Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion

Naoki Saito

Schlumberger-Doll Research
Old Quarry Road, Ridgefield, CT 06877-4108
and
Department of Mathematics
Yale University
10 Hillhouse Avenue, New Haven, CT 06520

ABSTRACT

We describe an algorithm to estimate a discrete signal from its noisy observation, using a library of orthonormal bases (consisting of various wavelets, wavelet packets, and local trigonometric bases) and the information-theoretic criterion called minimum description length (MDL). The key to effective random noise suppression is that the signal component in the data may be represented efficiently by one or more of the bases in the library, whereas the noise component cannot be represented efficiently by any basis in the library. The MDL criterion gives the best compromise between the fidelity of the estimation result to the data (noise suppression) and the efficiency of the representation of the estimated signal (signal compression): it selects the "best" basis and the "best" number of terms to be retained out of various bases in the library in an objective manner. Because of the use of the MDL criterion, our algorithm is free from any parameter setting or subjective judgments.
This method has been applied usefully to various geophysical datasets containing many transient features.

1. INTRODUCTION

Wavelet transforms and their relatives such as wavelet packet transforms and local trigonometric transforms are becoming increasingly popular in many fields of applied sciences. So far their most successful application area seems to be data compression.[24,8,13,5,32,28] Meanwhile, several researchers claimed that wavelets and these transforms are also useful for reducing noise in (or denoising) signals/images.[14,9,19] In this paper, we take advantage of both sides: we propose an algorithm for *simultaneously* suppressing random noise in data and compressing the signal, i.e., we try to "kill two birds with one stone."

Throughout this paper, we consider a simple degradation model: observed data consists of a signal component and additive white Gaussian noise (WGN). Our algorithm estimates the signal component from the data using a library of orthonormal bases (including various wavelets, wavelet packets, and local trigonometric bases) and the information-theoretic criterion called the Minimum Description Length (MDL) criterion for discriminating signal from noise.

The key motivation here is that the signal component in the data can often be efficiently represented by one or more of the bases in the library whereas the noise component cannot be represented efficiently by any basis in the library. The use of the MDL criterion frees us from any subjective parameter setting such as threshold selection. This is particularly important for real, field data where the noise level is difficult to obtain or estimate *a priori*.

The organization of this paper is as follows. In Section 2, we formulate our problem. We view the problem of simultaneous noise suppression and signal compression as a model selection problem out of models generated by "a library of orthonormal bases." In Section 3, we review the MDL principle which plays a critical role in this paper. We also give some simple examples to help understand its concept. In Section 4, we develop an actual algorithm of simultaneous noise suppression and signal compression. We also give the computational complexity of our algorithm. Then, we extend our algorithm for higher dimensional signals (images) in Section 5. In Section 6, we apply our algorithm to several geophysical datasets, both synthetic and real, and compare the results with other competing methods. We discuss the connection of our algorithm with different approaches in Section 7, and finally, we conclude in Section 8.

2. THE FORMULATION OF THE PROBLEM

Let us consider a discrete degradation model

$$\boldsymbol{d} = \boldsymbol{f} + \boldsymbol{n}, \tag{2.1}$$

where $\boldsymbol{d}, \boldsymbol{f}, \boldsymbol{n} \in \mathbb{R}^N$ and $N = 2^n$. The vector $\boldsymbol{d}$ represents the noisy observed data and $\boldsymbol{f}$ is the unknown true signal to be estimated. The vector $\boldsymbol{n}$ is the WGN, i.e., $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. Let us assume that σ^2 is unknown.

We now consider an algorithm to estimate $\boldsymbol{f}$ from the noisy observation $\boldsymbol{d}$. To do this, we need an efficient representation scheme for the signal $\boldsymbol{f}$ which may consist of various transient features or edges. For this purpose, we prepare *a library of orthonormal bases* including the standard Euclidean basis of $\mathbb{R}^N$, the Haar-Walsh bases, various wavelet bases and wavelet packet best-bases generated by Daubechies's QMFs, their less asymmetric versions (i.e., coiflets), and local trigonometric best-bases. This collection of bases is highly adaptable and versatile for representing various transient signals. For example, if the signal consists of blocky functions such as acoustic impedance profiles of subsurface structure, the Haar-Walsh bases capture those discontinuous features both accurately and efficiently. If the signal consists of piecewise polynomial functions of order p, then the Daubechies wavelets/wavelet packets with the filter length $L > 2(p+1)$ or the coiflets with the filter length $L > 3(p+1)$ would be efficient because of the vanishing moment property. If the signal has a sinusoidal shape or highly oscillating characteristics, the local trigonometric bases would do the job. Moreover, computational efficiency of this library is also attractive; the most expensive expansion in this library, i.e., the local trigonometric expansion, costs about $O(N[\log_2 N]^2)$ as explained in the previous section. More detailed properties of these bases can be found in the literature.[1,8,12,21,20,31]

Let us denote this library $\mathcal{L} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_M\}$, where $\mathcal{B}_m$ represents one of the orthonormal bases in the library, and M (typically 5 to 20) is the number of bases in this library. If we want, we can add other orthonormal bases in this library such as the Karhunen-Loève basis or the prolate spheroidal wave functions. However, normally, the above-mentioned multiresolution bases are more than enough, considering their versatility and computational efficiency.

Since the bases in the library $\mathcal{L}$ compress signals/images very well, we make a strong assumption here: we suppose the unknown signal $\boldsymbol{f}$ can be *completely* represented by k ($< N$) elements of a basis $\mathcal{B}_m$, i.e.,

$$\boldsymbol{f} = \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}, \tag{2.2}$$

where $\boldsymbol{W}_m \in \mathbb{R}^{N \times N}$ is the orthogonal matrix whose column vectors are the basis elements of $\mathcal{B}_m$, and $\boldsymbol{\alpha}_m^{(k)} \in \mathbb{R}^N$ is the vector of expansion coefficients of $\boldsymbol{f}$ with only k non-zero coefficients. At this point, we do not know the actual value of k and the basis $\mathcal{B}_m$. However, we regard (2.2) as a *model* which should be selected from the library $\mathcal{L}$. Now the problem of simultaneous noise suppression and signal compression can be stated as follows: *find the "best" k and m given the library $\mathcal{L}$.* In other words, we translate the estimation problem into a model selection problem where models are the bases $\mathcal{B}_m$ and the number of terms k under the additive WGN assumption.

For data compression purposes, we want to have k as small as possible. At the same time, we want to minimize the distortion between the estimate and the true signal by choosing the most suitable basis $\mathcal{B}_m$, where the larger k normally gives smaller value. How can we satisfy these seemingly conflicting demands?

3. THE MINIMUM DESCRIPTION LENGTH PRINCIPLE

To satisfy the above mentioned conflicting demands, we need a model selection criterion. One of the most suitable criteria for our purpose is the so-called *Minimum Description Length* (MDL) criterion proposed by Rissanen,[25],[26].[27] The MDL principle suggests that the "best" model among the given collection of models is the one giving the shortest description of the data *and* the model itself. For each model in the collection, the length of description of the data is counted as a codelength of encoding the data using that model in binary digits (bits). The length of description of a model is a codelength of specifying that model, e.g., the number of parameters and their values if it is a parametric model.

To help understand what "code" or "encoding" means, we give some simple examples. We assume that we want to transmit data by first encoding (mapping) them into a bitstream by an encoder, then receive the bitstream by a decoder, and finally try to reconstruct the data. Let $L(\boldsymbol{x})$ denote the codelength (in bits) of a vector $\boldsymbol{x}$ of deterministic or probabilistic parameters of either real-valued, integer-valued, or taking values in a finite alphabet.

