ABSTRACT

MARTELL, LEAH. Data Reduction and Model Selection with Wavelet Transforms. (Under the direction of Dr. Jye-Chyi Lu.)

With modern technology massive quantities of data are being collected continuously. The purpose of our research has been to develop a method for data reduction and model selection applicable to large data sets and replicated data. We propose a novel wavelet shrinkage method by introducing a new model selection criterion. The proposed shrinkage rule has at least two advantages over the current shrinkage methods. First, it is adaptive to the smoothness of the signal regardless of whether it has a sparse wavelet representation, since we consider both the deterministic and the stochastic cases. The wavelet decomposition not only catches the signal components for a pure signal, but de-noises and extracts these signal components for a signal contaminated by external influences. Second, the proposed method allows for fine "tuning" based on the particular data at hand. Our simulation study shows that the methods based on the model selection criterion have better mean square error (MSE) over the methods currently known. Two aspects make wavelet analysis the analytical tool of choice. First, the largest in magnitude wavelet coefficients in the discrete wavelet transform (DWT) of the data, extract the relevant information, while discarding the rest eliminates the noise component. Second, the DWT allows for a fast algorithm calculation of computational complexity $O(n)$.

For the deterministic case we derive a bound on the approximation error of the nonlinear wavelet estimate determined by the largest in magnitude discrete wavelet coefficients. Upper bounds for the approximation error and the rate of increase of the number of wavelet coefficients in the model are obtained for the new wavelet shrinkage estimate. When the signal comes from a stochastic process, a bound for the MSE is found, and for the bias of its estimate. A corrected version of the model selection criterion is introduced and some of its properties are studied.

The new wavelet shrinkage is employed in the case of replicated data. An algorithm for model selection is proposed, based on which a manufacturing process can be automatically supervised for quality and efficiency. We apply it to two real life examples.

Data Reduction and Model Selection with Wavelet Transforms

by

Leah Martell

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Department of Statistics

Raleigh
2000

APPROVED BY:

Dr. B. B. Bhattacharyya

Dr. G. L. Bilbro

Dr. T. Johnson

Dr. Jye-Chyi Lu
Co-chair of Advisory Committee

Dr. Jacqueline Hughes-Oliver
Co-chair of Advisory Committee

Biography

Leah Martell is born on January 29, 1961 in Bulgaria, a country located on the Balkan peninsula in Europe. She moved to the United States and became a naturalized American citizen.

Leah Martell attended high school in her native country Bulgaria. There she earned the degree "A Candidate of the Mathematical Sciences", in the year 1987; this degree is considered equivalent to a Ph.D. in Mathematics here. Her enrollment at North Carolina State University began in the fall of 1995. While attending the university Leah Martell was inducted into two honor societies, Sigma Mu Rho and Sigma Xi. She graduated in 1997 with a Master of statistics from the Department of Statistics at NCSU. After that she continued her education by enrolling in the doctoral program of the same department.

Presently Leah Martell is working as a visiting professor at the Department of Mathematical Sciences at Johns Hopkins University.

Acknowledgments

I would like to express my gratitude to my advisor, Professor Dr. Jye-Chyi Lu, for introducing me to the areas of the statistical signal processing and wavelet transforms. His continuous support and encouragement, prudent guidance and advise, have made the research of this dissertation, an interesting and challenging process. In addition I would like to thank him, particularly, for his unfailing belief in me and his invaluable financial support.

I extend my gratitude to the entire Department of Statistics at North Carolina State University for the high level of educational standards. I also, would like to mention in my list of acknowledgments, the Director of Graduate Programs, Professor Dr. S. Pantula, for giving me the opportunity to enroll as a student at his department and for his efficient running of the program, a result of long working hours.

This research was in part supported by the NSF-VIGRE grant.

The final thanks go to my family in Bulgaria, who has missed me greatly, but have stood stoically by my side all throughout the years of education.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Importance of data reduction

Recent advances in computer technology allowed manufacturers and companies to start building various databases with the purpose of extracting valuable information aimed at improving operation efficiency. Many companies start exploring ways to learn more about their business operations, their customers and suppliers, based on the data they collect daily on aspects like customer orders, material inventory levels, production schedules, service records, etc.. In the last few years a large number of software packages and programs are created to satisfy the need for database building, data mining, and factory operation simulation. They are all meant to facilitate access, editing and manipulation of the existing and accumulating data. In this environment it becomes imperative to create tools for efficient and timely data

analysis which go beyond the traditional statistical methods.

Several characteristicss of the data so collected create the need for a new approach. The sheer size of the data requires methods which use the most recent developments in applied mathematics, signal processing and statistics, allowing for computationally simple and economical ways to synthesize and extract the information locked in the vast volumes of numbers. Particular types of data which, on the one hand have sharp curves, changing trends, numerous cusps, but on the other hand, exhibit distinct patterns, is a good candidate for data reduction with the idea of further decision making. In this regard, wavelets present an excellent tool for data processing. Wavelets can handle data with dynamic trends and irregular data patterns such as sharp jumps better than Fourier transform and standard statistical procedures. Wavelet transforms are by construction intrinsically multi-resolution function approximations which makes them suitable for hierarchical data processing. In addition, like with the Fast Fourier Transform, fast algorithms with computational complexity of order $O(n)$ are developed for efficient and fast computation of the discrete wavelet transform (DWT).

For storage, statistical processing or other types of manipulation, it is desirable to be able to reduce the size of the data set while preserving the valuable information contained in them. For example, due to high cost constraints typically associated with processes in the semiconductor and electronic industries, recent equipment and process problem-solving techniques have employed sophisticated process information synthesis tools to handle

complicated data such as non-stationary and dynamically changing trends due to potential process faults. Methods involving artificial neural networks, wavelet neural networks, spatial time series and others, are applied in a variety of different situations with the hope of revealing a hidden path for process improvement, control and prediction. However, these methods are not suitable for handling large volumes of data. Once again it is necessary that at first, the data are reduced to a manageable size without essential loss of information, and then be processed with the existing method or sophisticated new analytical tools. In other words, the process of synthesizing information involves two parts: first extracting the most salient and important features of the data, or data reduction, and second, analysis for further intelligent decision making. In the present thesis we develop a new method for efficient and computationally effective way to reduce large data sets. The methodology that lies in the heart of our research is wavelet analysis.

1.2 A historical note on wavelets

The development of wavelets is fairly recent in applied mathematics, but wavelets have already had a remarkable impact. The wavelet transforms have been one of the most exciting development in the last decade to bring together researchers in several different fields, such as signal processing, image processing, communications, computer science, to name a few. The beginning of the wavelet transform as a specialized field can be traced to the work

of Grossman and Morlet [1984]. Morlet knew that the modulated pulses sent underground have a duration that is too long at high frequencies to separate the returns of fine, closely spaced layers. Instead emitting pulses of equal duration, he thus thought of sending shorter waveforms at high frequencies. Such waveforms are simply obtained by scaling a single function called wavelet.

The recent unprecedented expansion of the use of wavelet analysis in a wide variety of fields is due, in large part, to its different approach to the time-frequency aspect of signal processing. Like a windowed Fourier transform, a wavelet transform can measure the time-frequency variations of spectral components, but it has a different time-frequency resolution. Due to the fact that the wavelet transform originates in a single function, through scaling and translation, the time resolution increases while the frequency resolution decreases. This allows the wavelet transform to detect and characterize transient phenomena with a zooming procedure across scales. Similarly to the windowed Fourier transform, the wavelet transform detects various frequency bands (scales), but while the first is suitable for time-invariant, stationary processes, the wavelet analysis is particularly adept for dealing with transient phenomena, localized in time.

1.3 Overview of the thesis

In this thesis we combine three lines of research to propose a novel approach to the problem of data reduction and de-noising of a signal, all of which play a vital role in analyzing large volumes of data. The three directions are:

- the recent development of wavelet bases and their application in extracting relevant features from a signal while discarding the irrelevant information i.e. de-noising it;
- the principle of minimum description length (MDL) and its statistical applications;
- the use of information criteria like AIC, C_p, and BIC, proposed, respectively, by Akaike [1970], Akaike [1973], Mallows [1973], Schwarz [1978], as criteria for statistical model selection.

As Rissanen [1984] points out, there are three main aspects to signal processing, namely, prediction, data reduction and estimation. These three aspects, Rissanen asserts, are intrinsically connected by a common link – the information in the string of data observations defined as:

$$\min_{k,\theta}\{-\log_2 P_\theta(x) + \frac{1}{2}k\log_2 N\}$$

where N denotes the number of observations in the data set, $P_\theta(x)$ is a parametric statistical model which we believe explains the data (i.e. the

likelihood function), k is the number of parameters needed to define the statistical model. The focus of this study is the branch concerned with data reduction, the search for an efficient and compact representation of the signal while preserving most of its relevant information.

Saito [1994] observes that all practical methodologies for feature extraction in signal analysis may be classified into two groups: statistical or decision-theoretic approach, and structural or synthetic approach. Examples for the statistical approach are methods like Fast Fourier transform (FFT), Principal Component analysis (PCA) (also known as Karhunen-Lòeve transform, (KLT)), Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), etc. The basic principle for most of these methods is providing an efficient coordinate system tailored for the specific problem; a common drawback is computational complexity and restricted ability to capture features localized, in both, the frequency and the time domains. The structural approach is based on the principle of synthesis, i.e. recovering the signal with the help of pre-defined primitive components used as building blocks. If the set of these primitive components is looked upon as a formal language, this then provides a direct link to results found in information theory. The difficulty in the structural approach lies in choosing a suitable set of building blocks or a basis. A development which appears to be very promising in this context is the introduction of wavelet bases.

In this thesis we first introduce our main analytical tool, that is the wavelet transform for the discrete and continuous case. We also have a

discussion on the notion of non-linear wavelet estimates together with a literature review of the methods currently in use.

The original contribution of this thesis starts from Chapter III. First, we state a general result for the wavelet estimate based on the K largest in absolute value wavelet coefficients, when certain restrictions on the underlying function are imposed.

Guided by the principle that in the process of model selection we are concerned with, on the one hand, precision of the estimate, and with, on the other hand, number of coefficients included in the model, we next propose a novel criterion for selecting the wavelet estimate. Certain properties of thus chosen wavelet estimates are established under the assumption that the underlying function obeys some conditions for regularity. Under the same regularity conditions we find a bound for the rate of increase in the number of coefficients selected for inclusion in the wavelet estimate, as the number of observations increases. We show that in this case the wavelet estimate is equivalent to a wavelet shrinkage found by thresholding the vector of wavelet coefficients, and establish a bound on the corresponding threshold value.

In the second half of Chapter III we consider the case of noisy data. We propose a modified version of the model selection criterion appropriate for the case when the signal is modeled as a stochastic process. Properties concerning the mean value of the mean square error (MSE) and the bias of the estimate of the MSE are established. Another result finds a suitable estimate for the parameter involved in the definition of the model selection

criterion.

The following Chapter IV starts with a comparative study. We present a set of synthesized signals found in most papers on the topic. The wavelet shrinkage estimates considered by various methods, presently in use, are found for this set of synthesized signals, as well as the estimates found by the application of the methods proposed in this work. We compare the performance of the different methods based on their precision and levels of data reduction. Once again we regard separately, the cases of deterministic and stochastic signals. In the second half of Chapter IV we propose an algorithm for fault detection based on the model selection method. Then we apply the algorithm to two sets of replicated data taken from different manufacturing processes.

We conclude in Chapter V with a discussion about our future projects.

Chapter 2

Background and Literature Review

This chapter provides an overview of wavelets, wavelet transforms, wavelet series and their estimates. For these wavelet estimates we summarize statistical properties and model selection methods. We also review the major results concerning our line of research. Although all applications of wavelet analysis concern finite sequences (data sets), the majority of the theoretical results are obtained considering continuous time functions. The connection between continuous and discrete signals is not straightforward; the findings for the continuous time functions reflect the asymptotic properties of discrete sequences sampled at intervals of ever decreasing lengths. A wavelet basis in $L^2(\mathbb{R})$ is constructed by dilating and translating a single wavelet function; dilations are not defined over discrete sequences, as well as, differentiability

or other types of regularity do not apply to discrete signals. Also, continuous time models do not transform directly into discrete signal processing algorithms; for example, uniform sampling a continuous wavelet basis does not produce a discrete basis. For these reasons we will treat discrete and continuous wavelet transforms separately. Our review is brief and focused on the aspects of wavelet analysis which relate to the research of interest. For a better exposition with precise definitions and detailed explanations the reader is referred to books like Daubechies [1988], Chui [1992], Meyer [1990], Wickerhauser [1994], Mallat [1998] among the many treatments of the subject.

2.1 Continuous wavelet transform

2.1.1 Basic wavelet.

Let $L^2(\mathbb{R})$ denote the space of measurable functions f defined on the real line $\mathbb{R}$, that satisfy

$$\|f\|^2 = \int_{-\infty}^{\infty} |f(t)|^2 dt < \infty \, .$$

These are functions which "decay" fast to zero at $\pm\infty$. With the goal of building an orthonormal basis in $L^2(\mathbb{R})$, Meyer [1990] considered functions $\psi(t)$ with zero average: $\int_{-\infty}^{\infty} \psi(t)dt = 0$, that is, its Fourier transform $\hat{\psi}(w)$ vanishes at zero. The graph of such a function looks like a small wave or

wavelet. Usually it is normalized $\|\psi(t)\| = 1$ and centered in a neighborhood of $t = 0$. A commonly used wavelet function is the so called Mexican hat wavelet which is obtained as the second derivative of a Gaussian function. Figure 2.1 displays it along with its Fourier transform.

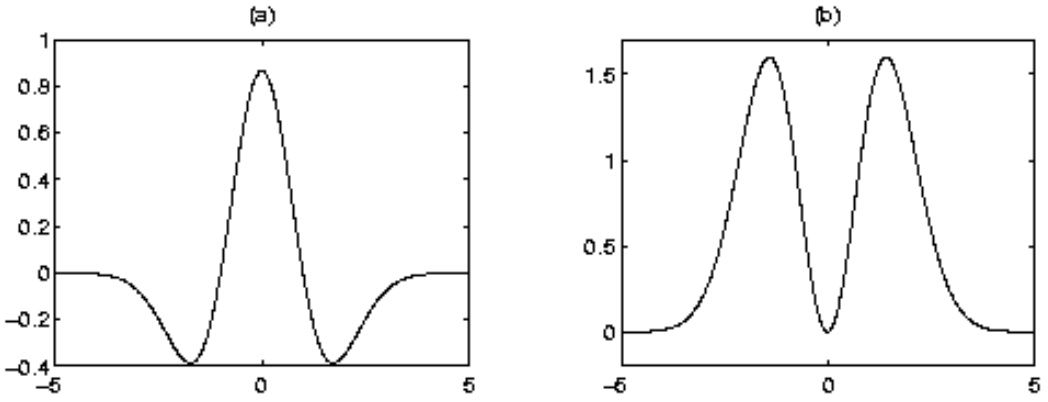

Figure 2.1: Plot of the (a) *Mexican-hat* wavelet $(-\psi(t))$ and (b) its Fourier transform for $\sigma = 1$.

2.1.2 Time-frequency wavelet atoms.

To consider wavelets with different frequency bands, the basic wavelet $\psi(t)$ is dilated by a scale parameter s: $\frac{1}{\sqrt{s}}\psi(\frac{t}{s})$. Now to cover the entire real line the basic wavelets and its scaled versions are shifted by a location parameter u. The resulting function

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)$$

is called a time-frequency atom. The frequency of a signal is directly proportional to the length of the cycle; that is for high-frequency spectral information the time-interval should be relatively small and for low-frequency spectral information, the time-interval should be relatively wide to allow for complete information. The time-frequency atoms have exactly this zoom-in and zoom-out capability.

The inner product in $L^2(\mathbb{R})$ is defined as:

$$\langle f, g \rangle = \int_{-\infty}^{\infty} f(t)\bar{g}(t)dt \,,$$

where $\bar{g}(t)$ is the conjugate of the function $g(t)$.

Definition 1. *The integral or continuous wavelet transform (CWT) is defined as the inner product of the function f with an wavelet atom $\psi_{u,s}(t)$:*

$$Wf(u,s) = \int_{-\infty}^{\infty} f(t)\bar{\psi}_{u,s}(t)dt \,.$$

Since ψ has a zero average, the wavelet transform $Wf(u,s)$ measures the variation of f in a neighborhood of u, whose size is proportional to s.

If the basic wavelet $\psi(t)$ satisfies what is known as the admissibility condition, that is:

$$C_\psi = \int_0^{\infty} \frac{|\hat{\psi}(w)|^2}{w} dw < \infty \,,$$

then any function $f \in L^2(\mathbb{R})$ can be recovered from its CWT by the following

inversion formula:

$$f(t) = \frac{1}{C_\psi} \int_0^\infty \int_{-\infty}^\infty Wf(u,s)\psi_{u,s}(t) du \frac{ds}{s^2} .$$

2.2 Decay of the wavelet coefficients

We are interested in results connecting the regularity of a function with the rate of decay of the wavelet coefficients in its CWT. We will list few of these results but, first, we need to introduce some definitions.

2.2.1 Lipschitz regularity.

Definition 2. *(i) A function is point-wise Lipschitz $\alpha \geq 0$ at ν, if there exist $C > 0$ and a polynomial p_ν of degree $m = \lfloor \alpha \rfloor$ such that*

$$|f(t) - p_\nu(t)| \leq C|t - \nu|^\alpha \quad for \quad t \in \mathbb{R} . \tag{2.1}$$

(ii) A function f is uniformly Lipschitz α over $[a,b]$ if it satisfies (2.1) for all $\nu \in [a,b]$, with a constant C that is independent of ν.

(iii) The Lipschitz regularity of f at ν or over $[a,b]$ is defined as the

$$\sup_\alpha \{ f \text{ is Lipschitz } \alpha \} .$$

If f is uniformly Lipschitz $\alpha > m$ in the neighborhood of ν, then this

implies that f is m times continuously differentiable in this neighborhood. The most interesting case is when $0 \leq \alpha < 1$. Then $p_\nu(t) = f(\nu)$ is a constant and condition (2.1) becomes

$$|f(t) - f(\nu)| \leq C|t - \nu|^\alpha \quad \text{for} \quad t \in \mathbb{R}\,.$$

In this case f is not differentiable at ν and α characterizes the singularity type.

2.2.2 Wavelet vanishing moments.

Definition 3. *A wavelet $\psi(t)$ has n vanishing moments if it is orthogonal to polynomials of degree $n-1$:*

$$\int_{-\infty}^{\infty} t^k \psi(t) dt = 0 \quad \textit{for} \quad 0 \leq k < n\,.$$

Theorem 1. *A wavelet $\psi(t)$ has n vanishing moments if and only if there exists a function $\theta(t)$, whose derivative of order n exists, and such that*

$$\psi(t) = (-1)^n \frac{d^n \theta(t)}{dt^n}\,.$$

Moreover, $\psi(t)$ has no more than n vanishing moments if and only if

$$\int_{-\infty}^{\infty} \theta(t) dt \neq 0\,.$$

This characterization of a wavelet with n vanishing moments leads to a result (Jaffard [1991]) which directly connects the regularity of a function to the decay of the coefficients in its wavelet transform.

