613-627, (1995).

[3] D. Donoho, I. M. Johnstone. "Ideal Spatial Adaptation by Wavelet Shrinkage", *Biometrika*, pp. 425-455, (1994).

[4] H. Krim, D. Tucker, S. Mallat, D. Donoho. "On Denoising and Best Signal Representation", *IEEE Trans. on Information Theory,* vol.45, pp. 2225-2238, (1999)

[5] J. Rissanen. "MDL Denoising", *IEEE Trans. on Information Theory,* vol.46, pp. 2537-2543, (2000).