Example 4.1: Codelength of symbols drawn from a finite alphabet. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$ be a string of symbols drawn from a finite alphabet $\mathcal{X}$ with the independently and identically distributed (i.i.d.) probability mass function $p(x), x \in \mathcal{X}$. In this case, clearly the frequently occurring symbols should have shorter codelengths than rarely occurring symbols for efficient communication. This leads to the so-called Shannon code[10] whose codelength (if we ignore the integer requirement for the codelength) can be written as

$$L(x) = -\log p(x) \qquad \text{for } x \in \mathcal{X}. \tag{3.1}$$

(From now on, we denote the logarithm of base 2 by "$\log$", and the natural logarithm, i.e., base e by "$\ln$".) The Shannon code has the shortest codelength *on average*, and satisfies the so-called the Kraft inequality[10]:

$$\sum_{x \in \mathcal{X}} 2^{-L(x)} \leq 1, \tag{3.2}$$

which is necessary and sufficient for the existence of an instantaneously decodable code, i.e., a code such that there is no codeword which is the prefix of any other codeword in the coding system. The shortest codelength on average for the whole sequence $\boldsymbol{x}$ becomes

$$L(\boldsymbol{x}) = \sum_{i=1}^{N} L(x_i) = -\sum_{i=1}^{N} \log p(x_i). \tag{3.3}$$

Example 4.2: Codelength of deterministic integers. For a deterministic parameter $j \in \mathbb{Z}_N = (0, 1, \dots, N-1)$ (i.e., both the encoder and decoder know N), the codelength of describing j is written as $L(j) = \log N$ since $\log N$ bits are required to specify N. This can also be interpreted as a codelength using Shannon code for a sample drawn from the uniform distribution over $(0, 1, \dots, N-1)$.

Example 4.3: Codelength of an integer (universal prior for an integer). Suppose we do not know how large a natural number j is. Rissanen proposed in[25] that the code of such j should be the binary representation of j, preceded by the code describing its length $\log j$, preceded by the code describing the length of the code for $\log j, \dots$. This recursive strategy leads to

$$L^*(j) = \log^* j + \log c_0 = \log j + \log\log j + \cdots + \log c_0, \tag{3.4}$$

where the sum involves only the non-negative terms and the constant $c_0 \approx 2.865064$ which was computed to satisfy the Kraft inequality with equality, i.e., $\sum_{j=1}^{\infty} 2^{-L^*(j)} = 1$. This can be generalized for an integer j by defining

$$L^*(j) = \begin{cases} 1/2 & \text{if } j = 0, \\ 1/2\ (\log^* |j| + \log 2c_0) & \text{otherwise.} \end{cases} \tag{3.5}$$

(We can easily see that (3.5) satisfies $\sum_{j=-\infty}^{\infty} 2^{-L^*(j)} = 1$.)

Example 4.4: Codelength of a truncated real-valued parameter. For a deterministic real-valued parameter $v \in \mathbb{R}$, the exact code generally requires infinite length of bits. Thus, in practice, some truncation must be done for transmission. Let δ be the precision and v_δ be the truncated value, i.e., $|v - v_\delta| < \delta$. Then, the number of bits required for v_δ is the sum of the codelength of its integer part $[v]$ and the number of fractional binary digits of truncation precision δ, i.e.,

$$L(v_\delta) = L^*([v]) + \log(1/\delta). \tag{3.6}$$

Having gone through the above examples, now we can state the MDL principle more clearly. Let $\mathcal{M} = \{\boldsymbol{\theta}_m : m = 1, 2, \dots\}$ be a class or collection of models at hand. The integer m is simply an index of a model in the list. Let $\boldsymbol{x}$ be a sequence of observed data. Assume that we do not know the true model $\boldsymbol{\theta}$ generating the data $\boldsymbol{x}$. As in,[27],[22] given the index m, we can write the codelength for the whole process as

$$L(\boldsymbol{x}, \boldsymbol{\theta}_m, m) = L(m) + L(\boldsymbol{\theta}_m \mid m) + L(\boldsymbol{x} \mid \boldsymbol{\theta}_m, m). \tag{3.7}$$

This equation says that the codelength to rewrite the data is the sum of the codelengths to describe (1) the index m, (2) the model $\boldsymbol{\theta}_m$ given m, and (3) the data $\boldsymbol{x}$ using the model $\boldsymbol{\theta}_m$. The MDL criterion suggests picking the model $\boldsymbol{\theta}_{m^*}$ which gives the minimum of the total description length (3.7).

The last term of the right-hand side (RHS) of (3.7) is the length of the Shannon code of the data assuming the model $\boldsymbol{\theta}_m$ is the true model, i.e.,

$$L(\boldsymbol{x} \mid \boldsymbol{\theta}_m, m) = -\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_m, m), \tag{3.8}$$

and the maximum likelihood (ML) estimate $\widehat{\boldsymbol{\theta}}_m$ minimizes (3.8) by the definition:

$$L(\boldsymbol{x} \mid \widehat{\boldsymbol{\theta}}_m, m) = -\log p(\boldsymbol{x} \mid \widehat{\boldsymbol{\theta}}_m, m) \leq -\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_m, m). \tag{3.9}$$

However, in practice, we must truncate the real-valued parameters specifying the models in $\mathcal{M}$ as shown in Example 4 above. In other words, we must use the truncated version of the ML estimate. The finer truncation precision we use, the smaller the term (3.9), but the larger the term $L(\widehat{\boldsymbol{\theta}}_m \mid m)$ becomes. Suppose that the model $\boldsymbol{\theta}_m$ has k_m real-valued parameters, i.e., $\boldsymbol{\theta}_m = (\theta_{m,1}, \dots, \theta_{m,k_m})$. Rissanen showed in,[2527] that the optimized truncation precision (δ^*) is of order $1/\sqrt{N}$ and

$$\begin{aligned}\min_{\delta} L(\boldsymbol{x}, \boldsymbol{\theta}_{m,\delta}, m, \delta) &= L(m) + L(\widehat{\boldsymbol{\theta}}_{m,\delta^*} \mid m) + L(\boldsymbol{x} \mid \widehat{\boldsymbol{\theta}}_{m,\delta^*}, m) + O(k_m) \\ &\approx L(m) + \sum_{j=1}^{k_m} L^*([\widehat{\theta}_{m,j}]) + \frac{k_m}{2}\log N + L(\boldsymbol{x} \mid \widehat{\boldsymbol{\theta}}_m, m) + O(k_m),\end{aligned} \tag{3.10}$$

where $\widehat{\boldsymbol{\theta}}_m$ is the optimal non-truncated value given m, and $L^*(\cdot)$ is defined in (3.6). We note that the last term $O(k_m)$ in the approximation in (3.10) includes the penalty codelength necessary to describe the data $\boldsymbol{x}$ using the truncated ML estimate $\widehat{\boldsymbol{\theta}}_{m,\delta^*}$ instead of the true ML estimate $\widehat{\boldsymbol{\theta}}_m$. For sufficiently large N, the last term may be omitted, and instead of minimizing the ideal codelength (3.7), Rissanen proposed to minimize

$$MDL(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}_m, m) = L(m) + \sum_{j=1}^{k_m} L^*([\widehat{\theta}_{m,j}]) + \frac{k_m}{2}\log N + L(\boldsymbol{x} \mid \widehat{\boldsymbol{\theta}}_m, m). \tag{3.11}$$

The minimum of (3.11) gives the best compromise between the low complexity in model and high likelihood on the data.

The first term of the RHS of (3.11) can be written as

$$L(m) = -\log p(m), \tag{3.12}$$

where $p(m)$ is the probability of selecting m. If there is a prior information about m as to which m is more likely, we should reflect this in $p(m)$. Otherwise, we assume each m is equally likely, i.e., $p(m)$ is a uniform distribution.

Remark 1: Even though the list of models $\mathcal{M}$ does not include the true model, the MDL method achieves the best result among the available models. See Barron,[2] Barron and Cover[3] for detailed information on the error between the MDL estimate and the true model.