Theorem 2. *If $f \in L^2(\mathbb{R})$ is uniformly Lipschitz $\alpha \leq n$ over the interval $[a, b]$, then there exists $A > 0$ such that*

$$|Wf(u,s)| \leq As^{\alpha+\frac{1}{2}} \quad for \quad (u,s) \in [a,b] \times \mathbb{R}^+ \,. \tag{2.2}$$

Conversely, if $Wf(u,s)$ satisfies (2.2) and if $\alpha < n$ is not an integer then f is uniformly Lipschitz α on (a,b).

2.3 Discrete wavelet transform

In the discrete wavelet transform (DWT) the frequency bands are integral powers of 2, and the wavelet atom $\psi_{j,k}(t)$ is obtained from the basic wavelet function $\psi(t)$ by a binary dilation, i.e. dilation by 2^j, and a dyadic translation of $2^{-j}k$, where j, k are integers. In other words

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right) \,.$$

Then the discrete wavelet transform is defined similarly by:

$$w_{j,k} = Wf(2^{(-j)}k, 2^{(-j)}) = \langle f, \psi_{j,k} \rangle \quad \text{where} \quad j,k \in \mathbb{Z} \,.$$

The necessary condition for the existence of a reconstructing formula is the so called stability condition:

$$A \leq \sum_{n=-\infty}^{\infty} |\hat{\psi}(\frac{w}{2^j})|^2 \leq B\,;$$

here A and B are constants, $0 < A \leq B < \infty$, independent of w.

2.3.1 Orthonormal wavelet bases.

It is proven (Mallat [1989], Meyer [1990]) that there exists a wavelet function $\psi(t)$ such that the family of functions

$$\psi_{j,k}(t) = 2^{j/2}\psi\left(2^j t - k\right)\ ,$$

where j and k are arbitrary integers, is an orthonormal basis in the Hilbert space $L^2(\mathbb{R})$. An important contribution by Daubechies [1988] lies in showing that it is possible to find a basic wavelet with a compact support which satisfies the stability condition and its sequence of wavelet atoms forms an orthonormal basis in $L^2(\mathbb{R})$, as well. The support of the basic wavelet, in this case, is proportional to the number of times it is continuously differentiable.

Multiresolutional analysis. Orthonormal wavelets dilated by 2^j carry signal variations at the resolution 2^j. The connection between the orthonormal wavelet basis and the multi-resolution signal approximation is expressed in the notion of a multi-resolution analysis (MRA) in $L^2(\mathbb{R})$.

Definition 4. *An MRA in* $L^2(\mathbb{R})$ *is defined as a sequence of subspaces* V_j, $j \in \mathbb{Z}$ *with the following properties:*

(i) $\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots$;

(ii) $span\,(\cup_{j\in\mathbb{Z}} V_j) = L^2(\mathbb{R})$;

(iii) $\cap_{j\in\mathbb{Z}} V_j = \{0\}$;

(iv) $f(t) \in V_j$ *if and only if* $f(2^{(-j)}t) \in V_0$;

(v) $f(t) \in V_0$ *if and only if* $f(t-k) \in V_0$ *for all* $k \in \mathbb{Z}$;

(vi) there exists a function $\phi(t) \in V_0$, *called a scaling function, such that the system* $\{\phi(t-k)\}_{k\in\mathbb{Z}}$ *forms an orthonormal basis of* V_0.

Using the nested property $V_{j-1} \subset V_j$, let us denote by W_{j-1} the orthogonal complement of V_{j-1} in V_j, that is $V_j = V_{j-1} \oplus W_{j-1}$, where $\oplus$ indicates "direct sum". Now, $L^2(\mathbb{R})$ can be decomposed as follows:

$$\begin{aligned} L^2(\mathbb{R}) &= V_{j_0} \oplus \left(\bigoplus_{j\geq j_0} W_j \right) \\ &= V_0 \oplus \left(\bigoplus_{j\geq 0} W_j \right) \\ &= \oplus_{j\in\mathbb{Z}} W_j \,, \end{aligned}$$

where $\cdots \oplus W_{-k} \oplus W_{-k+1} \oplus \cdots \oplus W_0 \oplus \cdots \oplus W_k \oplus W_{k+1} \oplus \cdots = \oplus_{j\in\mathbb{Z}} W_j$.

This can be expressed by saying that the family of functions

$$\{\phi_{j_0,k}(t), \psi_{j,k}(t)\}_{(j\geq j_0, k\in\mathbb{Z})}$$

is an orthonormal basis in $L^2(\mathbb{R})$.

Respectively, any function $f(t)$ in $L^2(\mathbb{R})$ can have the following three representations:

$$\begin{aligned} f(t) &= \sum_{k\in\mathbb{Z}} s_{j_0,k}\phi_{j_0,k}(t) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t) \\ &= \sum_{k\in\mathbb{Z}} s_{0,k}\phi(t-k) + \sum_{j\geq 0}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t) \\ &= \sum_{j,k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)\ , \end{aligned}$$

where $s_{j,k} = \int_{\mathbb{R}} f(t)\phi_{j,k}(t)dt$ are considered to be the "smooth" coefficients and $d_{j,k} = \int_{\mathbb{R}} f(t)\psi_{j,k}(t)dt$ are viewed as the "detail" ones. In practice, the following finite version of the wavelet series estimate is used:

$$\hat{f}(t) = \sum_{k\in\mathbb{Z}} s_{L,k}\phi_{L,k}(t) + \sum_{j=L}^{J}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(t)\ ,$$

here $J > L$ and L corresponds to the coarsest level (resolution 2^L).

Orthonormal wavelet bases in $L^2[0,1]$. The so called Haar basis is a simple and well known example for an orthonormal basis of $L^2[0,1]$. The development of the wavelet bases in $L^2(\mathbb{R})$, prompted the study of orthonormal bases in $L^2[0,1]$. There are various methods to solve this problem, like the introduction of "boundary" wavelets, or "periodic" wavelets. For our purposes it will suffice to say that such basis exists. We will consider

an orthonormal basis arisen from a multiresolutional analysis of $L^2[0,1]$. Let $\mathcal{I}_j = \{I_{j,k} = [\frac{k}{2^j}, \frac{k+1}{2^j}], 0 \leq k \leq 2^j - 1\}$ are all 2^j dyadic intervals of length $\frac{1}{2^j}$ at a resolution level $j \geq 0$. We consider the family of functions $\left[\{\phi_{L,k}\}_{0\leq k\leq 2^L-1}, \{\psi_{j,k}\}_{j\geq L, 0\leq k\leq 2^j-1}\right]$ where ψ is a basic wavelet and ϕ is the corresponding scale function so that this family of functions forms an orthonormal basis of $L^2[0,1]$; the basic wavelet can be chosen so that it has a compact support and n vanishing moments.

Theorem 2 about the decay of the CWT has an analogous version in the context of a wavelet orthonormal basis in $L^2[0,1]$.

Theorem 3. *If $f \in L^2[0,1]$ is uniformly Lipschitz $\alpha \leq n$ over $[0,1]$, then there exists A such that*

$$|w_{j,k}| = |\langle f, \psi_{j,k}\rangle| \leq \frac{A}{2^{j(\alpha+1/2)}} ,$$

for all resolution levels j and $0 \leq k \leq 2^j - 1$.

2.3.2 Matrix representation of the DWT.

The function $f(t)$ which we are interested in approximating is usually represented by a discrete signal $f_1, f_2, \ldots, f_N$. Another way to describe the wavelet transform in this case, is as a linear transform defined by an orthogonal matrix $\mathbf{W}$, that is $\mathbf{d} = \mathbf{W}\mathbf{f}$, where $\mathbf{d}$ is the vector of N discrete wavelet coefficients. The discrete wavelet coefficients, $d_{j,k}$, are related to the wavelet

coefficients

$$w_{j,k} = \int f(t)\psi_{j,k}(t)\, dt$$

by the relationship $w_{j,k} = d_{j,k}/\sqrt{n} + O(1/n)$. The factor $\sqrt{n}$ arises because of the difference between the continuous and discrete orthonormality conditions. Similarly, the reconstructed signal obtained by the inverse discrete wavelet transform (IDWT), is equivalent to $\mathbf{f} = \mathbf{W^T d}$. In the case when $N = 2^J$ for some positive integer J, both the DWT and IDWT can be performed through an efficient algorithm that requires only $O(n)$ operations, so the DWT is computationally fast.

2.4 Nonparametric regression

In this section we focus our attention on one of the various statistical applications, such as nonparametric regression, density estimation, inverse problems, time series analysis, and change-point problems, which are currently studied in conjunction with the wavelet bases. We review the nonparametric regression in more detail, since this is the main area of interest concerning the research in this thesis.

Here we start with the standard nonparametric regression setting:

$$y_i = f(t_i) + \sigma\, z_i, \quad i = 1, \ldots, N\,,$$

where $\{z_i\}$ are independent, identically distributed (iid), usually, normal random variables with zero mean and variance one. The goal is to recover the underlying function f from the noisy data, $\{y_i\}$, without assuming any practical parametric structure for f. Without loss of generality we assume that $t_i = i/n$, and that the sample size is a power of 2: $N = 2^J$ for some positive integer J. This assumptions will make possible the application of the fast algorithm. A research is being developed to cover the situations when the observations are not equally spaced, or a power of 2, as well as, when the assumption for normality and independence of the random variables modeling the noise are invalid.

2.4.1 Linear estimates.

A linear approximation projects the signal f over a linear space span on a finite subset of vectors taken from a chosen orthonormal basis; if we consider a wavelet orthonormal basis then the expansion

$$f_M = \sum_{(j,k)\in M} w_{j,k}\psi_{j,k}(t)$$

considered over a finite subset M of indices is a linear estimate of f. In this setting the original nonparametric problem essentially transforms to linear regression and the corresponding sample estimates of the scaling coefficients

and the wavelet coefficients are given by:

$$\hat{w}_{j,k} = \frac{1}{N} \sum_{i=1}^{N} \psi_{j,k}(t_i) y_i \ .$$

alternatively the vector $\hat{\mathbf{w}}$ of empirical wavelet coefficients can be written as

$$\hat{\mathbf{w}} = \frac{1}{\sqrt{N}} \mathbf{W} \mathbf{y} \ ,$$

where $\mathbf{W}$ is the DWT matrix. The approximation error is given by the expression: $\epsilon(M) = \|f - f_M\|^2$. The performance of the truncated wavelet estimate clearly depends on the appropriate choice of M. Intuitively it is clear that the "optimal" M should depend on the regularity of the unknown response function. The accuracy of this approximation depends also on the properties of f relative to the particular basis, and it is efficient only if the approximation error decays fast to zero as the subset M of indices increases. It is a well established fact that linear approximations (fast Fourier transform), work well for Sobolev functions. At the same time it is to be expected that the linear estimate will face difficulties in estimating functions which are not homogeneous and contain local singularities. Accurate approximation of these singularities will require high frequency terms and, as a result, a large value of M, while the corresponding oscillating terms will damage the estimate in the smooth regions.

2.4.2 Non-linear wavelet estimates.

In contrast to the linear wavelet estimate introduced above, Donoho and Johnstone [1995], and Donoho et al. [1995] proposed a nonlinear estimate based on selective reconstruction of empirical wavelet coefficients. This approach is now widely accepted in the statistical research, particularly in signal processing and image analysis.

The role of the Sobolev space in the context of linear approximations is taken by the Besov spaces of functions when studying non-linear wavelet estimates.

Definition 5. *The p - modulus of continuity of a function f defined on $\mathbb{R}$, $1 \leq p \leq \infty$, is the function*

$$\omega_p(f;\delta) = \sup_{0<|h|<\delta} \|f(t) - f(t-h)\|_p$$

defined for $\delta > 0$.

The main idea of modulus of continuity is to measure the difference between the function and its translate. If we denote by $MC_p(\mathbb{R})$ the set of functions having a finite p - modulus of continuity (for some δ), then for each $\delta > 0$, $\omega_p(f;\delta)$ is a semi-norm on $MC_p(\mathbb{R})$.

Definition 6. *Suppose $0 < \alpha \leq 1$ and p, s are such that $1 \leq p, s \leq \infty$. The*

Besov norm $\|f\|_{p,\alpha,s}$ *of the function* f *is defined as*

$$\|f\|_{p,\alpha,s} = \begin{cases} \left(\int_0^\infty \left[u^{-\alpha}\omega_p(f;u)\right]^s \frac{du}{u}\right)^{1/s} & \text{when } 1 \le s < \infty \\ \sup_{0<u<\infty} u^\alpha \omega_p(f;u) & \text{when } s = \infty \end{cases}$$

Note that the norms defined above are only semi-norms since they vanish on a constant function.

Definition 7. *A Besov space (inhomogeneous)* $B^p_{\alpha,s}(\mathbb{R})$ *consists of those functions from* $L^p(\mathbb{R})$ *for which* $\|f\|_{p,\alpha,s} < \infty$. *Thus the natural norm on* $B^p_{\alpha,s}(\mathbb{R})$ *is* $\|f\|_p + \|f\|_{p,\alpha,s}$.

The family of Besov spaces contains both Lipschitz spaces and Sobolev spaces. Functions which are only piecewise smooth belong to Besov spaces with high smoothness index α. For functions of the latter type the Fourier based methods have slow convergence rates, while wavelet bases are optimal for compressing and recovering functions in such spaces.

2.4.3 Properties of the non-linear wavelet estimate.

Here we are concerned with the following problem of non-linear approximation:

> Suppose we have a wavelet ψ on $\mathbb{R}$. Given a function f and $\epsilon > 0$, find a "small" set of wavelet coefficients A so that the function f and the sum $\sum_{(j,k)\in A}\langle f, \psi_{j,k}\rangle\psi_{j,k}$ are within ϵ.

The most widely used measure for the distance between the function and its wavelet approximation is the L^2 - norm.

Theorem 4. *(DeVore et al. [1992]) Let α, $0 < \alpha \leq \frac{1}{2}$ be given and let $\tau = \frac{1}{\alpha + \frac{1}{2}}$. Suppose that $f \in L^2(\mathbb{R})$ and that $\|f\|_{\tau,\alpha,\tau} \leq C$. Then there exists a constant C' such that for every integer K we can find $A \subset \mathbb{Z} \times \mathbb{Z}$ with cardinality K such that*

$$\left\| f - \sum_{(j,k)\in A} \langle f, \psi_{j,k} \rangle \psi_{j,k} \right\|_2 \leq C' C K^{-\alpha} \tag{2.3}$$

The set A can be chosen by taking the K coefficients $\langle f, \psi_{j,k} \rangle$ with the biggest absolute values.

This result establishes that when f belongs to a particular family of Besov spaces with a smoothness index α, the best approximation of f with K coefficients is obtained by taking the K largest in absolute values. The rate of approximation for this wavelet estimate then is $o(K^{-\alpha})$.

Before we list few of the numerous minimax results established by Donoho and Johnstone, results which were ground-breaking in the context of non-linear wavelet estimates, we need to introduce a bit more notation.

Let $\mathcal{I} = \cup_{j \geq 0} \mathcal{I}_j$ be the set of all dyadic intervals on the interval $[0, 1]$. We will use the elements of $\mathcal{I}$ as an indexing set without changing the notation; that is if $I \in \mathcal{I}$ it corresponds to some dyadic interval $I_{j,k}$ and we will identify I with the corresponding pair of indexes (j, k). Suppose we observe

N noisy samples of a function f: $y_i = f(\frac{i}{N}) + \sigma z_i, \quad i = 1, \ldots, N$, where $\{z_i\}$ are iid $N(0,1)$. Let $\hat{f}$ be an estimate of f obtained from the observations $y_1, y_2, \ldots, y_N$. Let $R_N(\hat{f}, f) = E\|\hat{f} - f\|^2$ be the risk function. The desired goal is to obtain an estimate $\hat{f}$ attaining the minimax risk

$$\tilde{\mathcal{R}}(N, \mathcal{F}) = \inf_{\hat{f}} \sup_{f} R_N(\hat{f}, f) \,, \tag{2.4}$$

when f belongs to a certain family of smooth functions, like Besov or Triebel function spaces. Let $\{\psi_I, I \in \mathcal{I}\}$ is an orthonormal wavelet basis in $L^2[0,1]$ and $\theta_I = \langle \psi_I, f \rangle$ are the corresponding wavelet coefficients, that is $f = \sum_{I \in \mathcal{I}} \theta_I \psi_I$ (the equality is in terms of the $L^2[0,1]$-norm). Because we are considering an orthogonal basis in the Hilbert space $L^2[0,1]$, we have the following Parseval relation:

$$\|\hat{f} - f\|^2 = \sum_I (\hat{\theta}_I - \theta_I)^2 = \|\hat{\theta} - \theta\|^2 \,.$$

The right hand side is expressed in terms of the norm in $l^2(\mathbb{Z})$, the space of all square-summable, bi-infinite sequences, that is $\{c_n\} \in l^2(\mathbb{Z})$ if and only if $\|c\|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2 < \infty$.

Analogous relation between the Besov norm and, what Donoho and Johnstone [1998] call Besov body $\Theta_{p,q}^{\alpha}$, in the space of wavelet coefficients can be established. This allows for the replacement of the estimation problem (2.4) in the spaces of functions with an estimation problem in the sequence space.

In this we assume we observe sequence data:

$$w_I = \theta_I + \epsilon z_I, \quad I \in \mathcal{I}\,, \tag{2.5}$$

where z_I are iid $N(0,1)$ and $\theta = (\theta_I)_{I\in\mathcal{I}}$. Then the difficulty of estimation in this setting is measured by the following Bayes minimax risk:

$$\mathcal{B}(\epsilon, \Theta^\alpha_{p,q}) = \inf_{\hat\theta} \sup_{\Theta^\alpha_{p,q}} E\|\hat\theta - \theta\|^2\,. \tag{2.6}$$

Now we consider two types of non-linear estimates $\hat\theta_S$ and $\hat\theta_H$, based respectively on the following two types of shrinkage functions:

$$\eta_H(\lambda, x) = xI_{[|x|>\lambda]} \qquad \text{hard shrinkage,}$$
$$\eta_S(\lambda, x) = (x - sgn(x)\lambda)I_{[|x|>\lambda]} \qquad \text{soft shrinkage.}$$

Similarly to the Bayes minimax risk (2.6), we define $\mathcal{B}_S(\epsilon, \Theta^\alpha_{p,q})$ and $\mathcal{B}_H(\epsilon, \Theta^\alpha_{p,q})$ in which the $infimum$ is restricted over $\hat\theta_S$ and $\hat\theta_H$, respectively. Donoho and Johnstone [1998] prove the following theorem establishing the nearly minimax properties of the non-linear estimates obtained through a threshold procedure:

Theorem 5. *Let $\mathcal{F} = \mathcal{F}(C)$ be a ball of Besov space $B^\alpha_{p,q}$ or Triebel space $F^\alpha_{p,q}$ with $\alpha > 1/p$ and $1 \le p, q \le \infty$ or $\alpha = p = q = 1$. There are constants*

$\Lambda(p), K(p)$, *both finite, with:*

$$\mathcal{B}_S(\epsilon, \Theta^{\alpha}_{p,q}) \leq \Lambda(min(p,q))\mathcal{B}(\epsilon, \Theta^{\alpha}_{p,q}) \ , \quad \mathcal{B}_H(\epsilon, \Theta^{\alpha}_{p,q}) \leq K(min(p,q))\mathcal{B}(\epsilon, \Theta^{\alpha}_{p,q}) \ .$$

There exist thresholds which attain these performances.