Remark 2: We also would like to note that the MDL principle does not attempt to find the absolutely minimum description of the data. The MDL always requires an available collection of models and simply suggests picking the best model from that collection. In other words, the MDL can be considered as an "oracle" for model selection. This contrasts with the algorithmic complexities such as the Kolmogorov complexity which gives the absolutely minimum description of the data, however, in general, is impossible to obtain.[25]

Before deriving our simultaneous noise suppression and signal compression algorithm in the context of the MDL criterion, let us give a closely related example:

Example 4.5: A curve fitting problem using polynomials. Given N points of data $(x_i, y_i) \in \mathbb{R}^2$, consider the problem of fitting a polynomial through these points. The model class we consider is a set of polynomials of orders $0, 1, \dots, N-1$. In this case, $\boldsymbol{\theta}_m = (a_0, a_1, \dots, a_m)$ represents the $m+1$ coefficients of a polynomial of order m. We also assume that the data is contaminated by the additive WGN with unknown variance σ^2, i.e.,

$$y_i = f(x_i) + e_i,$$

where $f(\cdot)$ is an unknown function to be estimated by the polynomial models, and $e_i \sim \mathcal{N}(0, \sigma^2)$. To invoke the MDL formalism, we pose this question in the information transmission setting. First we prepare an encoder which computes the ML estimate of coefficients of the polynomial, $(\widehat{a}_0, \dots, \widehat{a}_m)$ of the given degree m from the data. (In the additive WGN assumption the ML estimate coincides with the least square estimate.) This encoder transmits these m coefficients as well as the estimation errors. We also prepare a decoder which receives the coefficients of the polynomial and residual errors and reconstruct the data. (We assume that the abscissas $\{x_i\}_{1 \le i \le N}$ and the noise variance σ^2 are known to both the encoder and the decoder.) Then we ask how many bits of information should be transmitted to reconstruct the data. If we used polynomials of degree $N-1$, we could find a polynomial passing through all N points. In this case, we could describe the data extremely well. In fact, there is no error between the observed data and the ones reconstructed by the decoder. However, we do not gain anything in terms of data compression/transmission since we also have to encode the model which requires N coefficients of the polynomial. In some sense, we did not "learn" anything in this case. If we used the polynomial of degree 0, i.e., a constant, then it would be an extremely efficient model, but we would need many bits to describe the deviations from that constant. (Of course, if the underlying data is really a constant, then the deviation would be 0.)

Let us assume there is no prior preference on the order m. Then we can easily see that the total codelength (3.11) in this case becomes

$$MDL(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}_m, m) = \log N + \sum_{j=0}^{m} L^*([\widehat{a}_j]) + \frac{m+1}{2}\log N + \frac{N}{2}\log 2\pi\sigma^2 + \frac{\log \mathrm{e}}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{m}\widehat{a}_j {x_i}^j\right)^2. \tag{3.13}$$

The MDL criterion suggests to pick the "best" polynomial order m^* by minimizing this approximate codelength.

The MDL criterion has been successfully used in various fields such as signal detection,[30] image segmentation,[17] and cluster analysis[29] where the optimal number of signals, regions, and clusters, respectively should be determined. Most of these applications use very rigid models such as a set of sinusoids, piecewise polynomials, etc., which either lack adaptability to transient features or are time-consuming to estimate their parameters.

We would like to note that compared to the set of polynomials or sinusoids, the library of wavelets, wavelet packets, and local trigonometric transforms provide a versatile set of models which efficiently describe real-life signals which are full of transients or edges.

4. A SIMULTANEOUS NOISE SUPPRESSION AND SIGNAL COMPRESSION ALGORITHM

We carry on our development of the algorithm based on the information transmission setting as the polynomial curve fitting problem described in the previous section. We consider again an encoder and a decoder for our problem. Given (k, m) in (2.2), the encoder expands the data $\boldsymbol{d}$ in the basis $\mathcal{B}_m$, then transmits the number of terms k, the specification of the basis m, and k expansion coefficients, the variance of the WGN model σ^2, and finally the estimation errors. The decoder receives this information in bits and tries to reconstruct the data $\boldsymbol{d}$.

In this case, the total codelength to be minimized may be expressed as the sum of the codelengths of (1) two natural numbers (k, m), (2) $(k+1)$ real-valued parameters $(\boldsymbol{\alpha}_m^{(k)}, \sigma^2)$ given (k, m), and (3) the deviations of the observed data $\boldsymbol{d}$ from the (estimated) signal $\boldsymbol{f} = \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}$ given $(k, m, \boldsymbol{\alpha}_m^{(k)}, \sigma^2)$. The approximate total description length (3.11) now becomes

$$MDL(\boldsymbol{d}, \widehat{\boldsymbol{\alpha}}_m^{(k)}, \widehat{\sigma}^2, k, m) = L(k, m) + L(\widehat{\boldsymbol{\alpha}}_m^{(k)}, \widehat{\sigma}^2 \mid k, m) + L(\boldsymbol{d} \mid \widehat{\boldsymbol{\alpha}}_m^{(k)}, \widehat{\sigma}^2, k, m), \tag{4.1}$$

where $\widehat{\boldsymbol{\alpha}}_m^{(k)}$ and $\widehat{\sigma}^2$ are the ML estimates of $\boldsymbol{\alpha}_m^{(k)}$ and σ^2, respectively.

Let us now derive these ML estimates. Since we assumed the noise component is additive WGN, we have the probability of observing the data given all model parameters

$$P(\boldsymbol{d} \mid \boldsymbol{\alpha}_m^{(k)}, \sigma^2, k, m) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{\|\boldsymbol{d} - \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}\|^2}{2\sigma^2}\right), \tag{4.2}$$

where $\| \cdot \|$ is the standard Euclidean norm on $\mathbb{R}^N$. For the ML estimate of σ^2, first consider the log likelihood of (4.2)

$$\ln p(\boldsymbol{d} \,| \boldsymbol{\alpha}_m^{(k)}, \sigma^2, k, m) = -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{\|\boldsymbol{d} - \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}\|^2}{2\sigma^2}. \tag{4.3}$$

Taking the derivative with respect to σ^2 and setting it to zero, we easily obtain

$$\widehat{\sigma}^2 = \frac{1}{N} \|\boldsymbol{d} - \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}\|^2. \tag{4.4}$$

Insert this equation back to (4.3) to get

$$\ln p(\boldsymbol{d} \,| \boldsymbol{\alpha}_m^{(k)}, \widehat{\sigma}^2, k, m) = -\frac{N}{2} \ln\left(\frac{2\pi}{N} \|\boldsymbol{d} - \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}\|^2\right) - \frac{N}{2}. \tag{4.5}$$

Let $\widetilde{\boldsymbol{d}}_m = \boldsymbol{W}_m^T \boldsymbol{d}$ denote a vector of the expansion coefficients of $\boldsymbol{d}$ in the basis $\mathcal{B}_m$. Since this basis is orthonormal, i.e., $\boldsymbol{W}_m$ is orthogonal, and we use the ℓ^2 norm, we have

$$\|\boldsymbol{d} - \boldsymbol{W}_m \boldsymbol{\alpha}_m^{(k)}\|^2 = \|\boldsymbol{W}_m (\boldsymbol{W}_m^T \boldsymbol{d} - \boldsymbol{\alpha}_m^{(k)})\|^2 = \|\widetilde{\boldsymbol{d}}_m - \boldsymbol{\alpha}_m^{(k)}\|^2. \tag{4.6}$$

Putting this back to (4.5) and then we find that maximizing (4.5) is equivalent to minimizing

$$\|\widetilde{\boldsymbol{d}}_m - \boldsymbol{\alpha}_m^{(k)}\|^2. \tag{4.7}$$

Considering that the vector $\boldsymbol{\alpha}_m^{(k)}$ only contains k nonzero elements, we can easily conclude that the minimum of (4.7) is achieved by taking the largest k coefficients in magnitudes of $\widetilde{\boldsymbol{d}}_m$ as the ML estimate of $\boldsymbol{\alpha}_m^{(k)}$, i.e.,

$$\widehat{\boldsymbol{\alpha}}_m^{(k)} = \boldsymbol{\Theta}^{(k)} \widetilde{\boldsymbol{d}}_m = \boldsymbol{\Theta}^{(k)} (\boldsymbol{W}_m^T \boldsymbol{d}), \tag{4.8}$$

where $\boldsymbol{\Theta}^{(k)}$ is a thresholding operation which keeps the k largest elements in absolute value intact and sets all other elements to zero. Finally, inserting (4.8) into (4.4), we obtain

$$\widehat{\sigma}^2 = \frac{1}{N} \|\boldsymbol{W}_m^T \boldsymbol{d} - \boldsymbol{\Theta}^{(k)} \boldsymbol{W}_m^T \boldsymbol{d}\|^2 = \frac{1}{N} \|(\boldsymbol{I} - \boldsymbol{\Theta}^{(k)}) \boldsymbol{W}_m^T \boldsymbol{d}\|^2, \tag{4.9}$$

where $\boldsymbol{I}$ represents the N dimensional identity operator (matrix).

Let us further analyze (4.1) term by term. Let us assume now that we do not have any prior information on (k, m) so that the cost $L(k, m)$ is the same for all cases, i.e., we can drop the first term of (4.1) for minimization purpose. If one has some prior preference about the choice of basis, knowing some prior information about the signal $\boldsymbol{f}$, $L(k, m)$ should reflect this information. For instance, if we happen to know that the original function $\boldsymbol{f}$ consists of linear combination of dyadic blocks, then we clearly should use the Haar basis. In this case, we may use the Dirac distribution, i.e., $p(m) = \delta_{m,m_0}$, where m_0 is the index for the Haar basis in the library $\mathcal{L}$. By (3.12), this leads to

$$L(k,m) = \begin{cases} L(k) & \text{if } m = m_0, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.10}$$

On the other hand, if we either happen to know *a priori* or want to force the number of terms retained (k) to satisfy $k_1 \leq k \leq k_2$, then we may want to assume the uniform distribution for this range of k, i.e.,

$$L(k,m) = \begin{cases} L(m) + \log(k_2 - k_1 + 1) & \text{if } k_1 \leq k \leq k_2, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.11}$$

As for the second term of (4.1), which is critical for our algorithm, we have to encode k expansion coefficients $\widehat{\boldsymbol{\alpha}}_m^{(k)}$ and $\widehat{\sigma}^2$, i.e., $(k+1)$ real-valued parameters. However, in this case, by normalizing the whole sequence by $\|\boldsymbol{d}\|$, we can safely assume that the magnitude of each coefficient in $\widehat{\boldsymbol{\alpha}}^{(k)}$ is strictly less than one; in other words, the integer part of each coefficient is simply zero. Hence we do not need to encode the integer part as in (3.11) if we transmit the real-valued parameter $\|\boldsymbol{d}\|$. Now the description length of these becomes approximately $\frac{k+2}{2} \log N + L^*([\widehat{\sigma}^2]) + L^*([\|\boldsymbol{d}\|])$ bits since there are $k+2$ real-valued parameters: k nonzero coefficients, $\widehat{\sigma}^2$, and $\|\boldsymbol{d}\|$. After normalizing by $\|\boldsymbol{d}\|$, we clearly have $\widehat{\sigma}^2 < 1$, so that $L^*([\widehat{\sigma}^2]) = 1/2$ (see (3.5)). For each expansion coefficient, however, we still need to specify the index of the coefficient, i.e., where the k non-zero elements are in the vector $\widehat{\boldsymbol{\alpha}}_m^{(k)}$. This requires $k \log N$ bits. As a result, we have

$$L(\widehat{\boldsymbol{\alpha}}_m^{(k)}, \widehat{\sigma}^2 \mid k, m) = \frac{3}{2} k \log N + c, \tag{4.12}$$

where c is the constant independent of (k,m).

Since the probability of observing $\boldsymbol{d}$ given all model parameters is given by (4.2), we have for the last term in (4.1)

$$L(\boldsymbol{d} \mid \widehat{\boldsymbol{\alpha}}_m^{(k)}, \widehat{\sigma}^2, k, m) = \frac{N}{2} \log \|(\boldsymbol{I} - \boldsymbol{\Theta}^{(k)}) \boldsymbol{W}_m^T \boldsymbol{d}\|^2 + c', \tag{4.13}$$

where c' is the constant independent of (k,m). Using (4.12) and (4.13), now we can state our simultaneous noise suppression and signal compression algorithm:
Pick

$$(k^*, m^*) = \operatorname*{argmin}_{\substack{0 \leq k < N \\ 1 \leq m \leq M}} \left(\frac{3}{2} k \log N + \frac{N}{2} \log \|(\boldsymbol{I} - \boldsymbol{\Theta}^{(k)}) \boldsymbol{W}_m^T \boldsymbol{d}\|^2 \right). \tag{4.14}$$

Then reconstruct the signal estimate

$$\hat{\boldsymbol{f}} = \boldsymbol{W}_{m^*} \boldsymbol{\alpha}_{m^*}^{(k^*)}. \tag{4.15}$$

Let us call the objective function to be minimized in (4.14), the approximate MDL (AMDL). Let us now show a typical behavior of the AMDL value as a function of the number of terms retained (k) in Figure 1. (In fact, this curve is generated using Example 7.1 below.) We see that the $\log(\text{residual energy})$ always decreases as k increases. By adding the penalty term of retaining the expansion coefficients, i.e., $(3/2)k \log N$ (which is just a straight line), we have the AMDL curve which typically decreases for the small k, then starts increasing because of the penalty term, then finally decreases again at some large k near from $k = N$ because the residual error becomes very small. Now what we really want is the value of k achieving the minimum in the beginning of the k-axis, and we want to avoid searching for k beyond the maximum occurring for k near N. So, we can safely assume that $k_1 = 0$ and $k_2 = N/2$ in (4.10) to avoid searching more than necessary. (In fact, setting $k_2 > N/2$ does not make much sense in terms of data compression purpose either.)