Corollary 1. *A nearly minimax estimate can be constructed for any of the function spaces $\mathcal{F}$ covered by Theorem 5, by appropriate thresholding of the wavelet coefficients of the function and inverting the wavelet transform.*

2.4.4 Convergence of the threshold estimates.

Before ending this section we would like to mention one more result, concerning the non-linear estimate obtained through a threshold procedure.

Theorem 6. *(Tao [1991]) If $f \in L^p(\mathbb{R})$ for some $1 < p < \infty$, then for almost every x one has*

$$\lim_{\lambda \to 0} T^S_\lambda f(x) = \lim_{\lambda \to 0} T^H_\lambda f(x) = f(x) \ ,$$

where $T^S_\lambda f(x), T^H_\lambda f(x)$ are the threshold estimates for soft and hard threshold, correspondingly.

2.5 Literature Review

2.5.1 Best basis selection algorithm.

In the process of feature extraction and data reduction, a major step is the selection of the appropriate basis. Coifman and Wickerhauser [1992] came up with the idea of a library of orthonormal bases from which an algorithm based on minimizing an information cost functional selects the one most suitable for a particular signal. Specifically, they consider two types of libraries of modulated wave forms, one comprised of local trigonometric bases, and the other containing the wavelet bases, Walsh functions, and the smooth versions of Walsh functions, called wavelet packets. In both libraries, a partial ordering can be defined by moving from finer to coarser partitions of $L^2(\mathbb{R})$, the graph of the partial order can be made into a tree, and the tree can then be searched efficiently for a "best basis".

An information cost functional is defined as a real-valued functional M which maps the sequences $\{x_i\}$ to $\mathbb{R}$ and satisfies the following additivity condition:

$$M(\{x_i\}) = \sum_i \mu(x_i) \quad \text{where} \quad \mu(0) = 0\,.$$

Here $\mu(x)$ is a real valued function defined on $[0, \infty)$. Few examples of information cost functional are as follows:

- Number above a threshold:

$$\mu(x) = \begin{cases} |x| & \text{if } |x| \geq c \\ 0 & \text{otherwise} \end{cases}$$

- Concentration in l^p: For p in $(0, 2)$ set $\mu(x) = |x|^p$, so that the information cost functional $M(\{x_i\}) = \|\{x_i\}\|_p^p$.

- Perhaps, the most noted of the information cost functionals is the Shannon entropy. Let $p_n = \frac{x_n^2}{\|\{x_i\}\|_2^2}$, then:

$$M(\{x_i\}) = -\sum_n p_n \log p_n \, .$$

- Logarithm of energy: Here we take $\mu(x) = \log(x^2)$ and then $M(\{x_i\}) = \sum_k \log(x_i^2)$. This is the function used in the KLT.

2.5.2 Minimum description length (MDL) principle.

This principle is introduced by Rissanen [1983] as a generalization of the maximum likelihood principle in the context of parameter estimation. MDL principle originates in the field of information theory from the desire to minimize the number of bits (binary digits), necessary for the encoding of a data set. By attempting to optimize the description length within a class of parametric distributions, Rissanen [1983] derives the following formula as a

measure of a description length:

$$L(x,\theta) = -\log P(x|\theta) + \log^*[C(k)(\|\theta\|_{M(\theta)})^k] \,.$$

Here, $P(x|\theta)$ denotes the likelihood function of the data x for the parameter vector θ with k components, $C(k)$ is the volume of the k-dimensional unit ball, and $\|\theta\|_{M(\theta)} = \sqrt{\theta^T M(\theta)\theta}$ denotes the natural norm induced by the quadratic form associated with the $k \times k$ matrix $M(\theta)$ of the second derivatives of $-\log P(x|\theta)$. The function $\log^*$ is defined as $\log^* y = \log y + \log\log y + \cdots$, where only the positive terms are included in the sum.

It should be noted that the MDL principle does not attempt to find the absolutely minimum description of the data, rather it picks, from a collection of available models, the one that provides that minimum description length. Saito [1994] derives a similar formula measuring the description length of a particular model in the context of best basis selection. We will provide more details later in the paper.

Another development is the minimax description length (MMDL) criterion proposed in Verdù and Poor [1984] and used in Krim and Schick [1999]. The initial assumption is that the noise distribution f is a scaled version of a distribution belonging to the family of ϵ-contaminated normal distributions $\mathcal{P}_\epsilon = \{(1-\epsilon)\Phi + \epsilon G : G \in \mathcal{F}\}$, where Φ is the standard normal distribution, $\mathcal{F}$ is the set of all distribution functions, and $\epsilon \in (0,1)$ is the known fraction of contamination. In accordance with the minimax principle, the

goal is to find the least favorable noise distribution and evaluate the MDL criterion for that distribution, in other words, this is equivalent to solving a minimax problem where the entropy is simultaneously maximized over all distributions in $\mathcal{P}_\epsilon$ and minimized over all estimators.

2.5.3 Model selection based on an information criterion.

The introduction of the information criterion, Akaike [1970], provided an alternative to the multiple hypotheses testing used previously for model selection. Here, a model is selected according to some optimal properties of an information criterion. Nishii [1984] provides an exhaustive list of the information criteria available presently, among them Akaike information criterion (AIC), Mallows [1973] C_p statistic, and the Bayesian information criterion found in Schwarz [1978]. They have been applied in the setting of a regression model with the assumption for independent, normally distributed errors of constant variance. Define $j = \{j_1, \ldots, j_k\}, 1 \leq j_1 < \cdots < j_k \leq K$ to be the model j if $\beta_{j_1}, \ldots, \beta_{j_k}$ are different from zero but the rest of the elements in the vector of coefficients β are zeros; let J be a set of models j under consideration, for example, J could be the hierarchic model $J_H = \{\bar{j}_1, \ldots, \bar{j}_K\}$ where $\bar{j}_t = \{1, \ldots, t\}$ for $t = 1, \ldots, K$. Below are given the information

criteria which are used often in practice:

$$\begin{aligned} AIC(j) &= N \log \hat{\sigma}^2(j) + ak(j) \\ C_p(j) &= N \frac{\hat{\sigma}^2(j)}{\hat{\sigma}^2(\bar{j_K})} + a(k(j) - 1) \\ BIC(j) &= N \log \hat{\sigma}^2(j) + a_N k(j) \,, \end{aligned}$$

where a is a positive constant, often taken to be 2.0, $a_N > 0$ is a sequence such that $\lim_{N\to\infty} a_N = \infty$; N is the number of data observations, $\hat{\sigma}$ is an estimate of the error variance σ, which is assumed unknown, and $k(j)$ is the number of parameters in the model.

Nishii [1984] derives some results about the asymptotic equivalence of these information criterion functions, based on the behavior of the risk function $R_n(j) = E_y[\|X\beta - X\hat{\beta}\|^2]$.

2.6 De-noising a signal

The idea of "noise-reduction" or "de-noising", as defined in Donoho [1995], is a thresholding procedure through which selected wavelet coefficients are annulled and the reconstructed signal via the IDWT is taken to be the estimate of the original signal. In this section we will list some of the thresholding procedures proposed in the literature presently, together with few different methods for finding an wavelet estimate of the underlying function that describes a given data set.

Donoho and Johnstone [1994] formulated the principle of selective wavelet reconstruction, based on only a subset of the empirical wavelet coefficients. Empirical wavelet coefficients are the ones calculated from the data, while the true wavelet coefficients, which we want to estimate, come from the DWT of the "true" underlying function. Two properties of the wavelet reconstruction support that principle:

Property 1. for a spatially inhomogeneous function the largest wavelet coefficients are concentrated in a small subset of the $\{(j,k)\}$-space;

Property 2. noise affects all wavelet coefficients equally.

In other words while every empirical coefficient contributes noise of variance, only a few contribute signal. Thus their algorithm for finding an estimate includes the following three steps:

Step 1. based on the data $\{y_1, y_2, \ldots, y_n\}$ find the empirical wavelet coefficients: $\mathbf{w} = \mathbf{W}\mathbf{y}$;

Step 2. establish a threshold value λ and a threshold procedure $\delta(.)$ which selects a subset of coefficients from $\mathbf{w}$, while annulling the rest to produce a modified vector of wavelet coefficients $\mathbf{w}^*$;

Step 3. using $\mathbf{w}^*$ calculate the estimate $\hat{\mathbf{f}}^* = \mathbf{W}^\mathbf{T}\mathbf{w}^*$.

2.6.1 *VisuShrink.*

We will start the description of the known choices for threshold values with the "universal" or *VisuShrink* procedure, proposed in Donoho and Johnstone [1994]. The universal threshold is taken to be $\lambda = \hat{\sigma}\sqrt{2\log n}$, where $\hat{\sigma}$ is an estimate of the error variance, which is usually unknown. In practice, the coarsest few resolution levels are left intact and the thresholding is applied only to the remaining empirical wavelet coefficients.

Theorem 7. *Assume model (2.5). The estimator*

$$\hat{\theta}_i^u = \eta_S(w_i, \epsilon(2\log n)^{\frac{1}{2}}) \quad (i = 1, \ldots, n)$$

satisfies

$$E\|\hat{\theta}^u - \theta\|_{2,n}^2 \leq (2\log n + 1)\left\{\epsilon^2 + \sum_{i=1}^{n} \min(\theta_i^2, \epsilon^2)\right\} \tag{2.7}$$

for all $\theta \in \mathbb{R}^n$.

2.6.2 *RiskShrink.*

When the factor $(2\log n + 1)$ in (2.7) is replaced by a constant Λ_n^* calculated as a minimax quantity:

$$\Lambda_n^* \equiv \inf_{\lambda} \sup_{\mu} \frac{\rho(\lambda, \mu)}{\frac{1}{n} + \min(\mu^2, 1)} \tag{2.8}$$

then another threshold value λ^* can be obtained as

$$\lambda_n^* = \text{the largest } \lambda \text{ attaining } \Lambda_n^* \text{ above.} \tag{2.9}$$

This last shrinkage estimation is known as the *RiskShrink*. The function $\rho(\lambda, \mu)$ is defined as $\rho(\lambda, \mu) = E[\delta_\lambda^S(Y) - \mu]^2$ for Y, a random variable from a $N(\mu, 1)$ distribution.

Theorem 8. *Assume model (2.5). The minimax threshold λ_n^* defined at (2.9) yields an estimate*

$$\hat{\theta}_i^* = \eta_S(w_i, \lambda_n^* \epsilon) \quad (i = 1, \ldots, n)$$

which satisfies

$$E\|\hat{\theta}^* - \theta\|_{2,n}^2 \leq \Lambda_n^* \left\{ \epsilon^2 + \sum_{i=1}^{n} \min(\theta_i^2, \epsilon^2) \right\} \tag{2.10}$$

for all $\theta \in \mathbb{R}^n$. The coefficient Λ_n^* defined at (2.8), satisfies $\Lambda_n^* \leq 2 \log n + 1$, and the threshold $\lambda_n^* \leq (2 \log n)^{\frac{1}{2}}$. Asymptotically, as $n \to \infty$,

$$\Lambda_n^* \sim 2 \log n, \quad \lambda_n^* \sim (2 \log n)^{\frac{1}{2}} .$$

If we go back to the sequence data in the function space, the performance of the wavelet estimator is measured by the risk $R(\hat{f}, f) = \frac{1}{n} E\|\hat{f} - f\|^2$; the last norm is the usual l_N^2 norm, that is, if $v = \{v_1, \ldots, v_N\} \in l_N^2$ then

$\|v\| = \sum_{i=1}^{N} v_i^2$. Let

$$\mathcal{R}_{n,\sigma}(T_{SW}, f) = \inf_{\delta} R(T_{SW}(y, \delta), f)$$

denote the ideal risk; here $T_{SW}(y, \delta)$ stands for the estimator $\hat{f}$ obtained through a thresholding procedure.

Corollary 2. *For all f and all $n = 2^{J+1}$,*

$$R(\hat{f}, f) \leq \Lambda_n^* \left\{ \frac{\sigma^2}{n} + \mathcal{R}_{n,\sigma}(T_{SW}, f) \right\} .$$

Moreover, no estimate can satisfy a better inequality than this for all f and n, in the sense that for no measurable estimate can such an inequality hold with Λ_n^* replaced by $\{2 - \epsilon + o(1)\} \log n$.

2.6.3 *SureShrink.*

For signals which are assumed "spatially homogeneous" and smooth, Donoho and Johnstone [1995] proposed a different shrinkage procedure which has different threshold values for different resolution levels in the wavelet transform. This approach uses Stein's unbiased estimate of the risk function (SURE) and the threshold procedure is called *SureShrink*, respectively. The explicit form of SURE is as follows:

$$SURE(x, \lambda) = N - 2 \sum_{i=1}^{N} I_{[|x_i| \leq \lambda]} + \sum_{i=1}^{N} \min(x_i^2, \lambda^2) .$$

Let $d_j = (d_{j,1}, d_{j,2}, \ldots, d_{j,2^j})^T$ be the vector of wavelet coefficients at a resolution level j. Then the *SureShrink* threshold at a resolution level j is given by;

$$\lambda_j^{SURE} = \arg \min_{0 \le \lambda \le \lambda^U} SURE(d_j, \lambda)$$

where $\lambda^U = \sqrt{2\sigma \log N}$ is the universal threshold.

Theorem 9. *Let the discrete wavelet analysis correspond to a wavelet ψ having n vanishing moments and n continuous derivatives, $n > max(1, \sigma)$. Then the estimator SureShrink $\hat{f}^*$ is simultaneously nearly minimax,*

$$\sup_{B^\sigma_{p,q}(C)} R(\hat{f}^*, f) \asymp R(N, B^\sigma_{p,q}(C)) \quad N \to \infty$$

for all $p, q \in [1, \infty]$, for all $C > 0$ and for all $\sigma_0 < \sigma < n$.

2.6.4 Hybrid Algorithm.

The authors of *SureShrink* found that this procedure doesn't perform as well in situations of extreme sparsity of the wavelet coefficients. To remedy this deficiency an algorithm called the Hybrid scheme is suggested instead. It contains the following three steps:

Step 1. for a specific resolution level j calculate a criterion for sparsity:

$$(sp)_j^2 = \frac{1}{2^j} \sum_{k=1}^{2^j} \left(\frac{d_{j,k}^2}{\sigma_j^2} - 1 \right) ;$$

Step 2. choose a critical value e.g. $\gamma_j = (\log_2 2^{\frac{j}{2}})^{3/2}$;

Step 3. now, if $(sp)_j^2 \leq \gamma_j$ use the universal threshold; otherwise use the threshold defined by *SureShrink*.

2.6.5 Cross-validation procedures.

A different type of wavelet estimate which de-noises the data is proposed by Nason [1996] and is based on cross-validation. Two ways of cross-validation are suggested: *two-fold* cross-validation algorithm and *leave-one-out* cross-validation. The initial assumptions are slightly different in that the noise is not assumed to come from a normal distribution but rather comes from a distribution with mean 0 and constant variance σ^2. Let $y_1, y_2, \ldots, y_n$ be the data points, where $n = 2^M$. For the *two-fold* cross-validation, all the odd-indexed y_i are removed from the set. This leaves 2^{M-1} evenly indexed y_i which are re-indexed from $j = 1, 2, \ldots, \frac{n}{2}$. A function estimate $\hat{f}_t^E$ is then constructed from the re-indexed y_i by using a particular threshold t. To compare the function estimate with the left-out noisy data an interpolated version of $\hat{f}_t^E$ is formed:

$$\bar{f}_{t,j}^E = \frac{1}{2}(\hat{f}_{t,j+1}^E + \hat{f}_{t,j}^E), \qquad j = 1, 2, \ldots, \frac{n}{2},$$

setting $\hat{f}_{t,n/2+1}^E = \hat{f}_{t,1}^E$ because f is assumed periodic. Then $\hat{f}_t^O$ is computed for the odd-indexed points and the interpolant $\bar{f}_t^O$ computed as above. The full estimate for the mean integrated square error (MISE)$M(t) = E\int(\hat{f}_t(x) -$

$f(x))^2 dx$ compares the interpolated wavelet estimates and the left-out points:

$$\hat{M}(t) = \sum_{j=1}^{n/2} \left\{ (\bar{f}_{t,j}^{E} - y_{2j+1})^2 + (\bar{f}_{t,j}^{O} - y_{2j})^2 \right\}.$$

Note that the estimate $\hat{M}$ relies on two estimates of f_t based on $\frac{n}{2}$ data points. These estimates are calculated with the use of the universal threshold (*VisuShrink*), modified to accommodate $\frac{n}{2}$ data points, by the following formula:

$$T_{UV}(n) \approx \frac{T_{UV}(\frac{n}{2})}{\sqrt{1 - log_2 n}}.$$

After the estimate $\hat{M}(t)$ has been minimized the correction from above is applied, to obtain the final cross-validated wavelet estimate.

The *leave-one-out* cross-validation algorithm is applicable to a data set of any size n, not necessarily a power of 2. Given the data set $Y = \{y_1, y_2, \ldots, y_n\}$, $n > 1$, choose i between 1 and n. Remove y_i and split the remaining points into two groups:

$$Y_L = \{y_1, \ldots, y_{i-1}\}$$
$$Y_R = \{y_{i=1}, \ldots, y_n\}.$$

Form Y_{LRE} and Y_{RRE} by reflection at the left-hand and right-hand ends of Y_L and Y_R, respectively. Next extend each set to the next largest power of 2

by filling with y_{i-1} for Y_{LRE} and y_{i+1} for Y_{RRE} to obtain:

$$Y_{LRE} = \{y_{i-1}, \ldots, y_{i-1}, y_{i-2}, \ldots, y_2, y_1, y_1, y_2, \ldots, y_{i-2}, y_{i-1}\}$$
$$Y_{RRE} = \{y_{i+1}, y_{i+2}, \ldots, y_{n-1}, y_n, y_n, y_{n-1}, \ldots, y_{i+2}, y_{i+1}, \ldots, y_{i+1}\}.$$

Denote the number of points in Y_{LRE} by n_L and in Y_{RRE} by n_R. Then n_L is the smallest power of 2 greater than or equal to $2(i-1)$ and n_R is the smallest power of 2 greater than or equal to $2(n-i)$. Now form two wavelet shrinkage estimates $\hat{f}_{L,t}$ and $\hat{f}_{R,t}$ by using Y_{LRE} and Y_{RRE} and threshold value t. The removed point y_i is predicted by

$$\hat{y}_{t,-i} = \frac{1}{2}(\hat{f}_{L,t,n_L} + \hat{f}_{R,t,1})$$

where $\hat{f}_{L,t,n_L}$ is the rightmost point of $\hat{f}_{L,t}$, and $\hat{f}_{R,t,1}$ the leftmost point of $\hat{f}_{R,t}$. The cross-validation score is given by

$$\hat{M}(t) = \sum_{i=2}^{n-1}(y_i - \hat{y}_{t,-i})^2 .$$

2.6.6 False discovery rate of coefficients (FDR).

From a statistical point of view thresholding is closely related to another data-analytical approach to model building which utilizes multiple hypotheses testing. When the number of coefficients is large, as it is usually the case in the wavelet transform, the multiple hypothesis testing runs into the prob-

lem of accepting erroneously coefficients in the model if the error is controlled at an individual level. On the other hand if the error is controlled simultaneously for all hypotheses (Bonferroni's approach), then there is a great risk of throwing out coefficients which are statistically significant. A solution to the problem of the multiple hypotheses testing is proposed in Abramovich and Benjamini [1995] and is introduced under the name – false discovery rate of coefficients (FDR).

If R is the number of coefficients that are not dropped by the thresholding procedure for a given sample, then let S be the number of coefficients kept correctly, and V be the number of coefficients which are retained erroneously, where $R = S + V$. The error in such procedure is expressed in term of the random variable $Q = V/R$, that is the proportion of coefficients kept in the model that should have been left out. The FDR then is defined as the expected value of Q and the procedure is to take in the model as many coefficients as possible before FDR exceeds an a priori set level q.

Applying this procedure to the wavelet thresholding takes the following steps:

Step 1. for each coefficient $d_{j,k}$ calculate the corresponding two-sided p-value $p_{j,k}$ testing the hypothesis $H_{j,k} : d_{j,k} = 0$

$$p_{j,k} = 2(1 - \Phi(\frac{|d_{j,k}|}{\sigma}))\,;$$

Step 2. order the $p_{j,k}$'s according to their size, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, where

each of the $p_{(i)}$'s corresponds to some coefficient $d_{j,k}$;

Step 3. starting with $i = 1$, let k be the largest i for which $p_{(i)} \leq (i/m)q$; for this k calculate:

$$\lambda_k = \sigma\Phi^{-1}(1 - \frac{p_{(k)}}{2}) ;$$

Step 4. threshold all coefficients at level λ_k.

2.6.7 Ogden's *selection* thresholding and *data-analytic* thresholding.

Selection thresholding is based on hypothesis testing of coefficients level-by-level. Given a set of coefficients at a particular level Ogden [1994] describes a test statistic that if large will prompt the user to include the largest (in absolute values) coefficients into the reconstruction decomposition and then continue testing the remainder of the coefficients. If the test statistic is not large enough (when compared to some critical value) then the threshold is set to be the absolute value of the largest remaining coefficient.

Data-analytic thresholding is based on looking at plots of cumulative sums of the squares of the coefficients at a particular level. Coefficients are removed from the level (and marked for inclusion in the reconstruction) if some test based on Brownian bridge sampling is significant and then the remaining coefficients are tested. The test tries to ascertain if the remaining coefficients

are just white noise, by successively removing the larger coefficients until the test decides that the coefficients are indistinguishable from white noise. One important advantage of the *data-analytic* thresholding is that it does not separate coefficients that are close in time. For example, discontinuities can often cause two adjacent coefficients to be large (rather than just one) and this method identifies these together, which would not be possible with other methods that separate coefficients and sort according to size.

There are other methods for selecting a threshold, which are derived from a significance test Fan [1996], Lee [1997], but they involve testing one wavelet coefficient at a time and don't seem helpful for our purpose of minimizing the number of wavelet coefficients used for the calculation of the wavelet series estimate.

2.6.8 Bayesian approach.

Various Bayesian approaches for thresholding and nonlinear shrinkage in general, has been proposed recently as well. These have been shown to be relatively effective and it is argued that they are less ad-hoc than the proposals discussed above. In the Bayesian approach a prior distribution is imposed on the wavelet coefficients. The prior model is designed to capture the sparseness of wavelet expansions common to most applications. Then the function is estimated by applying a suitable Bayesian criteria to the resulting posterior distribution of the wavelet coefficients. Different choices of loss functions lead to different Bayesian rules and hence different nonlinear shrinkage. Compre-

hensive reviews on Bayesian wavelet regression are given by Vidakovic [1999], Abramovich and Sapatinas [1999].

2.7 Data reduction

The diversity of methods listed in the previous section is explained, in part, by the various goals that the corresponding series estimates are meant to achieve. For example, *VisuShrink* and *RiskShrink* aim to produce an estimate which is noise-free, and runs the risk of under-fitting the data (see also Fan et al. [1993]); their estimate has a near optimal minimax property in ensuring, with high probability, that the estimate is as smooth as the true underlying function. *SureShrink* and the cross-validation methods, on the other side, minimize the risk function, and this produces estimates with lower bias but more variance. The cross-validation method gives too low a threshold when errors are serially correlated (Lee [1997]). In addition, both types of methods use procedures which are essentially equivalent to multiple hypotheses testing, and harbor the danger of leaving out coefficients which might be relevant.

For the purpose of data reduction we need to find

(a) a series wavelet estimate,

(b) which is produced as a result of a model selection procedure,

(c) based on the optimization of a function analogous to the information

criterion function (as done in the context of linear regression),

(d) and which has the property of a near minimum description length.

In this section we list several methods for data reduction found in the literature, and then we describe a novel procedure reaching for the same goal.

2.7.1 Approximate minimum description length.

This algorithm, found in Saito [1994] takes advantage of two different approaches in data compression: one, the search for a best basis as in Coifman and Wickerhauser [1992] according to a given cost function, and two, the Minimum Description Length (MDL) criterion taken from the field of information theory (Rissanen [1984]). If $\mathcal{L} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_M\}$ is a library of bases, where $\mathcal{B}_m$ represents one of the orthonormal bases in the library, then Saito's algorithm finds m, an indicator for the orthonormal basis chosen from the library, and k, the number of non-zero elements retained for the reconstruction of the estimate, to be the values that minimize the following measure for MDL:

$$AMDL(k, m) = \frac{3}{2} k \log_2 N + \frac{N}{2} \log_2 \|(I - \Theta^{(k)}) W_m^T d\|^2 . \tag{2.11}$$

Here d is the data vector, W_m is the linear transformation that recovers the signal $f = W_m \alpha_m^{(k)}$, and $\Theta^{(k)}$ is the thresholding operation which keeps the k largest elements in absolute value intact and sets all other elements to zero.

The criterion AMDL is used also by Antoniadis et al. [1997] in the context of the wavelet orthonormal bases.

2.7.2 Tree-constrained thresholding.

The idea of adding a penalty term is found in Buckheit and Donoho [1995], as well. It is considered in the context of the tree-constrained thresholding, based on a search for a best tree pattern for the "surviving" wavelet coefficients. The best tree is defined by the optimization of the quantity:

$$CPRSS_\lambda = \|f - \hat{f}_T\|_2^2 + \lambda K_T$$

where K_T is the number of coefficients in the chosen tree, and λ could equal 2 for the standard AIC model selection, BIC uses $\lambda = \log n$, RIC uses $\lambda = 2\log n$. The selection of the best tree imitates the Coifman-Wickerhauser pruning algorithm, Coifman and Wickerhauser [1992].

2.7.3 Relative reconstruction error.

Another criterion is proposed in Lu et al. under the name of relative reconstruction error:

$$RRE(k) = \left(\frac{\sum_{i=1}^{N} (y_i - \hat{y}_{i,k})^2}{\sum_{i=1}^{N} y_i^2} \right)^{\frac{1}{2}} + \frac{k}{N} .$$

Chapter 3

Data Reduction – New Approach

In this chapter we introduce the main results in the thesis – the new model selection criterion, the properties of the nonlinear wavelet estimate obtained on the basis of this model selection criterion, as well as, the modified version pertaining to the case of data with noise. We start with a brief motivation for the need of a novel approach and why we chose this particular direction of research.

3.1 Motivation for the research

Due to the orthogonality of the matrix $\mathbf{W}$, the matrix of the discrete wavelet transform, the DWT of white noise is also an array of iid $N(0,1)$ random

variables, so the model

$$y_i = f(t_i) + \sigma z_i \,, \quad i = 1, \ldots, 2^J \,,$$

where $\{z_i\}$ are iid $N(0,1)$, leads to the model:

$$\hat{w}_{j,k} = w_{j,k} + \epsilon z_{j,k} \,, \quad j = 1, \ldots, J \,, \quad k = 1, \ldots, 2^j \,,$$

where $\epsilon^2 = \sigma^2/N$. The sparseness of the wavelet expansion makes it reasonable to assume that essentially only a few "large" $\hat{w}_{j,k}$ contain information about the underlying function f, while "small" $\hat{w}_{j,k}$ can be attributed to the noise which contaminates uniformly all wavelet coefficients. If we can decide which are the "significant" large coefficients, then we can retain them and set all others equal to zero, so obtaining an approximate wavelet representation of the underlying signal.

The thresholding procedure depends greatly on the appropriate choice of the threshold value. If we focus on the universal threshold, despite its simplicity and minimax properties, it depends only on the estimate of the variance of the data and the number of observations under consideration. This means it ignores the actual values of the observations in the data set and cannot be "tuned" to the specific problem of interest. To improve the finite sample properties of the universal threshold Donoho and Johnstone [1994] suggests that one should always keep the coefficients of the coarsest

level, even if they do not pass the threshold. Still for large samples the universal threshold while successful in de-noising the signal, removes part of the "real" signal as well, that is it does not compromise well between signal and noise, and in practice it tends to over-smooth the signal. The performance of $VisuShrink$ is open for the same criticism, in that it is too aggressive in eliminating coefficients.

The various data-adaptive thresholding rules seem to lead to better results. These include $SureShrink$ and $AMDL$ methods, and this is where the effort of this research is directed, as well. The problem with $SureShrink$ is that it is a bit more complex from a computational point of view, which can be a stumbling block when dealing with massive data sets. The function which is minimized in the $AMDL$ method allows for more than one minimum which poses the question which one of them should the appropriate choice.

We consider another data-adaptive approach by introducing a new function, in the spirit of the information criteria, on the one side, similar to a loss function, on the other side. We then use the minimum of this function, simlarly to the method in SureShrink to determine the number of coefficients retained for use in the wavelet estimate.

3.2 Deterministic case

3.2.1 Notation and assumptions.

For the purpose of the following results we consider ψ, a basic wavelet and ϕ, its corresponding scale function, such that the family of functions

$$\left[\{\phi_{L,k}\}_{0\leq k\leq 2^L-1}, \{\psi_{j,k}\}_{j\geq L, 0\leq k\leq 2^j-1}\right]$$

forms an orthonormal basis in $L^2[0,1]$; the basic wavelet is chosen so that it has a compact support, n vanishing moments and n continuous derivatives. To simplify the notation we write $\phi_{L,k} = \psi_{L-1,k}$. The vector of wavelet coefficients is $\mathbf{w}$. Let $w^2_{(1)} \geq w^2_{(2)} \geq \ldots \geq w^2_{(N)} \geq \ldots$ be the ordered squared values of the wavelet coefficients. Denote $\mathbf{w}_N = (w_1, w_2, \ldots, w_N)$ the DWT of a sample $f(k/N), k = 1, \ldots, N$ of size N and the corresponding non-linear estimate $\hat{\mathbf{w}}(k)$ is obtained from $\mathbf{w}_N$ by retaining the k, $1 \leq k \leq N$ largest in absolute value wavelet coefficients and setting the rest to 0.0.

Model assumptions (Deterministic case)

We assume here that the wavelet coefficients obtained from the DWT of the original signal, are the true parameters and that

$$f = \sum_{j,k} \langle f, \psi_{j,k}\rangle \psi_{j,k} = \sum_{j,k} w_{j,k}\psi_{j,k}\,,$$

the sum is taken over $j \geq L-1, 0 \leq k \leq 2^j - 1$ and the equality is in terms

of the norm in $L^2[0,1]$.

3.2.2 General results.

First we state two results of general importance and then we will proceed with the introduction of the new model selection criterion. The proofs of all results can be found in the Appendix.

Lemma 1. *Let $f \in L^2[0,1]$ is uniformly Lipschitz $\alpha < n$. Denote by*

$$w_{(1)}^2 \geq w_{(2)}^2 \geq \cdots \geq w_{(N)}^2 \geq \cdots \tag{3.1}$$

the values of the ordered, squared wavelet coefficients. If $i = 2^j + m$ for $m \in 1, 2, \ldots, 2^j$, then there exists a constant A such that

$$w_{(i)}^2 \leq \frac{A}{2^{j(2\alpha+1)}} \, .$$

Theorem 10. *Let $f \in L^2[0,1]$ is uniformly Lipschitz $\alpha < n$. Then there exists a constant C such that*

$$\left\| f - \sum_{(j,k)\in M} \langle f, \psi_{j,k} \rangle \psi_{j,k} \right\| \leq C \, K^{-\alpha} \, , \tag{3.2}$$

where M is the set of indexes of the K largest in absolute value wavelet coefficients.

This theorem establishes a result very close in nature to Theorem 4 (De-

Vore et al. [1992]).

3.2.3 New model selection criterion.

Let us consider the following expression:

$$InCr(k,h) = \frac{\|\mathbf{w}_N - \hat{\mathbf{w}}(k)\|^2}{\|\mathbf{w}_N\|^2} + \frac{(k+2)ln(k+2) - (k+2)}{h\,N^h}\,. \qquad (3.3)$$

Here N is a fixed number, $k, (1 \leq k \leq N)$ reflects the number of non-zero coefficients selected for the wavelet estimate $\hat{\mathbf{w}}(k)$, and $h > 0$ is a parameter which allows us to control the number of non-zero wavelet coefficients used in the wavelet vector $\hat{\mathbf{w}}(k)$. That is, if h assumes values close to 0.0 then only few non-zero wavelet coefficients are selected for the estimate $\hat{\mathbf{w}}(k)$, and if h assumes values away from 0.0 then a larger number of non-zero wavelet coefficients are used in the wavelet estimate.

The first term in (3.3) measures the distance between the wavelet estimate and the original signal it approximates. We normalize this expression so the changes in the variability doesn't influence it unduly. The second term serves as a penalty term against the possibility of including too many wavelet coefficients and the risk of undersmoothing the underlying function by including too much of the high frequency spectrum. As mentioned earlier this formula is very similar in spirit to the measure of minimum description length and the $AMDL$ method. As we can see from (2.11) the second term is proportional to the distance between the wavelet estimate and the original

function, while the first term is a function of both the number of coefficients chosen for inclusion and the number of observations in the data set. Once again this first term serves as a buffer against getting to estimate the observed function too closely, at the expense of missing the smooth underlying function.

The motivation for the particular choice of the function in the penalty term can be found in the proof of Proposition 1. The idea is that at first, as we are adding to the model more and more wavelet coefficients, (the wavelet coefficients are being ordered according to their absolute values), the approximation error decreases rapidly. Then a point comes when the rate of decrease of the approximation error slows down significantly and the addition of further wavelet coefficients does not bring a noticeable improvement of the estimate. The function of the second term was chosen so that $InCr$ begins to increase right around that point, that is the point at which the $InCr$ achieves its minimum.

Definition 8. *We consider* $\hat{\mathbf{w}}(K)$ *as an estimate of* $\mathbf{w}_N$*. Here* K *is the value for which (3.3) is minimized when* N *is fixed. We will denote it by* $InCr$*-estimate. Similarly, we will call* $InCr$*-estimate* $\hat{f} = \sum_{I \in M} w_I \psi_I$*, where* M *is the set of indices corresponding to the indices of the non-zero elements in* $\hat{\mathbf{w}}(K)$*.*

Next we will state several results concerning this wavelet estimate and establish some desirable properties.

3.2.4 Properties of the $InCr$-estimate.

Lemma 2. *InCr(k, h) starts out as a decreasing function of k till it reaches a single minimum and begins increasing as k changes from* 1 *to* N.

Proposition 1. *The procedure of selecting the InCr-estimate is equivalent to an estimate obtained through a thresholding procedure (hard threshold function).*

The proof is based on showing that the same models which are considered for selection through the threshold procedure can be selected through the new procedure by choosing appropriate values of h. The full length proof is found in Section A.1.2 of the Appendix.

Corollary 3. *For* $f \in L^p(\mathbb{R})$ *the InCr-estimate converges point-wise to the the underlying function* f *almost everywhere.*

Proof: This follows immediately from Proposition 1 and Theorem 6 (Tao [1991]).

□

What follows is a result providing an estimate for the rate of change of the number of non-zero coefficients in $InCr$-estimate and its corresponding threshold value as the number of observation N increases.

Theorem 11. *Let* $f \in L^2[0,1]$ *be uniformly Lipschitz* $\alpha \leq n$ *on the interval* $[0,1]$. *Let* $\mathbf{w}_N$ *be the DWT of a sample* $f(i/N), i = 1, 2, \ldots, N$.

(i) If $w^2_{(1)} \geq w^2_{(2)} \geq \ldots \geq w^2_{(K)}$ *are the ordered squared values of the non-zero wavelet coefficients in the InCr-estimate* $\hat{\mathbf{w}}(K)$*, then*

$$K = O\left(N^{h/(s(N)+2\alpha+1)}\right) \quad where \quad s(N) = \frac{ln(ln(N-1))}{ln(N)} \; .$$

(ii) If T_{InCr} *is the corresponding threshold value, then*

$$T_{InCr} = O\left(\frac{ln(N)}{(s(N)+2\alpha+1)N^h}\right) \; .$$

The proof is in Section A.1.2 of the Appendix. This result shows that when the function f is relatively smooth, that is $\alpha \geq 1$, for example, the number of non-zero coefficients in the $InCr$-estimate increases slower with the increase in the number of observations N, compared to a function which is less smooth, $\alpha < 1$.

3.3 Stochastic case

3.3.1 Notation and assumptions.

Model assumptions (Stochastic case) Based on the equivalence of the model stated in terms of the functional values and the data observations with the model set in terms of wavelet coefficients and their estimates, we can focus our attention to the following model. Let $w_1, w_2, \ldots, w_N$ be the wavelet coefficients based on N noisy, equally spaced, observations from a

function $f(t)$. Assume also that

$$w_i = \theta_i + \epsilon z_i \tag{3.4}$$

where $\{\theta_i\}_{i=1}^N$ are the true wavelet coefficients and $\{z_i\}_{i=1}^N$ are independent standard normal variables, $N(0,1)$, with mean 0 and variance 1.

3.3.2 General results.

First we state two results of general importance and then we proceed with the introduction of the corrected model selection criterion, used in the case when the data comes from noisy observations. The proofs of all results can be found in the Appendix.

Lemma 3. *If $Z_1^2, Z_2^2, \ldots, Z_n$ are iid χ_1^2 then we have the following expression for the mean of the largest order statistic:*

$$E[Z_{(n)}^2] = 2^{n+1} \int_0^\infty u\,(1 - \Phi(u))^n \, du = G(n)\,, \tag{3.5}$$

where $\Phi(u)$ is the cumulative distribution function of the standard normal distribution.