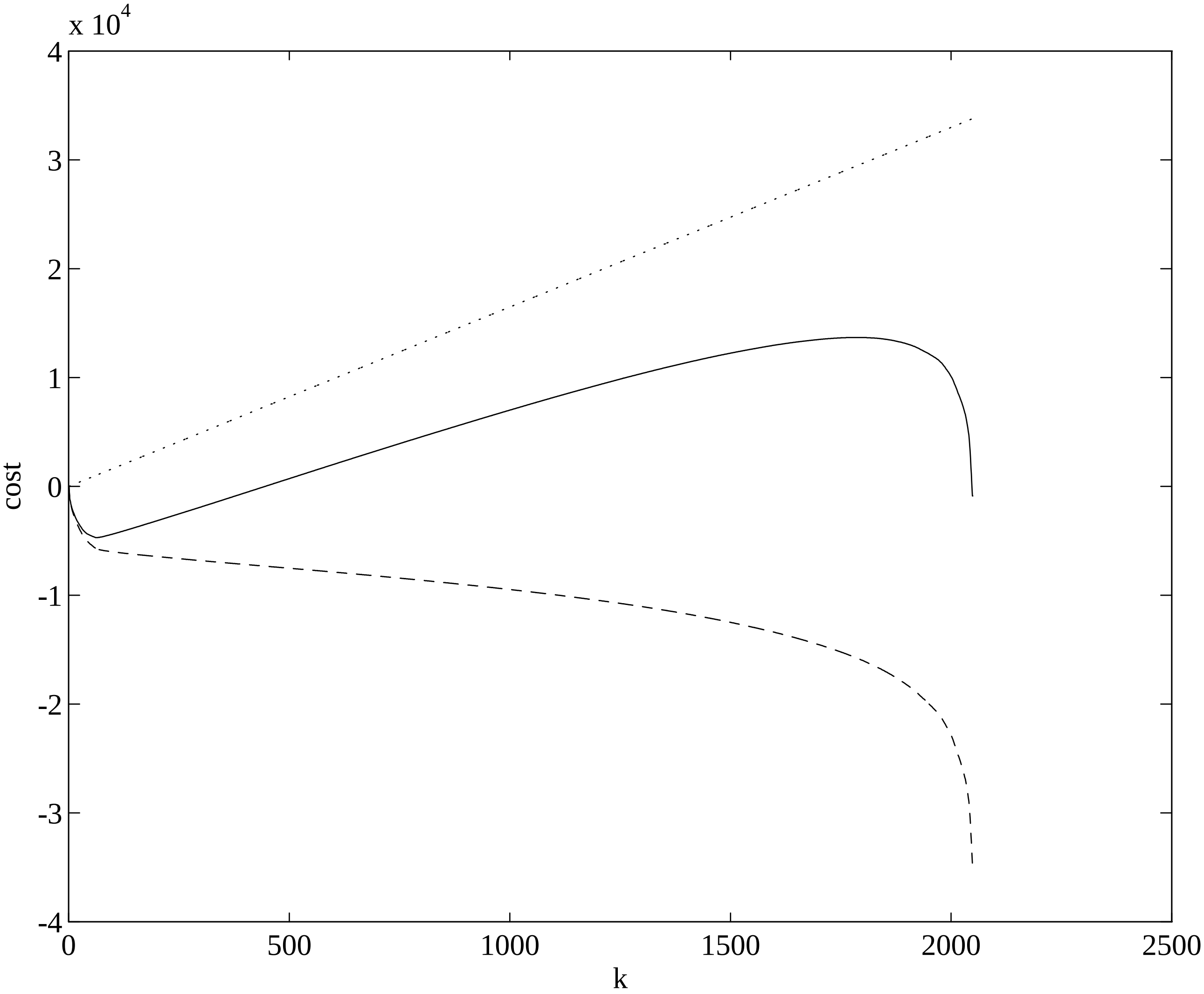

Figure 1. Graphs of AMDL versus k: AMDL [solid line] which is sum of the $(3/2)k$ log N term [dotted line] and the $(N/2)$ log(residual energy) term [dashed line].

We briefly examine the computational complexity of our algorithm here. To obtain (k^*, m^*), we proceed as follows:

Step 1: Expand the data $\boldsymbol{d}$ into bases $\mathcal{B}_1, \ldots, \mathcal{B}_M$. Each expansion (including the best-basis selection procedure) costs $O(N)$ for wavelets, $O(N \log N)$ for wavelet packets best-bases, and $O(N[\log N]^2)$ for local trigonometric best-bases.

Step2: Let $K (= k_2 - k_1 + 1)$ denote the length of the search range for k. For $k_1 \le k \le k_2$, $1 \le m \le M$, compute the expression in the parenthesis of the RHS in (4.14). This costs approximately $O(N + 3MK)$ multiplications and MK calls to the $\log$ function.

Step 3: Search the minimum entry in this table, which costs MK comparisons.

Step 4: Reconstruct the signal estimate (4.15), which costs $O(N)$ for wavelets, $O(N \log N)$ for wavelet packets best-bases, and $O(N[\log N]^2)$ for local trigonometric best-bases.

5. EXTENSION TO IMAGES

For images or multidimensional signals, we can easily extend our algorithm by using the multidimensional version of the wavelets, wavelet packets, and local trigonometric transforms. In this section, we briefly summarize the two-dimensional (2D) versions of these transforms. For the 2D wavelets, there are several different approaches. The first one is the tensor product of the one-dimensional (1D) wavelets, i.e., applying the wavelet expansion algorithm separately along two axes t_1 and t_2 corresponding to row (horizontal) and column (vertical) directions respectively. Let $\boldsymbol{f} \in \mathbb{R}^{N_1 \times N_2}$ and H_i, G_i be the 1D convolution-subsampling operations along axis $t_i, i = 1, 2$. Then this version of the 2D wavelet transform applies the convolution-subsampling operations along the t_1 axis to obtain $\boldsymbol{f}_1 = (G_1 \boldsymbol{f}, G_1 H_1 \boldsymbol{f}, \ldots, G_1 H_1^{J_1} \boldsymbol{f})$, then applies the convolution-subsampling operations along the t_2 axis to get the final 2D wavelet coefficients $(G_2 \boldsymbol{f}_1, G_2 H_2 \boldsymbol{f}_1, \ldots, G_2 H_2^{J_2} \boldsymbol{f}_1)$ of length $N_1 \times N_2$, where J_1 $(\leq \log N_1)$ and J_2 $(\leq \log N_2)$ are maximum levels of decomposition along t_1 and t_2 axes respectively. We note that one can choose different 1D wavelet bases for t_1 and t_2 axes independently. Given M different QMF pairs, there exist M^2 possible 2D wavelets using this approach.

The second one is the basis generated from the tensor product of the multiresolution analysis. This first decomposes an image $\boldsymbol{f}$ into four different sets of coefficients, $H_1 H_2 \boldsymbol{f}$, $G_1 H_2 \boldsymbol{f}$, $H_1 G_2 \boldsymbol{f}$, and $G_1 G_2 \boldsymbol{f}$, corresponding to “low-low”, “high-low”, “low-high”, “high-high” frequency parts of the two variables, respectively. The decomposition is iterated on the “low-low” frequency part and this ends up a “pyramid” structure of coefficients. Transforming the digital images by these wavelets to obtain the 2D wavelet coefficients are described in e.g.,,[18].[12]

There are also 2D wavelet bases which do not have a tensor product structure, such as wavelets on the hexagonal grids and wavelets with matrix dilations. See e.g.,,[16][15] for detail.

There has been some argument as to which version of the 2D wavelet bases should be used for various applications,[4].[12] Our strategy toward this problem is this: we can put as many versions of these bases in the library as we can afford in terms of computational time. Then minimizing the AMDL values automatically selects the most suitable one for our purpose.

Similarly to the 2D wavelets, there are two simple approaches to construct the 2D wavelet packet best-basis: one way is to select the best-basis of each column vector and then select the best-basis of the resulting column best-basis coefficients along row direction. The other way is to recursively decompose not only the “low-low” components but also the other three components. This process produces the “quad-tree” structure of the wavelet packet coefficients instead of the “binary-tree” structure for 1D wavelet packets. Then, finally the 2D wavelet packet best-basis coefficients are selected using the entropy criterion.[31]

The 2D version of the local trigonometric transforms proceeds just like the second way of constructing the 2D wavelet packet best-basis: the original image is smoothly folded and segmented into 4 subimages, 16 subimages, . . . , and in each subimage the separable DCT/DST is applied, and then the quad-tree structure of the coefficients is constructed. Finally, the local trigonometric best-basis is selected using the entropy criterion.[31]

For the image of the $N = N_1 \times N_2$ pixels, the computational costs are approximately $O(N)$, $O(N \log_4 N)$, $O(N[\log_4 N]^2)$ for a 2D wavelet, a 2D wavelet packet best-basis, a 2D local trigonometric best-basis, respectively.

6. EXAMPLES

In this section, we give several examples to show the usefulness of our algorithm.

Example 7.1: The Synthetic Blocky Function of Donoho-Johnstone's. We compared the performance of our algorithm in terms of the visual quality of the estimation and the relative ℓ^2 error with Donoho-Johnstone's method using the blocky function used in their experiments.[14] The results are shown in Figure 2. The true signal is the blocky function with $N = 2048$. In the Donoho-Johnstone method, we used the C06, i.e., 6-tap coiflet with 2 vanishing moments. We also specified the scale parameter $J = 7$, and supplied the *exact* value of σ^2. Next, we *forced* the Haar basis to be used in their method. Finally, we applied our algorithm without specifying anything. In this case, the Haar-Walsh best-basis with $k^* = 63$ was automatically selected. The relative ℓ^2 errors are 0.116, 0.089, 0.051, respectively. Although the visual quality of our result is not too different from Donoho and Johnstone's (if we *choose* the appropriate basis for their method), our method generated the estimate with smaller relative ℓ^2 error and slightly sharper edges.

Example 7.2: A Pure White Gaussian Noise. We generated the synthetic WGN with $\sigma^2 = 1.0$ and $N = 4096$. We also set the upper limit of search range $k_2 = N/2 = 2048$. Figure 3 shows the AMDL curves versus k for all bases in the library. As we can see, there is no single minimum in the graphs, and our algorithm satisfactorily decided $k^* = 0$, i.e., there is nothing to "learn" in this dataset.