Theorem 12. *Assuming model (3.4) we can obtain the following upper bound for the mean square error:*

$$E\|\mathbf{f} - \hat{\mathbf{f}}(K)\|^2 \leq \frac{G(N)}{N}\sigma^2\, K + \left(\sum_{i=1}^{N}\theta_i^4\right)^{1/2}(N-K)^{1/2}\ .$$

Here K indicates the number of wavelet coefficients retained for the wavelet estimate, and $\hat{\mathbf{f}}(K)$ is the IDWT of this wavelet estimate.

This bound reflects the fact that there is an optimal number of the largest in absolute value coefficients, which need to be included in the wavelet estimate; a lesser or larger number than this optimal one will produce a larger mean square error.

Corollary 4. *Let $f \in L^2[0,1]$ is uniformly Lipschitz $\alpha < n$, (n is the number of vanishing moments of the mother wavelet). Then under the assumption of Theorem 12 we have that*

$$E\|\theta - \hat{\mathbf{w}}(K)\|^2 \leq G(N)\epsilon^2\, K + C(\alpha)(N-K)^{1/2}\ .$$

Corollary 5. *Under the assumption of Theorem 12 with high probability we can assert that*

$$E\|\theta - \hat{\mathbf{w}}(K)\|^2 \leq 2ln\ N\epsilon^2\, K + \left(\sum_{i=1}^{N}\theta_i^4\right)^{1/2}(N-K)^{1/2}\ .$$

Theorem 13. *The bias of the estimate of the approximation error $\|\mathbf{w}_N -$*

$\hat{\mathbf{w}}(K)\|^2$ *is bounded as follows:*

$$\left| E\|\mathbf{w}_N - \hat{\mathbf{w}}(K)\|^2 - E\|\theta - \hat{\mathbf{w}}(K)\|^2 \right| \leq \epsilon^2(N-k) + \epsilon^2 k G(N) \, .$$

The proof of this result serves as a justification of considering the following corrected form of the model selection criterion.

3.3.3 Corrected model selection criterion.

For the stochastic case we consider the following corrected version of the model selection criterion (3.3) introduced earlier in this chapter.

$$\begin{aligned} InCr(k,h) = & \frac{\|\mathbf{w}_N - \hat{\mathbf{w}}(k)\|^2 + (N-2k)\hat{\epsilon}^2}{\|\mathbf{w}_N\|^2 - N\hat{\epsilon}^2} \\ & + \frac{(k+2)ln(k+2) - (k+2)}{h\, N^h} \, . \end{aligned} \tag{3.6}$$

Let us remind here that $\epsilon^2 = \sigma^2/N$, where σ^2 is the variance assumed in the nonparmetric regression model, and ϵ^2 is the variance of the discrete wavelet coefficients obtained through the fast pyramid algorithms. When no noise is assumed, that is $\epsilon = 0$, the corrected model selection criterion reduces to the form used in the deterministic case. The motivation for this formula comes from the heuristic argument that the first term, after the correction, is close to the value of the first term as if it were a deterministic case. In the stochastic case it is vital that we include only as many wavelet coefficients as it is indicated by the minimum value of the mean square error (MSE).

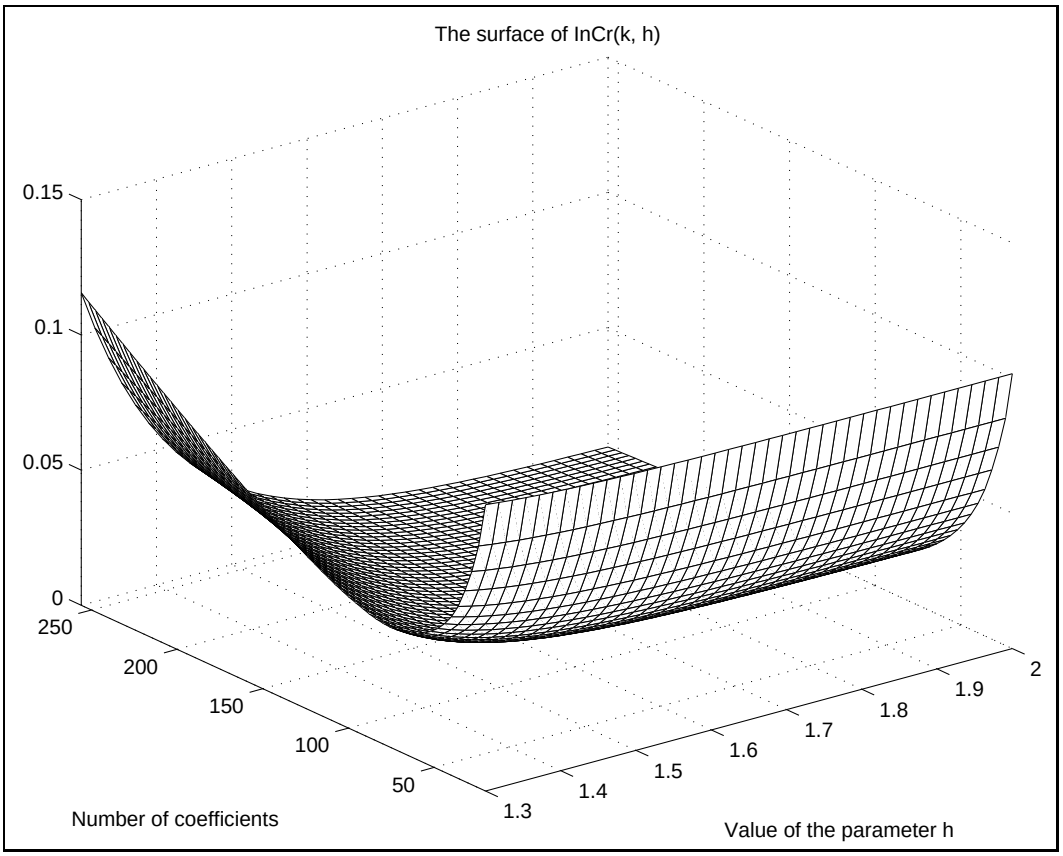

Figure 3.1: The $InCr$ surface.

Since we don't know the value of the actual MSE, our idea is to use the $InCr$ to imitate the deterministic case with the assumption that through the careful choice of the function in the second term, the value at which the $InCr$ achieves its minimum is the approximately the same as the one that minimizes the MSE. To land some credence to this heuristic argument we made a plot of the $InCr(k, h)$ surface as we change the number of coefficients k and vary the value of the parmeter h. Figure 3.1 shows how for a fixed

value of the parameter h the corresponding curve in the $InCr(k,h)$ surface begins by decreasing, then the change slows down and starts rising again at the other end of the values of k. Our assertion is that where this change takes place is approximately at the point which minimizes the MSE.

Following the notation established earlier, $\mathbf{w}_N$ is the vector of noisy wavelet coefficients, while $\hat{\mathbf{w}}(K)$ is the estimate of the true wavelet coefficients $\theta = \{\theta_1, \theta_2, \ldots, \theta_N\}^t$, obtained by keeping the K largest in absolute value noisy wavelet coefficients. Here K_{min} is the value that minimizes (3.6) for an appropriately chosen value of the parameter h.

3.3.4 Properties of the $InCr$ - estimate.

Lemma 4. *InCr(k, h) starts out as a decreasing function of k till it reaches a single minimum and begins increasing as k changes, as long as* $w^2_{(k)} > 2\epsilon^2$ *holds.*

Proposition 2. *The expected value of the wavelet coefficient at which (3.6) attains its minimum, when we consider the parameter h fixed, is given by:*

$$E\, w^2_{(K_{min})} = \frac{ln(a+2)}{hN^h}\|\theta\|^2\,,$$

where $a \in (k-1, k)$ *is a constant.*

Corollary 6. *There exists a value of the parameter h such that:*

$$E\|\theta - \hat{\mathbf{w}}(K_{min})\|^2 \leq E\|\theta - \hat{\mathbf{w}}^T(\lambda)\|^2\,.$$

where $\hat{\mathbf{w}}^T(\lambda)$ *is the hard threshold estimate with a threshold value* λ.

Proposition 3. *For an estimate of the parameter* h *we can choose the value for which the following equality holds:*

$$\frac{(k+2)ln(k+2)-(k+2)}{h\,N^h} = \frac{(k+2)(\mathbf{w}^2_{(k+1)}-2\epsilon^2)}{\|\mathbf{w}_N\|^2-N\epsilon^2}\ .$$

Chapter 4

Applications

4.1 Comparative study (synthesized signals)

4.1.1 Description of the methods.

Below is a list of the methods we are going to compare in the next section. Note that all methods require a pre-determined level L for the coarsest resolution.

Method 1. The method *VisuShrink* defined in section 2.6.1 uses a threshold value which depends on the number of observations and the standard deviation of the signal, but not on the actual data values.

Method 2. The method *RiskShrink* defined in section 2.6.2 uses a threshold value which is more data dependent.

Method 3. The method *SureShrink* defined in section 2.6.3 uses a different threshold value for each resolution level.

Method 4. The method described under the name AMDL defined in section 2.7.1.

Method 5. Minimizing $InCr(k, h_0)$ (formula (3.3)) for k after an appropriate choice of $h = h_0$.

Method 6.
(i) Find k_1 which minimizes $InCr(k, h_0|f)$, with a given value for $h = h_0$;

(ii) find k_2 which minimizes $InCr(k, h_0|Res)$, where $Res = (r_1, \ldots, r_n)^T$ is the vector of residuals $r_i = f_i - \hat{f}_i, i = 1, \ldots, n$;

(iii) the estimate $\tilde{f}$ is found as the IDWT of $\hat{\theta}_f + \hat{\theta}_{Res}$, where $\hat{\theta}_f, \hat{\theta}_{Res}$ are the $InCr$-vectors of wavelet coefficients for f and Res, correspondingly.

A brief argumentation for the last method can be made as follows. If $f = \hat{f} + Res$, where f is the observed signal, $\hat{f}$ is the wavelet estimate and Res is the vector of residuals, then $Wf = W\hat{f} + WRes$, where the orthogonal matrix W defines the DWT. The last statement can be written as: $\hat{\theta}_f = \hat{\theta}_{\hat{f}} + \hat{\theta}_{Res}$.

4.1.2 Quantities for the comparison.

For each of the methods listed in the previous section, and a set of different types of signals, we will calculate and compare the following quantities:

1. k the number of coefficients included in the reconstruction of the wavelet estimate $\hat{f}$;

2. relative error:

$$RelErr = \frac{\|f - \hat{f}\|}{\|f\|} = \sqrt{\frac{\sum_{i=1}^{n}(f_i - \hat{f}_i)^2}{\sum_{i=1}^{n} f_i^2}};$$

3. compression ratio (in percentages):

$$CR = (1 - \frac{k}{n}) \times 100$$

4. AMDL.

In all examples the number of observations N is chosen to be 1024 and the coarsest level is $L = 2$.

Figure 4.1 displays the plots of the signals for which we are comparing the different methods.

The first four examples are taken from signal making function provided by the WAVELET portion of Matlab; in particularly, we chose the signals known as *doppler*, *bumps*, *heavisine* and *blocks*, as seen in Donoho and

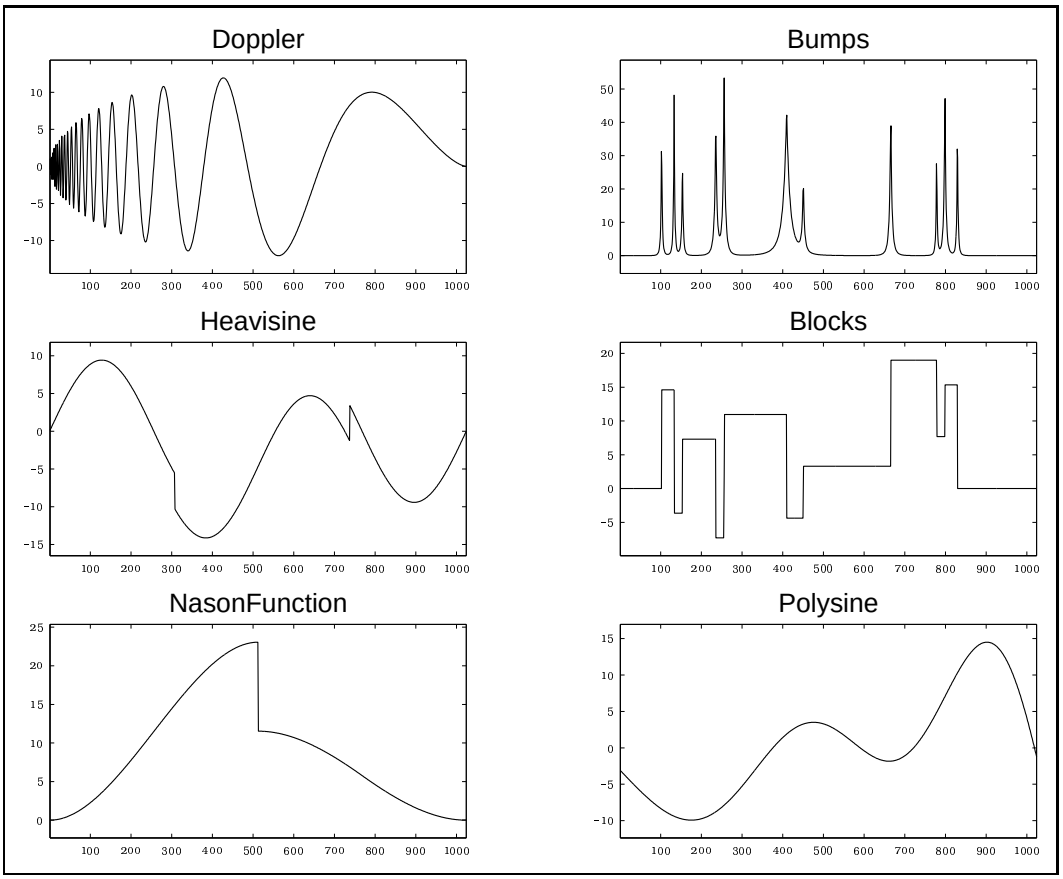

Figure 4.1: Synthesized signals (no noise).

Johnstone [1995]. The Nason's function is defined in Nason [1996] as:

$$f(t) = \begin{cases} 4t^2(3-4t) & \text{for } x \in [0, 0.5] \\ \frac{4}{3}t(4t^2 - 10t + 7) - \frac{3}{2} & \text{for } x \in [0.5, 0.75] \\ \frac{16}{3}t(t-1)^2 & \text{for } x \in [0.75, 1]. \end{cases}$$

The last signal is a product of a polynomial and a trigonometric polynomial, restricted to the interval $[0, 5]$:

$$f(t) = (t^3 - 3t^2 + t - 10)(sin(2t) + 0.5)\,.$$

A remark on the thresholding methods. We should like to point out that for the application of the methods *VisuShrink* and *RiskShrink* the signals need to be "normalized" so that the median of their set of wavelet coefficients equals 0.0. On the other side, the method *SureShrink* requires that the signal has a standard deviation 1. For all three methods it is essential that the added noise then comes for the standard normal distribution $N(0, 1)$.

The four signals *doppler*, *bumps*, *heavisine* and *blocks*, as created by the signal making function, have a median close to 0.0 and the procedure which "normalizes" them to make this median, exactly 0.0, multiplies the signals by a number of a large magnitude (10^{12}). On the other side, this makes the standard deviation large as well. In order to apply all methods to an identical collection of signals, we pre-processed all signals in the same way, as done in Donoho and Johnstone [1994]. Each signal is multiplied by the number $\frac{7}{\hat{\sigma}}$,

where $\hat{\sigma}$ is the standard deviation of the corresponding signal. When noise is added, it is taken from the standard normal distribution.

The importance of this information lies in the realization that the results produced by the methods $VisuShrink$, $RiskShrink$, and $SureShrink$ are highly influenced by the type of "normalization" they are made to undergo prior to the applications of these methods.

4.1.3 Results for signals without noise.

In the case when no noise is added to the original signal, the model taking in all 1024 wavelet coefficients, will be the one with the smallest relative error and the error is 0.0. This implies that the model using the K largest wavelet coefficients, $1 \leq K \leq 1024$ will have an increasing relative error as the number of non-zero coefficients K decreases. Thus the task of compressing the data is contingent upon the size of the relative error deemed acceptable. The results from the last three methods, the methods proposed in this work, reported in the tables below are aimed at showing that by varying the value of the parameter h we can achieve either a high compression ratio, or a high precision level, or a result that balances the two aspects in some desirable fashion.

Doppler **signal.** Table 4.1 shows the results obtained by the various methods applied to the *doppler* signal and Figure 4.2 displays the *doppler* signal with its wavelet estimates. We note that while the first two methods have

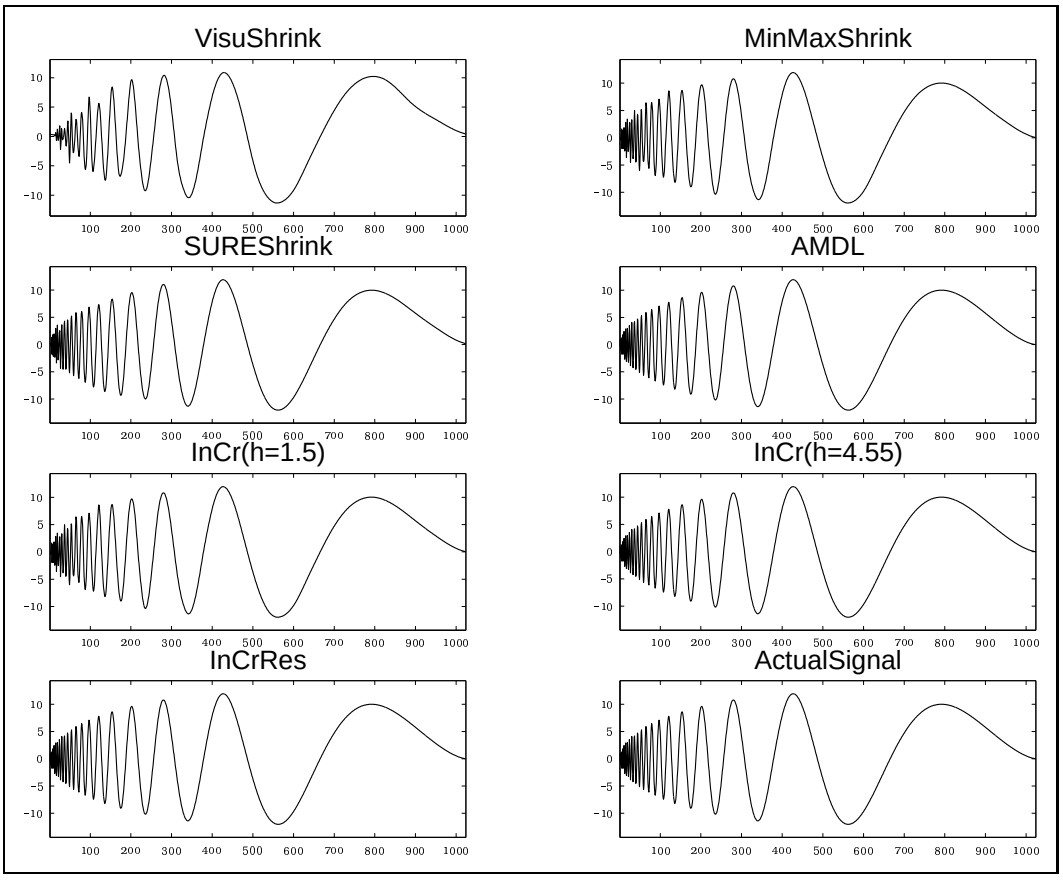

Figure 4.2: Reconstructions of the *doppler* signal.