Example 7.3: A Natural Radioactivity Profile of Subsurface Formation. We tested our algorithm on the actual field data which is a measurement of natural radioactivity of subsurface formation obtained at an oil-producing well. The length of the data is $N = 1024$. The results are shown in Figure 4. In this case, our algorithm selected the D12 best wavelet packet basis (Daubechies's 12-tap filter with 6 vanishing moments) with $k^* = 77$. The residual error is shown in Figure 4 (c) which consists mostly WGN-like high frequency component. The compression ratio is $1024/77 \approx 13.3$, i.e., about 92.5% of the data were discarded.

Example 7.4: A Migrated Seismic Section. In this example, the data is a migrated seismic section as shown in Figure 5 (a). The data consist of 128 traces of 256 time samples. The two-dimensional version of our program was used. The bases used were D02, C06, C12, C18, C24, C30. (D02 is the wavelet packet best-basis generated by the Haar-Walsh filter, Cn means the wavelet packet best-basis generated by the n-tap coiflet filter.) Figure 5 (b) shows the estimate by our algorithm. It automatically selected the filter C30 and the number of term retained $k^* = 1611$. If we were to choose the good threshold in this example, it would be fairly difficult since we do not know the accurate estimate of σ^2. The compression rate, in this case, is $(128 \times 256)/1611 \approx 20.34$. In other words, 95.1% of the original data were discarded. Figure 5 (c) shows the residual error between the orignal and the estimate. We can clearly see that the random noise and some strange high frequency patterns (which are considered to be numerical artifacts from the migration algorithm applied).

7. DISCUSSIONS

Our algorithm is intimately connected to the "denoising" algorithm described in[9] and.[6] However, the decision mechanism for signal/noise separation is quite different from[9] and[6] where the "theoretical compression rate" was used as the threshold for such a separation.

Our algorithm can also be viewed as a simple yet flexible and efficient realization of the complexity regularization method for estimation of functions proposed by,[2] or the automatic version of the Donoho-Johnstone denoising algorithm via wavelet shrinkage.[14]

The main differences between our algorithm and that of Donoho and Johnstone are

- Our method automatically select the most suitable basis from a collection of bases whereas their method only uses a *fixed* basis specified by the user.
- Our method includes adaptive expansion by means of wavelet packets and local trigonometric bases whereas their method only uses a wavelet transform.
- The Donoho-Johnstone method requires the user to set the coarsest scale parameter $J \leq n$ and a good estimate of σ^2, and the resulting quality depends on these parameters. On the other hand, our method does not require any such parameter setting.
- The Donoho-Johnstone method thresholds the coefficients *softly* whereas our method can be said to threshold *sharply*. This might cause some Gibbs-like effects in the reconstruction using our method.

The future direction of this research is to incorporate noise models other than Gaussian noise, extend the algorithm for highly nonstationary signals by segmenting those smoothly and adaptively, and investigate the effect of sharp thresholding.

8. CONCLUSIONS

We have described an algorithm for simultaneously suppressing the additive WGN component in the data and compressing the signal component in the data. One or more of the bases in the library, consisting of wavelets, wavelet packets, and local trigonometric bases, compress the signal component quite well, whereas the WGN component cannot be compressed by any basis in the library. Based on this observation, we have tried to estimate the “best” basis and the “best” number of terms to retain for estimating the signal component in the data using the MDL criterion. Both synthetic and real field data examples have shown the wide applicability and usefulness of this algorithm.

9. ACKNOWLEDGEMENTS

The author would like to thank Prof. R. Coifman and Prof. A. Barron of Yale University for fruitful discussions.

10. REFERENCES

[1] P. Auscher, G. Weiss, and M. V. Wickerhauser, *Local sine and cosine bases of Coifman and Meyer and the construction of smooth wavelets*, Wavelets: A Tutorial in Theory and Applications (C. K. Chui, ed.), Academic Press, 1992, pp. 237–256.

[2] A. R. Barron, *Complexity regularization with application to artificial neural networks*, Proceeding NATO ASI on Nonparametric Functional Estimation, Kluwer, 1991.

[3] A. R. Barron and T. M. Cover, *Minimum complexity density estimation*, IEEE Trans. Inform. Theory **37** (1991), no. 4, 1034–1054.

[4] G. Beylkin, R. Coifman, and V. Rokhlin, *Fast wavelet transforms and numerical algorithms I*, Comm. Pure Appl. Math. **44** (1991), 141–183.

[5] J. N. Bradely and C. M. Brislawn, *Image compression by vector quantization of multiresolution decompositions*, Physica D **60** (1992), 245–258.

[6] R. R. Coifman and F. Majid, *Adapted waveform analysis and denoising*, Progress in Wavelet Analysis and Applications (Y. Meyer and S. Roques, eds.), Editions Frontieres, B.P.33, 91192 Gif-sur-Yvette Cedex, France, 1993, Proceedings of the International Conference “Wavelets and Applications”, Toulouse, France, Jun., 1992, pp. 63–76.

[7] R. R. Coifman and Y. Meyer, *Remarques sur l’analyse de fourier à fenêtre*, Comptes Rendus Acad. Sci. Paris, Série I **312** (1991), 259–261.

[8] R. R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 713–719.

[9] ______, *Wavelets and adapted waveform analysis*, Wavelets: Mathematics and Applications (J. Benedetto and M. Frazier, eds.), CRC Press, Boca Raton, Florida, 1993.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, 1991.

[11] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math. **41** (1988), 909–996.

[12] ______, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, SIAM, Philadelphia, 1992.

[13] R. A. DeVore, B. Jawerth, and B. J. Lucier, *Image compression through wavelet transform coding*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 719–746.

[14] D. L. Donoho and I. M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, preprint, Dept. of Statistics, Stanford University, Stanford, CA, Jun. 1992, revised Apr. 1993.

[15] K. Gröchenig and W. R. Madych, *Mulitresolution analysis, Haar bases, and self-similar tilings of R^n*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 556–568.

[16] J. Kovačević and M. Vetterli, *Nonseparable multidimensional perfect reconstruction banks and wavelet for R^n*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 533–555.

[17] Y. G. Leclerc, *Constructing simple stable descriptions for image partitioning*, Intern. J. Computer Vision **3** (1989), 73–102.

[18] S. Mallat, *A theory for multiresolution signal decomposition*, IEEE Trans. Pattern Anal. Machine Intell. **11** (1989), no. 7, 674–693.

[19] S. Mallat and Z. Zhang, *Matching pursuit with time-frequency dictionaries*, IEEE Trans. Signal Processing, the special issue on wavelets and signal processing **41** (1993), no. 12, To appear.

[20] Y. Meyer, *Wavelets: Algorithms and Applications*, SIAM, Philadelphia, PA, 1993, Translated and revised by R. D. Ryan.

[21] ______, *Wavelets and Operators*, Cambridge Studies in Advanced Mathematics, vol. 37, Cambridge Univerity Press, New York, 1993, Translated by D. H. Salinger.

[22] W. Niblack, *MDL Methods in Image Analysis and Computer Vision*, New York, Jun. 1993, IEEE Conf. Comput. Vision, Pattern Recognition, Tutorial note.

[23] O. Rioul, *Regular wavelets: a discrete-time approach*, IEEE Trans. Signal Processing, the special issue on wavelets and signal processing **41** (1993), no. 12, To appear.

[24] O. Rioul and M. Vetterli, *Wavelets and signal processing*, IEEE SP Magazine **8** (1991), no. 4, 14–38.

[25] J. Rissanen, *A universal prior for integers and estimation by minimum description length*, Ann. Statist. **11** (1983), no. 2, 416–431.

[26] ______, *Universal coding, information, prediction, and estimation*, IEEE Trans. Inform. Theory **30** (1984), no. 4, 629–636.

[27] ______, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[28] J. M. Shapiro, *Image coding using the embedded zerotree wavelet algorithm*, Proc. SPIE Conf. on Mathematical Imaging: Wavelet Applications in Signal and Image Processing, vol. 2034 (A. F. Laine, ed.), SPIE, 1993.

[29] R. S. Wallace, *Finding natural clusters through entropy minimization*, Ph.D. thesis, School of Comput.

Science, Carnegie Mellon Univ., Pittsburgh, PA 15213, Jun. 1989.

[30] M. Wax and T. Kailath, *Detection of signals by information theoretic criteria*, IEEE Trans. Acoust., Speech, Signal Processing **ASSP–33** (1985), no. 2, 387–392.

[31] M. V. Wickerhauser, *Lectures on wavelet packet algorithms*, preprint, Dept. of Mathematics, Washington University, St. Louis, Missouri, Nov. 1991.

[32] ______, *High-resolution still picture compression*, preprint, Dept. of Mathematics, Washington Univ., St. Louis, Missouri, 1992.

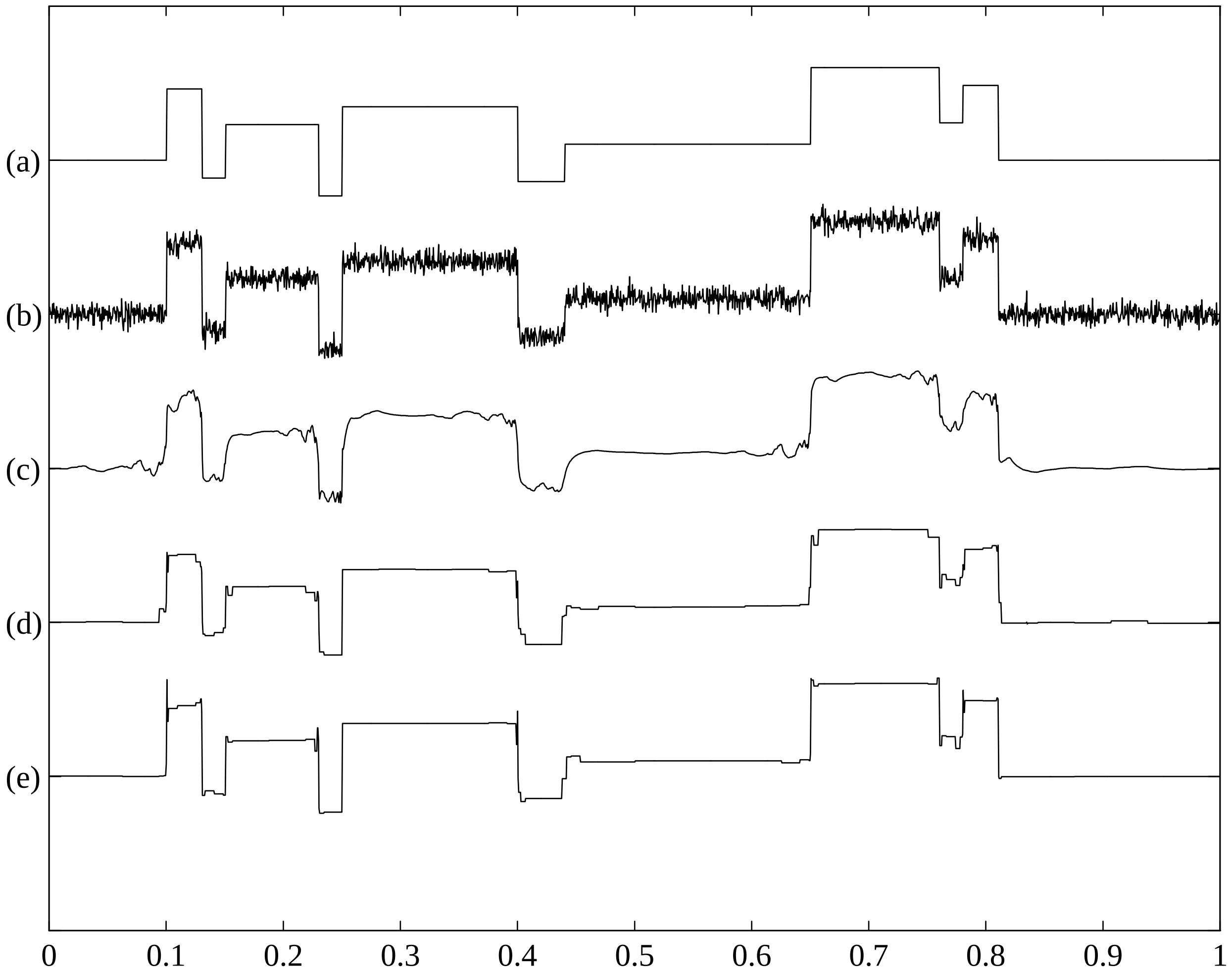

Figure 2. Results for the synthetic blocky function: (a) Original blocky function. (b) Noisy observation with $\|\boldsymbol{f}\|/\|\boldsymbol{n}\| = 7$. (c) Estimation by the Donoho-Johnstone method using coiflets C06. (d) Estimation by the Donoho-Johnstone method using Haar basis. (e) Estimation by the proposed method.

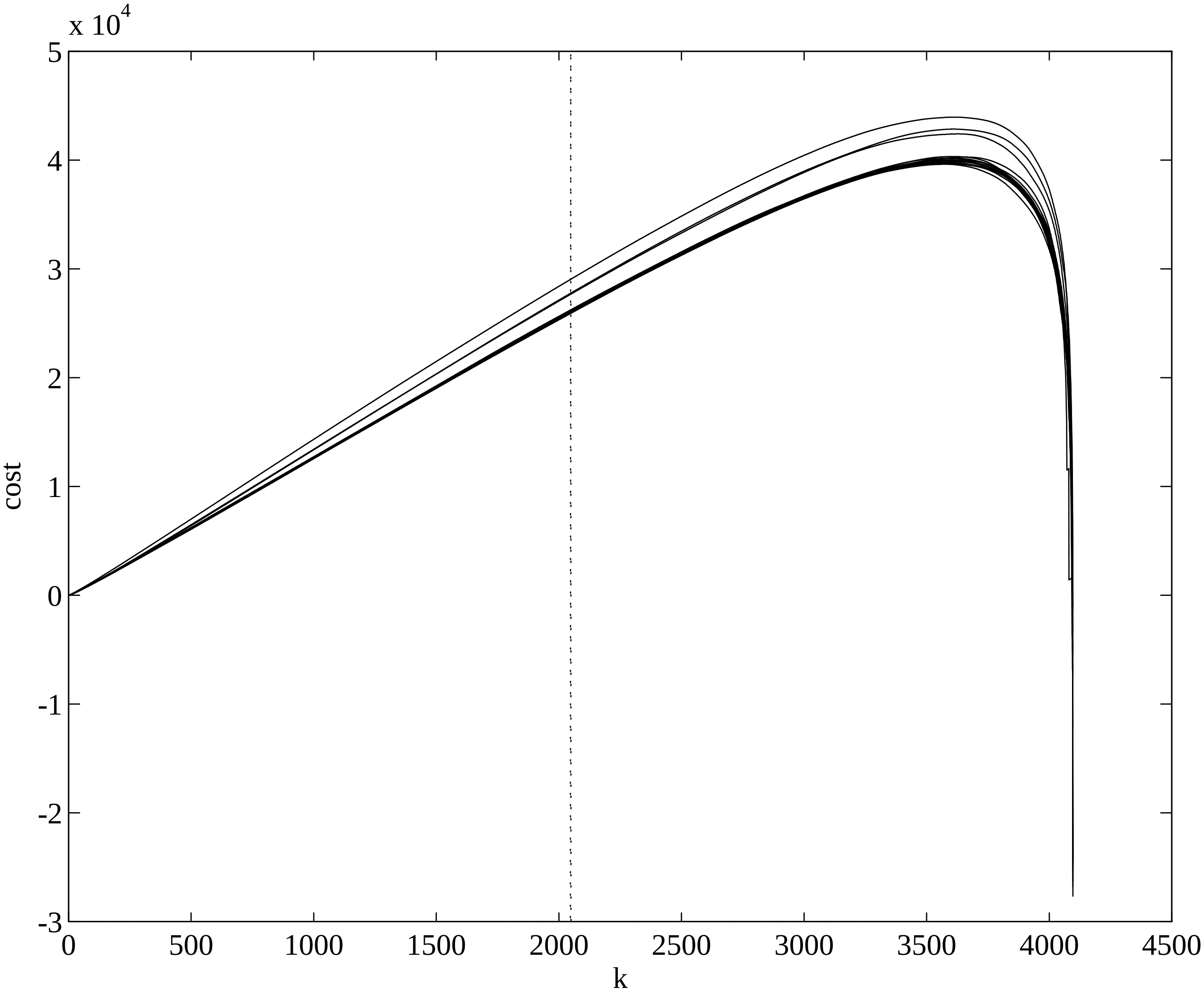

Figure 3. The AMDL curves of the White Gaussian Noise data for all bases. For each basis, $k = 0$ is the minimum value. The vertical dotted line indicates the upper limit of the search range for k.