Table 4.1: Results for the *doppler* signal.

#	**METHOD**	**k**	**RelErr**	**CR**	**ADML**
1	$VisuShrink$	42	0.114641	96%	5444.6115
2	$RiskShrink$	54	0.030114	95%	3649.7033
3	$SureShrink$	704	0.023173	31%	13690.9566
4	$AMDL$	619	0.000000	40%	-23264.2716
5	$InCr(k, h = 1.5)$	54	0.030114	95%	3649.7033
6	$InCr(k, h = 4.55)$	252	0.000000	75%	-9762.1275
7	$InCr(k, h = 2.58\|f)+$ $InCr(k, h = 2.45\|Res)$	252	0.000000	75%	-9762.1275

Table 4.2: Results for the *bumps* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	113	0.178159	89%	7261.4860
2	$RiskShrink$	143	0.044974	86%	5677.8135
3	$SureShrink$	719	0.015946	30%	11169.8261
4	$AMDL$	1023	0.000000	0%	-17846.4106
5	$InCr(k, h = 1.55)$	143	0.044974	86%	5677.8135
6	$InCr(k, h = 4.6)$	480	0.000000	53%	-6163.4984
7	$InCr(k, h = 2.6\|f)+$ $InCr(k, h = 2.55\|Res)$	481	0.000000	53%	-6208.8457

good compression ratio but poor relative error, the next two methods retain a larger number of coefficients for the model. Method $AMDL$ compensates for this by attaining a good precision. Line 5 in the table shows that the $InCr$-estimator can match the compression ratio of method $RiskShrink$. Line 6 and 7 show that we can attain a high precision level, as well. Yet, a more balanced approach might be a result which uses a relatively low number of non-zero coefficients while keeping the relative error less than $10^{(-3)}$, for example. When we set $h = 2.7$ the number of non-zero coefficients is $k = 136$ and the relative error is **RelErr**= 0.000400 with a compression ratio **CR**= 87%. That is, by including $((136-54)/1024)*100\% = 8\%$ more coefficients we decrease the relative error by $((0.030114-0.000400)/0.030114)*100\% = 98.7\%$.

Bumps **signal.** Table 4.2 shows the results obtained by the various methods applied to the *bumps* signal and Figure 4.3 displays the signal *bumps* with its wavelet estimates. Similarly, here the first three methods have rather poor relative error, the method $AMDL$ includes too many coefficients in

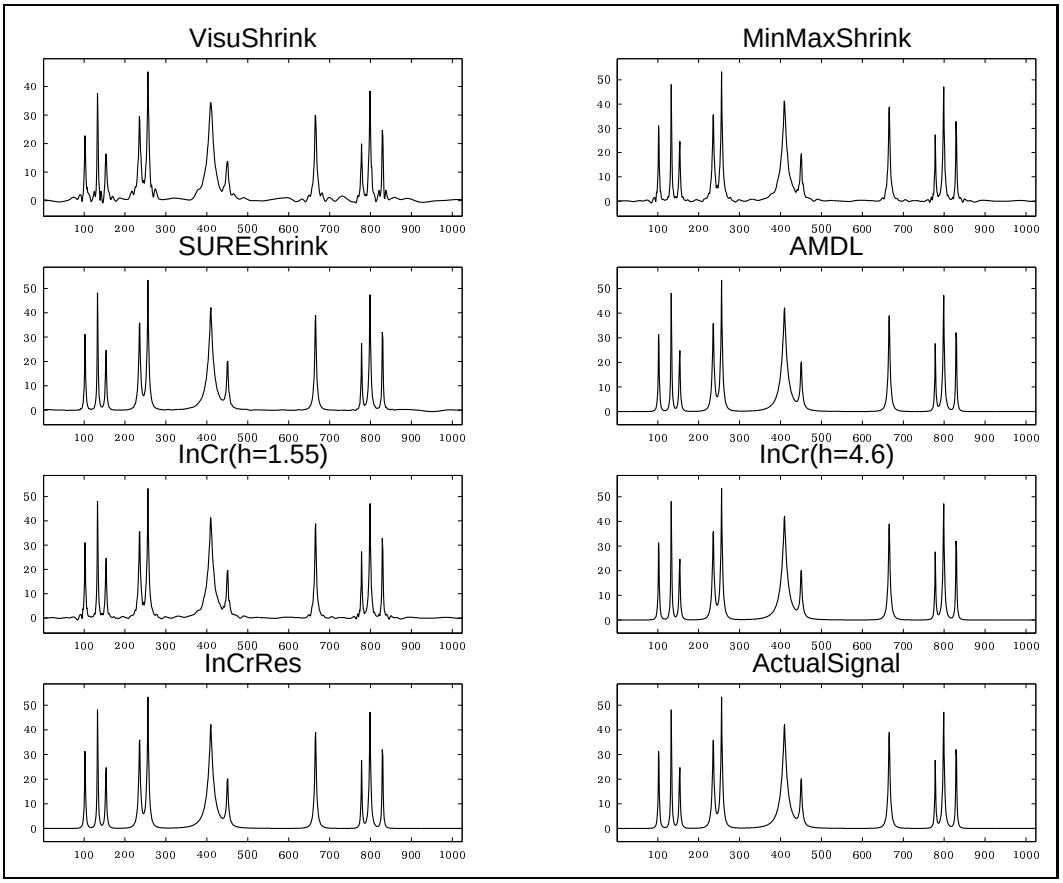

Figure 4.3: Reconstructions of the *bumps* signal.

Table 4.3: Results for the *heavisine* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	14	0.068612	99%	4302.6886
2	$RiskShrink$	24	0.027648	98%	3109.9411
3	$SureShrink$	419	0.404897	59%	10554.9726
4	$AMDL$	194	0.000000	81%	-26470.4345
5	$InCr(k, h = 1.48)$	24	0.027648	98%	3109.9411
6	$InCr(k, h = 4.35)$	133	0.000000	87%	-11403.9114
7	$InCr(k, h = 0.26\|f)+$ $InCr(k, h = 0.2\|Res)$	134	0.000000	87%	-11730.4448

the reconstruction but attains a high precision level. Method $SureShrink$ works poorly in both aspects. The results in the last three rows show that our methods can be adjusted to achieve either a high compression ratio or low relative error, depending on what the needs of the user might be. One more time, a more balanced result can be obtained by setting $h = 2.75$, for which **k**= 319, **RelErr**= 0.000316 and **CR**= 69%. Thus by adding on $((319 - 143)/1024) * 100\% = 17.2\%$ more coefficients we improve the relative error by $((0.044974 - 0.000316)/0.044974) * 100\% = 99.3\%$.

Heavisine **signal.** Table 4.3 shows the results obtained by the various methods applied to the *heavisine* signal and Figure 4.4 displays the *heavisine* signal with its wavelet estimates. The first two methods use too high a threshold, and the third method, which chooses the threshold according to the the different resolution levels, does a poor job selecting their values. Method $AMDL$ together with the last two examples produce good results by succeeding to attain a high precision while using a fraction of all wavelet

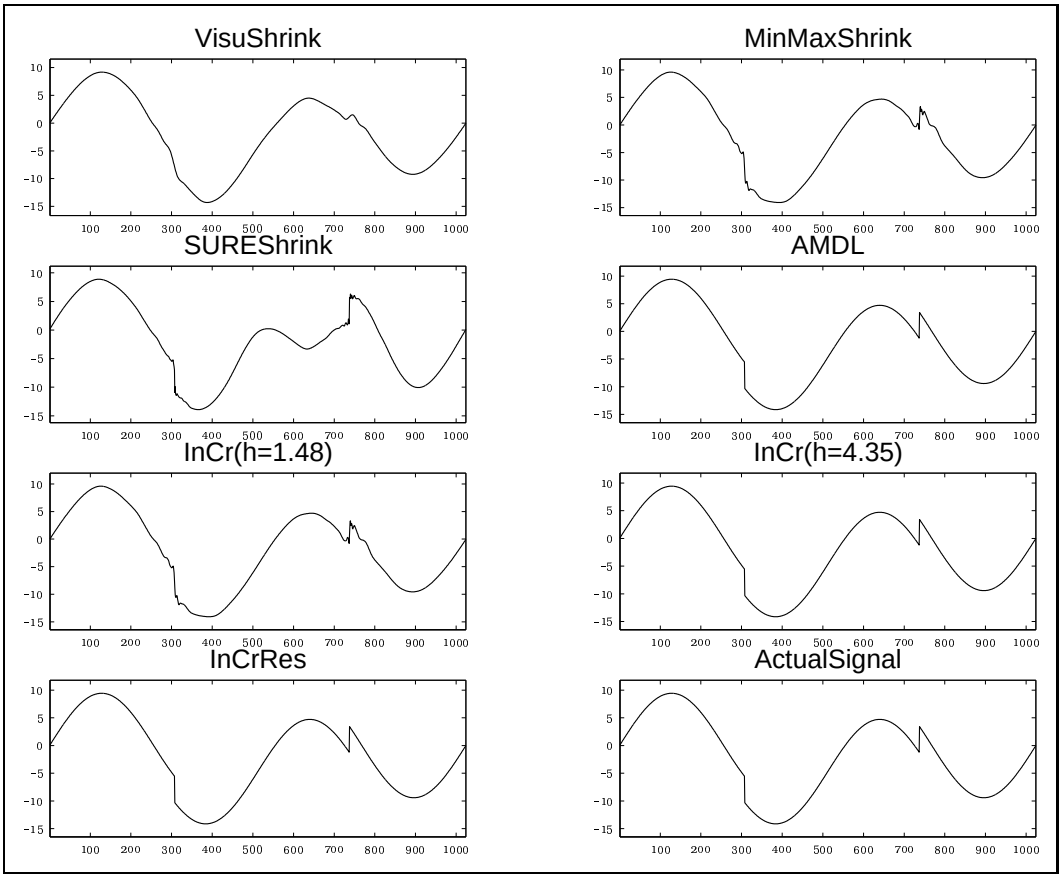

Figure 4.4: Reconstructions of the *heavisine* signal.

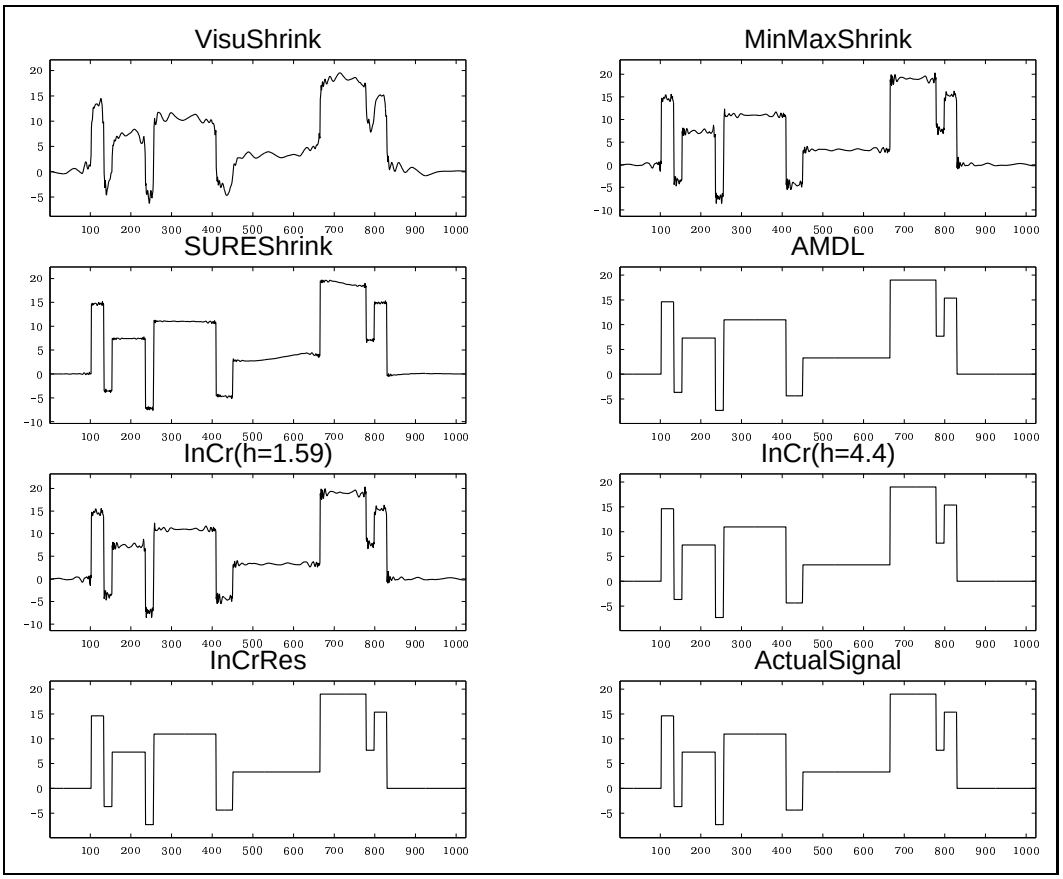

Figure 4.5: Reconstructions of the *blocks* signal.

coefficients. Line 5 shows that we can match the compression ratio of the method $RiskSrink$, but we feel the loss of precision is too high in this case.

Blocks **signal.** Table 4.4 shows the results obtained by the various methods applied to the *blocks* signal and Figure 4.5 displays the *blocks* signal with its wavelet estimates. This signal allows only a moderate compression. Once the number of wavelet coefficients included in the reconstruction falls bellow the

Table 4.4: Results for the *blocks* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	95	0.140093	91%	6889.5415
2	$RiskShrink$	123	0.040294	88%	5468.6390
3	$SureShrink$	645	0.037375	37%	11813.1356
4	$AMDL$	391	0.000000	62%	-26146.7557
5	$InCr(k, h = 1.59)$	123	0.040294	88%	5468.6390
6	$InCr(k, h = 4.4)$	373	0.000000	64%	-7666.9233
7	$InCr(k, h = 0.24\|f)+$ $InCr(k, h = 0.2\|Res)$	373	0.000000	64%	-7666.9233

Table 4.5: Results for the Nason's function.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	20	0.048572	98%	4509.6653
2	$RiskShrink$	27	0.010726	97%	2383.3056
3	$SureShrink$	418	0.038901	59%	10300.4220
4	$AMDL$	191	0.000000	81%	-28319.7502
5	$InCr(k, h = 1.6)$	27	0.010726	97%	2383.3056
6	$InCr(k, h = 4.45)$	118	0.000000	88%	-11089.3748
7	$InCr(k, h = 2.4\|f)+$ $InCr(k, h = 2.4\|Res)$	118	0.000000	88%	-11089.3748

350 largest in absolute values coefficients, the relative error starts growing rapidly. This might be an indication that the *symmlet* with 8 vanishing moments is not so suitable a choice when approximating a function which is piecewise constant.

Nason's function. Table 4.5 shows the results obtained by the various methods applied to the Nason's function and Figure 4.6 displays the Nason's function with its wavelet estimates. This is a piece-wise polynomial function with a discontinuity at $t = 0.5$. Noting that the relative error is a bit too

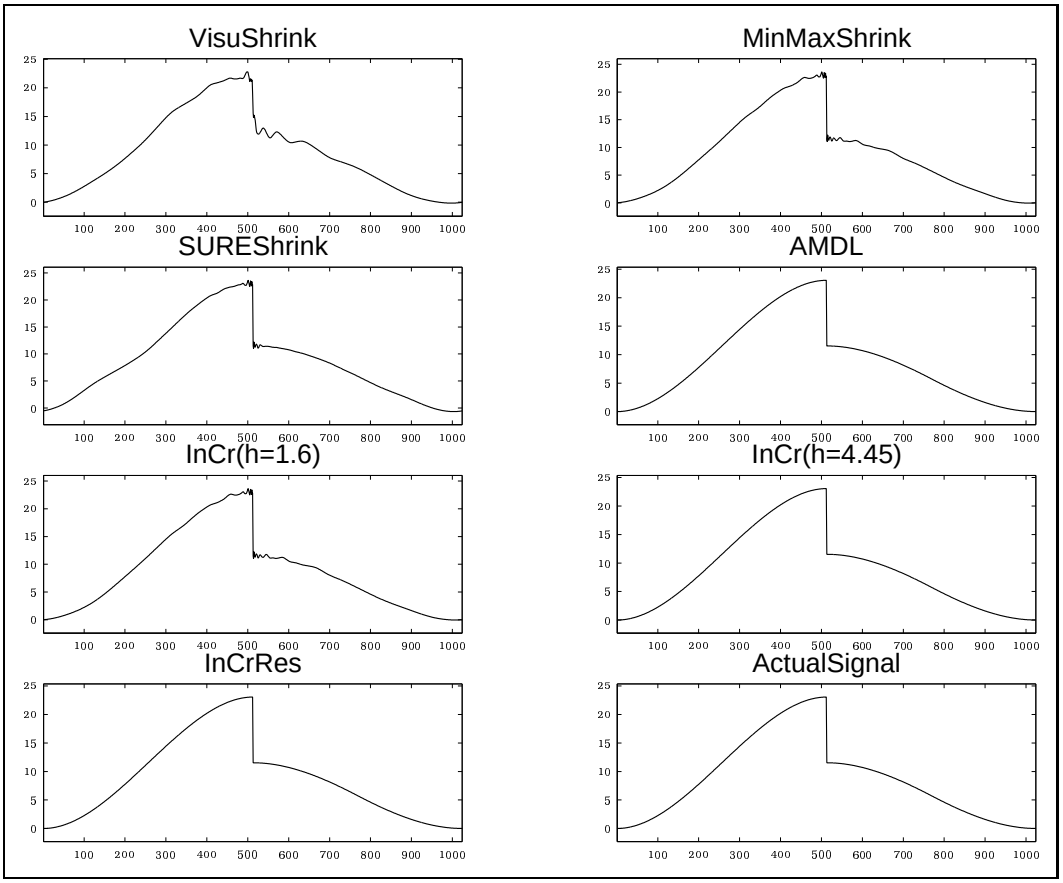

Figure 4.6: Reconstructions of the Nason's function.

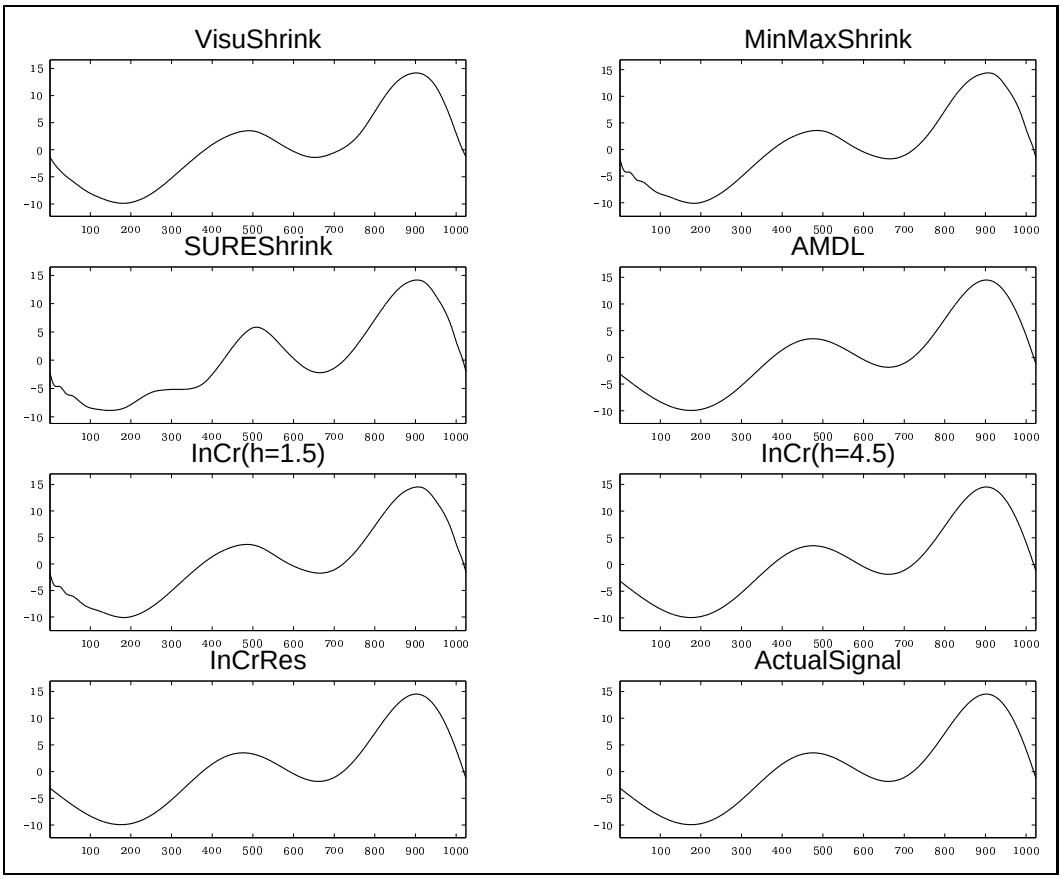

Figure 4.7: Reconstructions of the *polysine* function.

large in the estimates produced by the first three methods, the rest of the methods show good results in both aspects of compressing the data and approximating them closely.

Polysine **function.** Table 4.6 shows the results obtained by the various methods applied to the *polysine* function and Figure 4.7 displays the *polysine* function with its wavelet estimates.

Table 4.6: Results for the *polysine* function.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	12	0.055355	99%	3900.6316
2	$RiskShrink$	14	0.020728	99%	2479.5019
3	$SureShrink$	397	0.222424	61%	10426.4276
4	$AMDL$	258	0.000000	75%	-28328.1072
5	$InCr(k, h = 0.2)$	14	0.019917	99%	2479.5019
6	$InCr(k, h = 4.5)$	84	0.000000	92%	-12519.6302
7	$InCr(k, h = 2.35\|f)+$ $InCr(k, h = 2.42\|Res)$	84	0.000000	92%	-12519.6302

What we can notice here is that while the results of line 5 match the compression ratio of the method $RiskShrink$, the relative error is slightly smaller. This is due to the fact that in practice the shrinkage is not applied to the coarsest level of coefficients, which at times might not be such a bad idea, as this example shows.

As mentioned earlier, if we use all $N = 1024$ (the length of data set itself) wavelet coefficients obtained from the discrete wavelet transform, the reconstruction of the original signal is lossless. Compressing the data, or discarding some of the small in absolute values wavelet coefficients, will inevitably lead to worse fit, or losses in the information. Thus data reduction should balance out the decreasing precision of the fit with the increasing value of the compression ratio, in order to produce acceptable results. Below are some observations about the workings of the different methods considered in this section with regard to data compressions as a goal.

- The first two methods compress the data too much producing larger relative error. Let us point out once again, that for the application

of these two methods the signals need to be "normalized" so that the median of the set of the wavelet coefficients is 0.0. This is a step in which a hidden adjustment of the signal to the methods is achieved. Yet the fact that the results very so widely depending on the standard deviation of the data, might be considered somewhat troubling.

- Method $SureShrink$ estimates the threshold values according the level of resolution. Although this might be a good idea in general, this particular method produces so poor results in both the compression and the approximation aspects, that these selections of threshold values doesn't seem so suitable.

- Method $AMDL$ achieves a very high precision level, but if we deem some loss of information permissible with the idea of using a lower number of parameters in the model, then this method cannot give us the desired results.

- The $InCr$-estimator allows for adjustments of the method in order to accommodate results which the user might consider desirable. The methods based on the $InCr$ can produce an estimator with a very high compression ratio, or a very low approximation error, or still yet, an estimator which attains some preferable balance between these two aspects.

- The methods based on the model selection criterion are on the one side

computationally easy, on the other side are data dependent. *VisuShrink* while the simplest computationally, applies the same "pruning" for any signal of a fixed length, if the standard deviation remains the same. Methods 2 and 3 adapt to the particulars of the data but are computationally more intense. Method 4 is both data dependent and simple to calculate but may have more than one local minimum, and is keeping too many coefficients.

- The first three methods leave the top resolution levels intact not as a part of the derived algorithm but as a rule that appears to be working in practice. The methods based on the model selection criterion, although often have the same effect, achieve it as an integral part of the selection process.

4.1.4 Results for noisy signals.

In the research so far, when noise is added to a pure signal, it is usually of a constant variance throughout the length of the signal. The added error is normally distributed with mean 0.0 and standard deviation 1.0. We followed the example of Donoho and Johnstone [1994] in that the signal-to-noise ratio is chosen to be 7.0 for all signals.

In the case when noise is added to the signal it is no longer true that the larger the number of coefficients one uses in the model selection the better the precision of the estimator is going to be. The noise contamination prevents

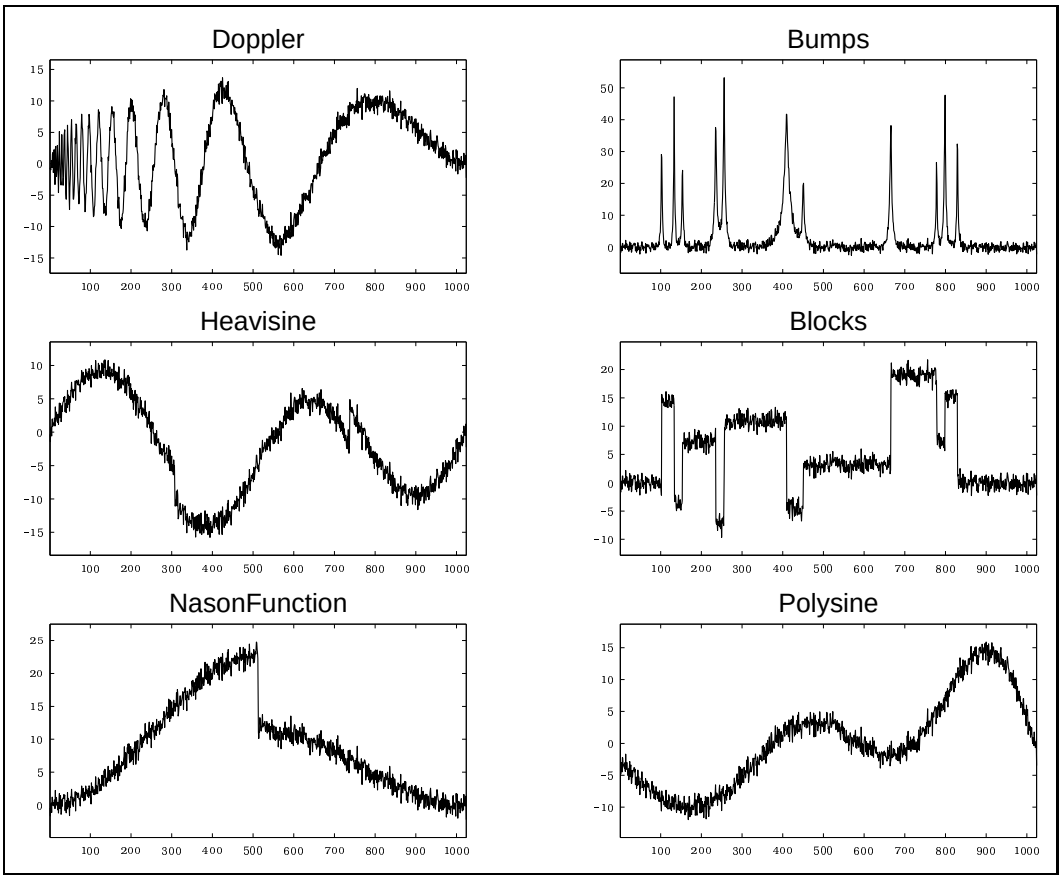

Figure 4.8: Synthesized signals (noise added).

the wavelet coefficients from their natural decay towards the value 0.0. Thus the precision level depends on the variance of the noise and cannot be made arbitrarily small. The difficulty then lies in finding just the right number of coefficients so that the mean square error (MSE) is close to its minimum.

Figure 4.8 displays the plots of the signals considered earlier with noise added to parts of them.

The following comments apply uniformly to all noisy signals.

- Methods 4, 5, and 6 have very similar results in that they eliminate successfully the artificially added noise, but lose relevant information in the process; comparing the number of coefficients required for the recovery of the pure signal and the number of coefficients used for that purpose in the case of the noisy signals, the former number is larger, that is the data are under-fitted.

- Method *VisuShrink* produces the smoothest estimates but they often under-fit the actual data.

- Methods *RiskShrink* and *SureShrink* fail to exclude the noise from their estimates, and the latter method seems to choose the coefficients retained for reconstruction somewhat poorly.

- The fact that the last methods show only a slight improvement of the results, reflected in the relatively smaller approximation error, might be explained by the observation that all wavelet coefficients are affected by noise. Although we might be able to re-capture relevant features of the signal, it comes with the price of adding on noise, as well.

- Methods 5 and 6, nevertheless, have an advantage in that by varying the values of the parameter h they provide a way to find the optimal number of coefficients which produces the lowest relative error.

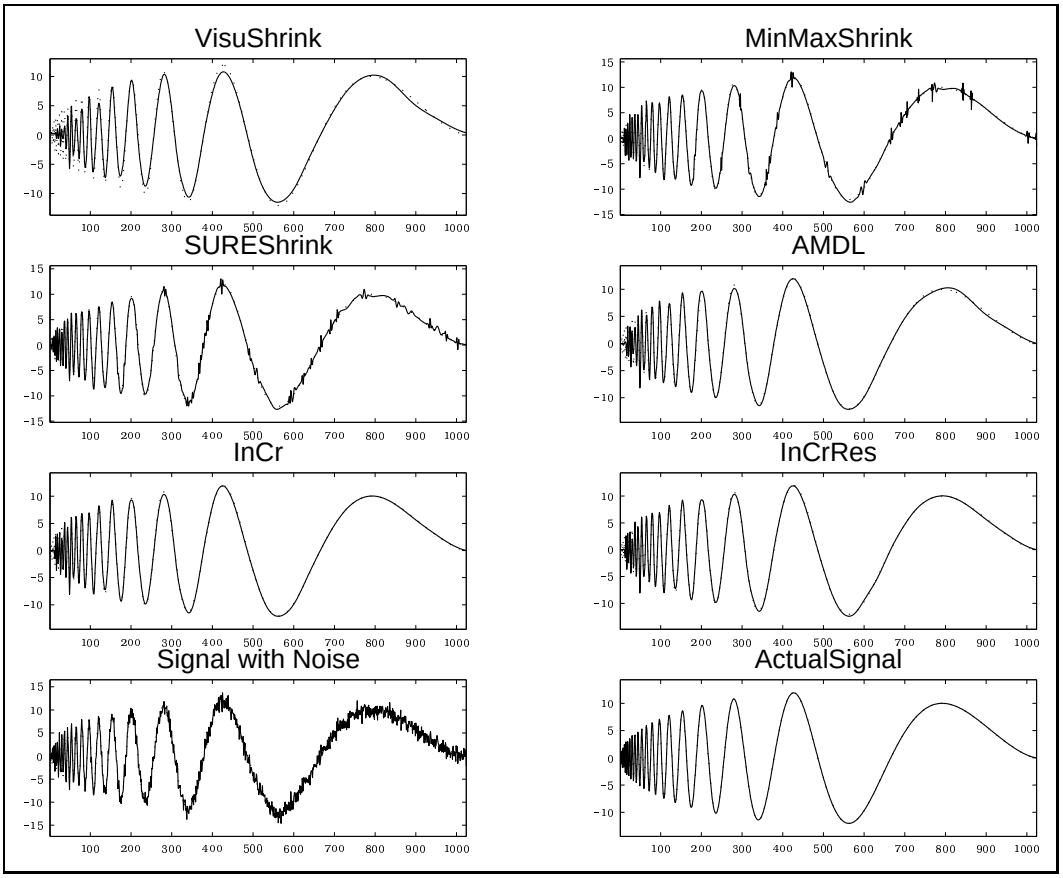

Figure 4.9: Reconstructions of the noisy *doppler* signal (dotted line is the original signal).

Table 4.7: Results for the noisy *doppler* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	43	0.114037	96%	5451.8187
2	$RiskShrink$	81	0.071533	92%	5332.8542
3	$SureShrink$	103	0.073128	90%	5695.4314
4	$AMDL$	37	0.075625	96%	4755.0240
5	$InCr(k, h = 1.4)$	47	0.053055	95%	4381.3804
6	$InCr(k, h = 1.4\|f)+$ $InCr(k, h = 0.4\|Res)$	48	0.053787	95%	4416.6085

Noisy *doppler* signal. Table 4.7 shows the results obtained by the various methods applied to the noisy *doppler* signal and Figure 4.9 displays the reconstructions of the noisy *doppler* signal.

Methods *Riskshrink* and *SureShrink* fail to completely de-noise the signal. Method *Visushrink* sets its estimate on the other extreme; that is it does de-noise the signal but cuts some of the existing signal components as well. Method $AMDL$ denoises the signal but suffers from accessive removing of coefficients which is most noticeable in the part where the signal oscillates at a higher frequency. Out two methods do the best job of estimating the true signal while successfully de-noising it at the same time. the reduction rate of all signals is good but clearly method *SureShrink* picks its coefficients most inappropriately. The fact that the $AMDL$ statistic achieves its minimum at a point which is different from the point which minimizes the MSE, suggests that better estimates can be selected.

Figure 4.10 displays the residuals obtained from the reconstructions of the *doppler* signal when the various methods are applied. The magnitude

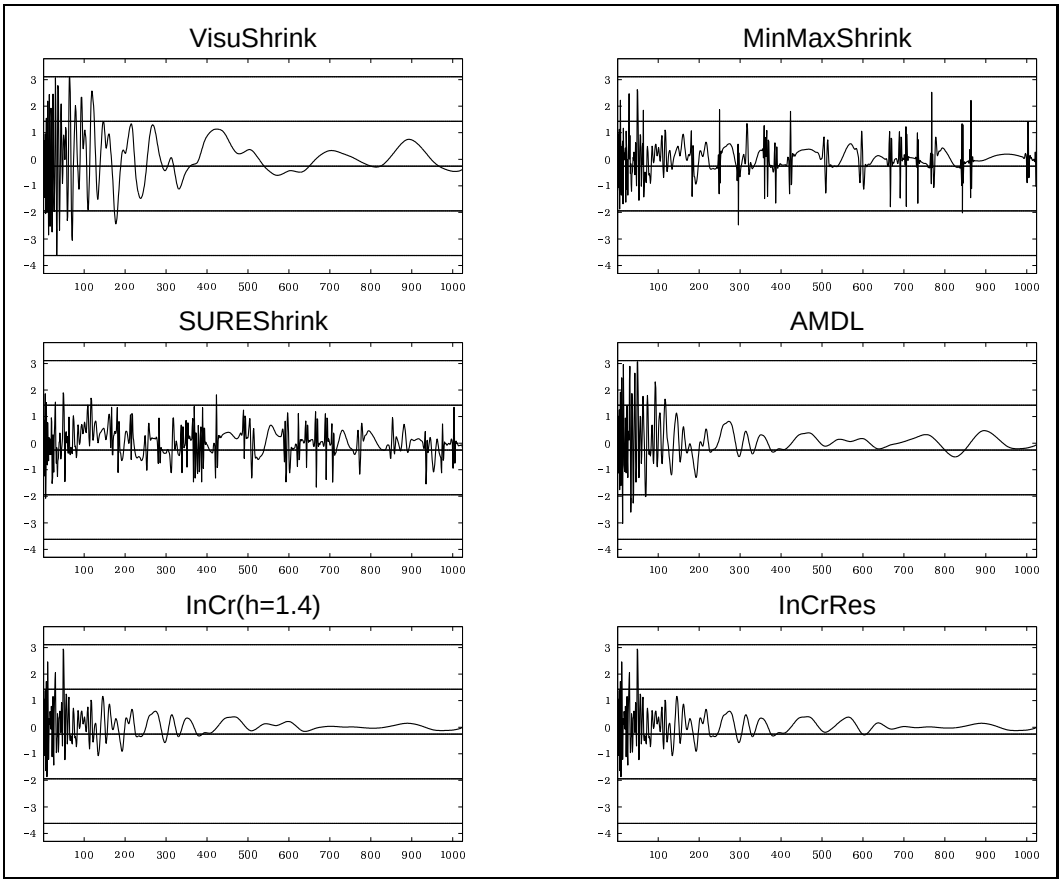

Figure 4.10: Residuals of the *doppler* signal.

Table 4.8: Results for the noisy *bumps* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	113	0.178136	89%	7261.2963
2	$RiskShrink$	173	0.082008	83%	7015.2925
3	$SureShrink$	300	0.093073	71%	9107.2654
4	$AMDL$	99	0.115177	90%	6407.0617
5	$InCr(k, h = 1.5)$	142	0.076524	86%	6448.0520
6	$InCr(k, h = 1.5\|f)+$ $InCr(k, h = 0.8\|Res)$	144	0.075976	86%	6467.4159

and location of the residuals lends additional support to the assertions from above. All methods exhibit larger residuals at the beginning, where the oscillation of the original signal is rapid and of high amplitude. At the other end $VisuShrink$, and the last three methods de-noise the signal successfully. Yet $VisuShrink$ allows for larger deviations from the original signal than the other three methods.

Noisy *bumps* signal. Table 4.8 shows the results obtained by the various methods applied to the noisy *bumps* signal and Figure 4.11 displays the reconstructions of the noisy *bumps* signal.