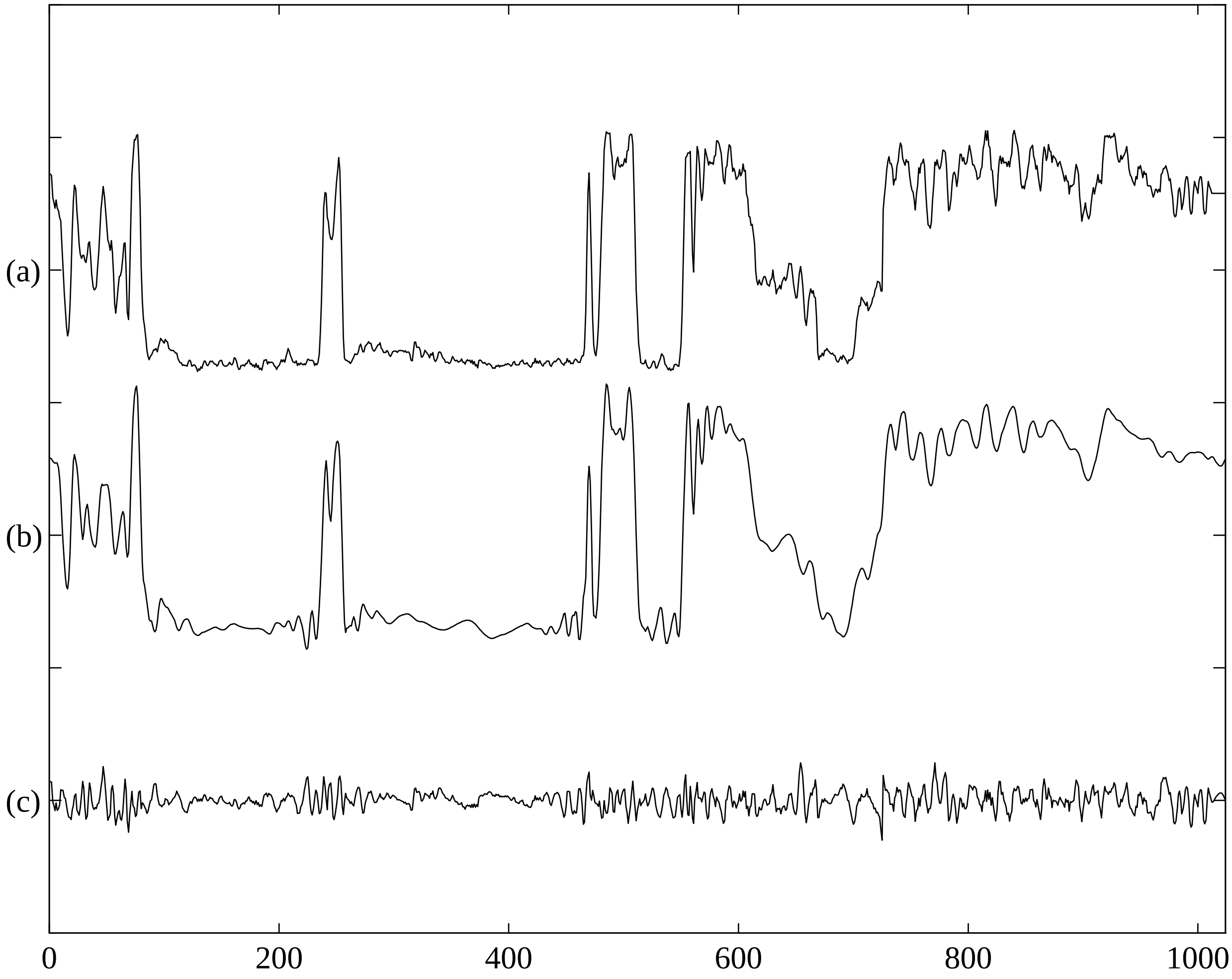

Figure 4. The estimate of the natural radioactivity profile of subsurface formation: (a) Original data which was measured in the borehole of an oil-producing well. (b) Estimation by our algorithm with 77 terms of D12 best wavelet packet basis. (c) Residual error between (a) and (b).

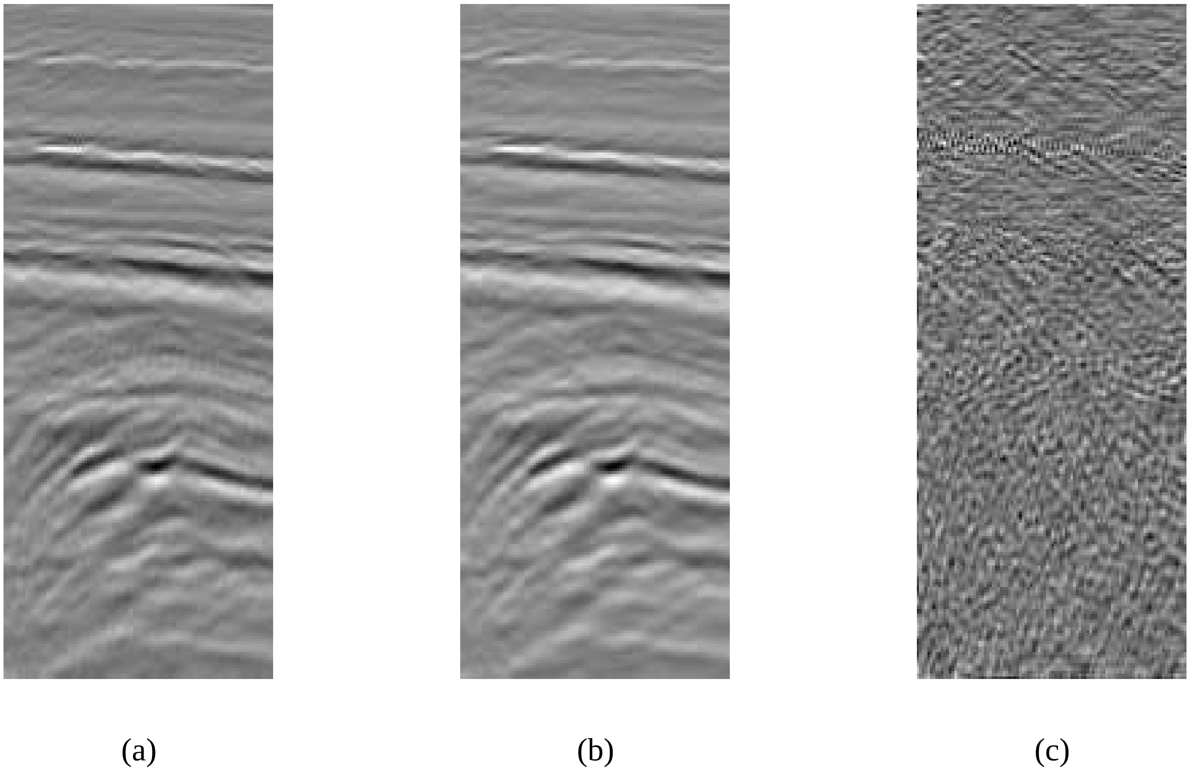

Figure 5. Results for the migrated seismic section: (a) Original seismic section with 128 traces and 256 time samples. (b) Estimation by the proposed method. (c) Residual error between (a) and (b). (Dynamic range of display (c) is different from those of (a) and (b).)