With the exception of $SureShrink$ all methods reduce the number of coefficient to acceptable levels, the trouble is that these coefficients are either not in the right location or not of the appropriate size, or both, since the relative errors are rather large. The flexibility of methods 5 and 6 allow them to find the optimal balance between relative error and compression ratio. Figure 4.12 displays the residuals obtained from the reconstructions of the *bumps* signal when the various methods are applied. The magnitudes of the

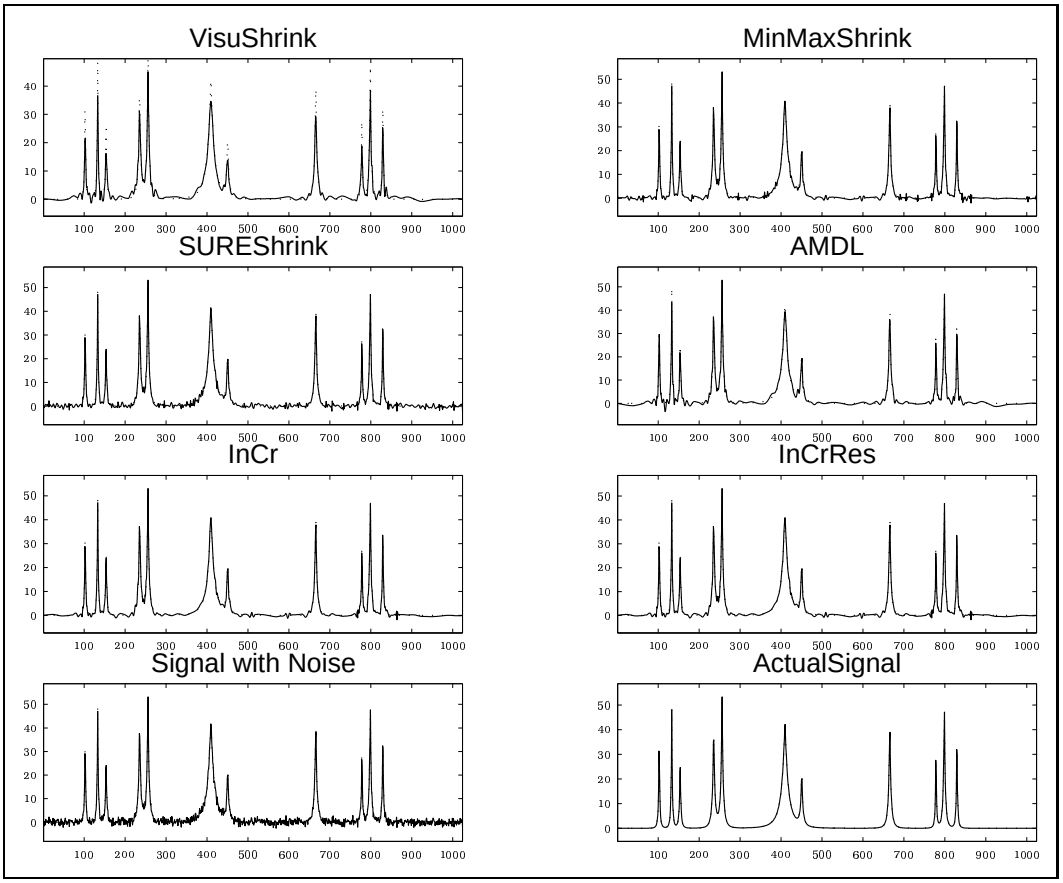

Figure 4.11: Reconstructions of the noisy *bumps* signal.

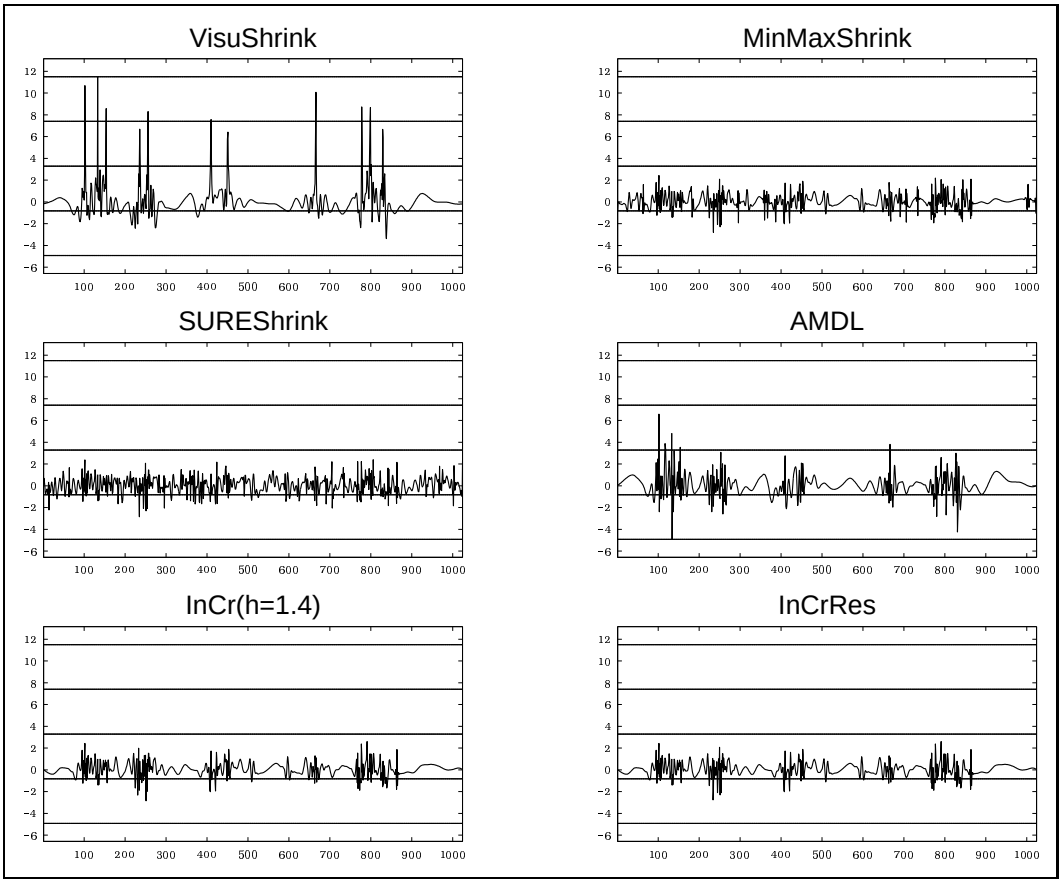

Figure 4.12: Residuals of the *bumps* signal.

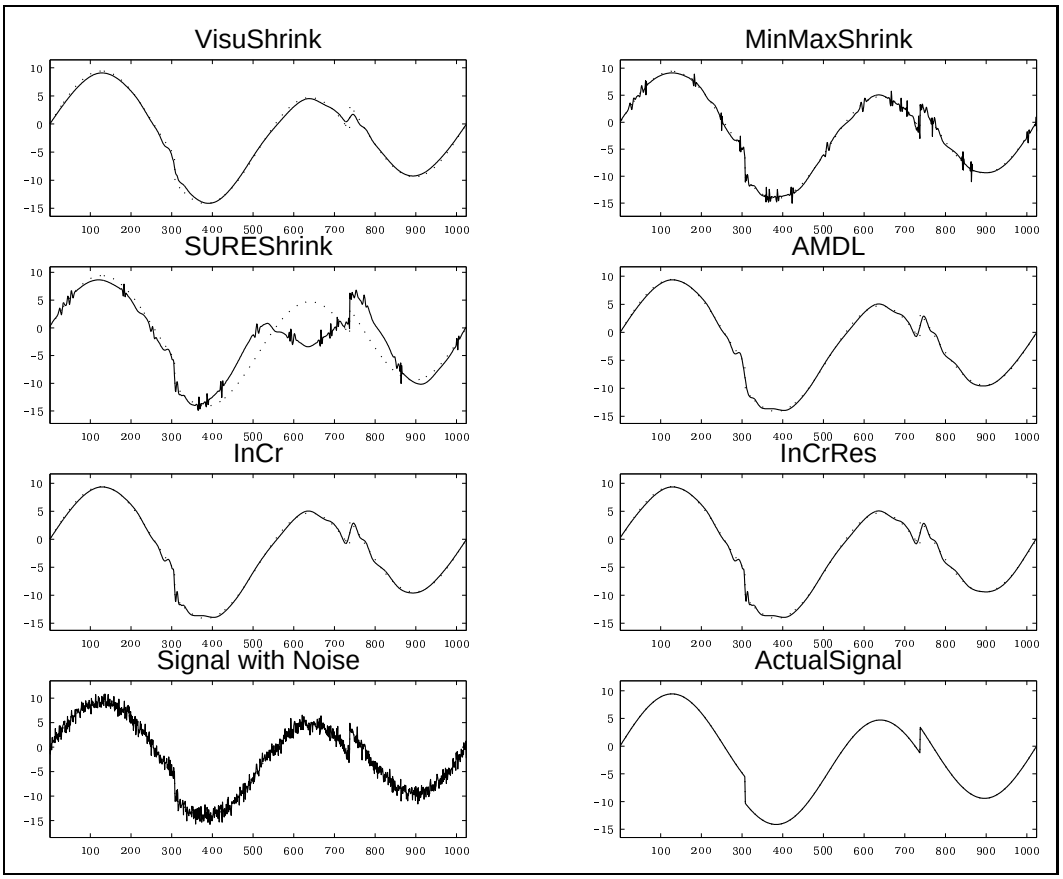

Figure 4.13: Reconstructions of the noisy *heavisine* signal.

residuals found in the last two methods, the methods proposed in this work, are consistently smaller than those found in the other reconstructions. This is yet another evidence that our methods perform better than the de-noising methods found in the literature recently.

Noisy *heavisine* signal. Table 4.9 shows the results obtained by the various methods applied to the noisy *heavisine* signal and Figure 4.13 displays

Table 4.9: Results for the noisy *heavisine* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	17	0.065253	98%	4273.5452
2	$RiskShrink$	54	0.063261	95%	4782.7366
3	$SureShrink$	56	0.409065	95%	7570.3114
4	$AMDL$	16	0.045724	98%	3733.1199
5	$InCr(k, h = 1.3)$	17	0.043949	98%	3689.6285
6	$InCr(k, h = 1.3\|f)+$ $InCr(k, h = 0.5\|Res)$	18	0.043457	98%	3688.0191

Table 4.10: Results for the noisy *blocks* signal.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	100	0.142343	90%	6988.0807
2	$RiskShrink$	164	0.064398	84%	6776.3336
3	$SureShrink$	297	0.084699	71%	9176.1621
4	$AMDL$	77	0.099412	92%	6112.7748
5	$InCr(k, h = 1.5)$	115	0.064243	89%	6037.7806
6	$InCr(k, h = 0.5\|f)+$ $InCr(k, h = 0.45\|Res)$	111	0.068485	89%	6072.2417

the reconstructions of the noisy *heavisine* signal.

All methods but $SureShrink$ perform well on that relatively smooth signal. What might be worth pointing out is that while methods 1 and 5 retain the same number of coefficients, the latter method seems to choose them more appropriately; an observation which is supported by the smaller relative error.

Noisy *blocks* signal. Table 4.10 shows the results obtained by the various methods applied to the noisy *blocks* signal and Figure 4.14 displays the reconstructions of the noisy *blocks* signal.

As mentioned earlier, this signal is difficult to manage when using a rather

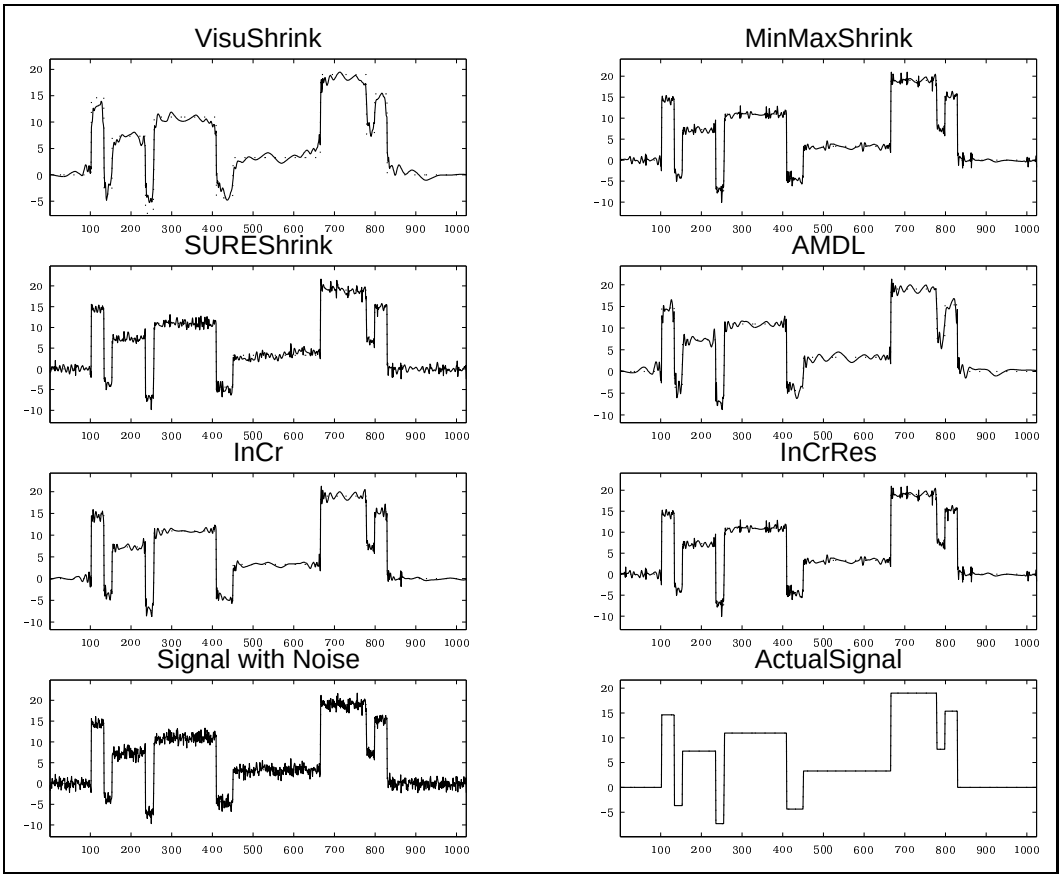

Figure 4.14: Reconstructions of the noisy *blocks* signal.

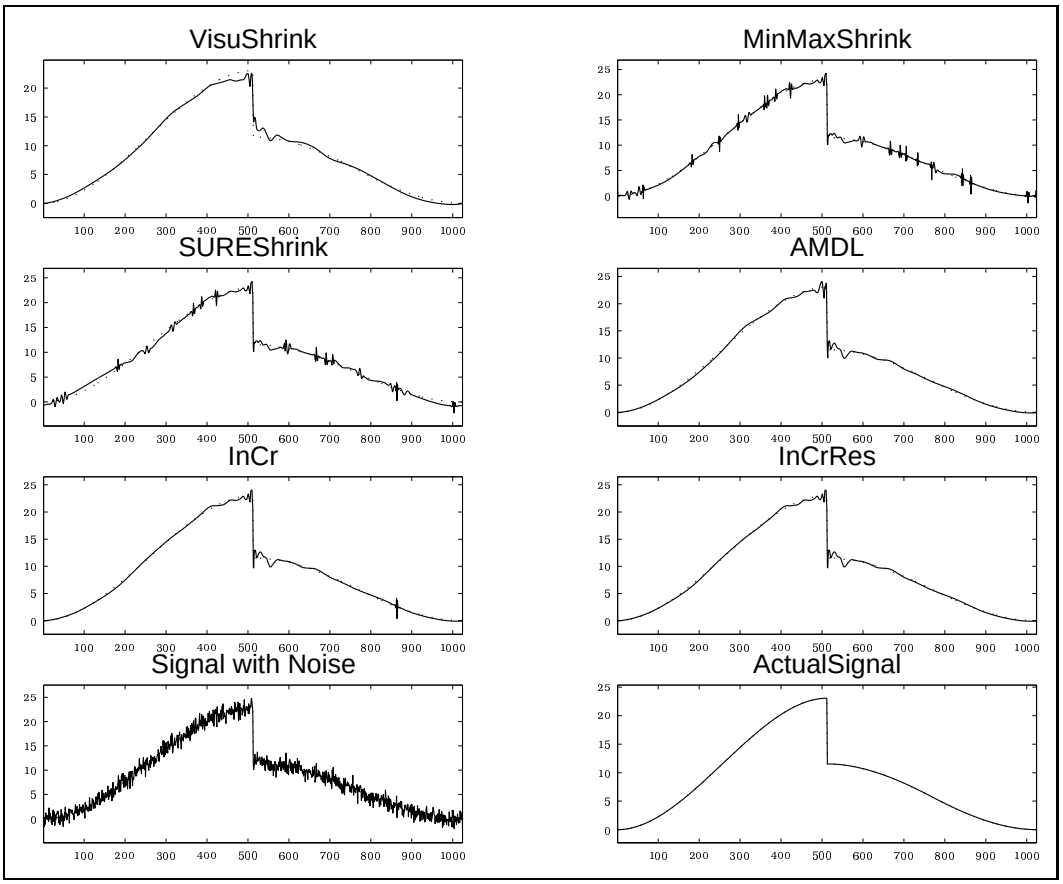

Figure 4.15: Reconstructions of the noisy Nason's function.

smooth wavelet basis. The large number of coefficients retained for the reconstruction by all methods seems to support that. Yet if an importance is placed on close approximation, the last two methods outperform the rest of the methods which is seen in the fact that their estimators produce smaller relative error.

Table 4.11: Results for the noisy Nason's function.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	22	0.050565	98%	4599.0887
2	$RiskShrink$	51	0.038897	95%	4646.5309
3	$SureShrink$	50	0.052958	95%	5087.3895
4	$AMDL$	19	0.032057	98%	3880.8099
5	$InCr(k, h = 1.5)$	23	0.027732	98%	3726.6899
6	$InCr(k, h = 1.4\|f)+$ $InCr(k, h = 0.5\|Res)$	22	0.026500	98%	3644.5506

Noisy Nason's function. Table 4.11 shows the results obtained by the various methods applied to the noisy Nason's function and Figure 4.15 displays the reconstructions of the noisy Nason's function.

One more time, all methods produce fairly good results in compressing and de-noising the data, but judging by the smaller relative error found in the estimates obtained by the last three methods, these three methods appear more adaptive to the original signal while leaving out the coefficients which carry the noise.

Noisy *polysine* function. Table 4.12 shows the results obtained by the various methods applied to the noisy *polysine* function and Figure 4.16 displays the reconstructions of the noisy *polysine* function.

Method $SureShrink$ although eliminating a sufficient number of coefficients appears to choose them inappropriately. The rest of the methods are very compatible with methods 4 and 5 showing identical results. This is a smooth function and it seems this fact makes it easier to capture the smooth shape while eliminating the noise.

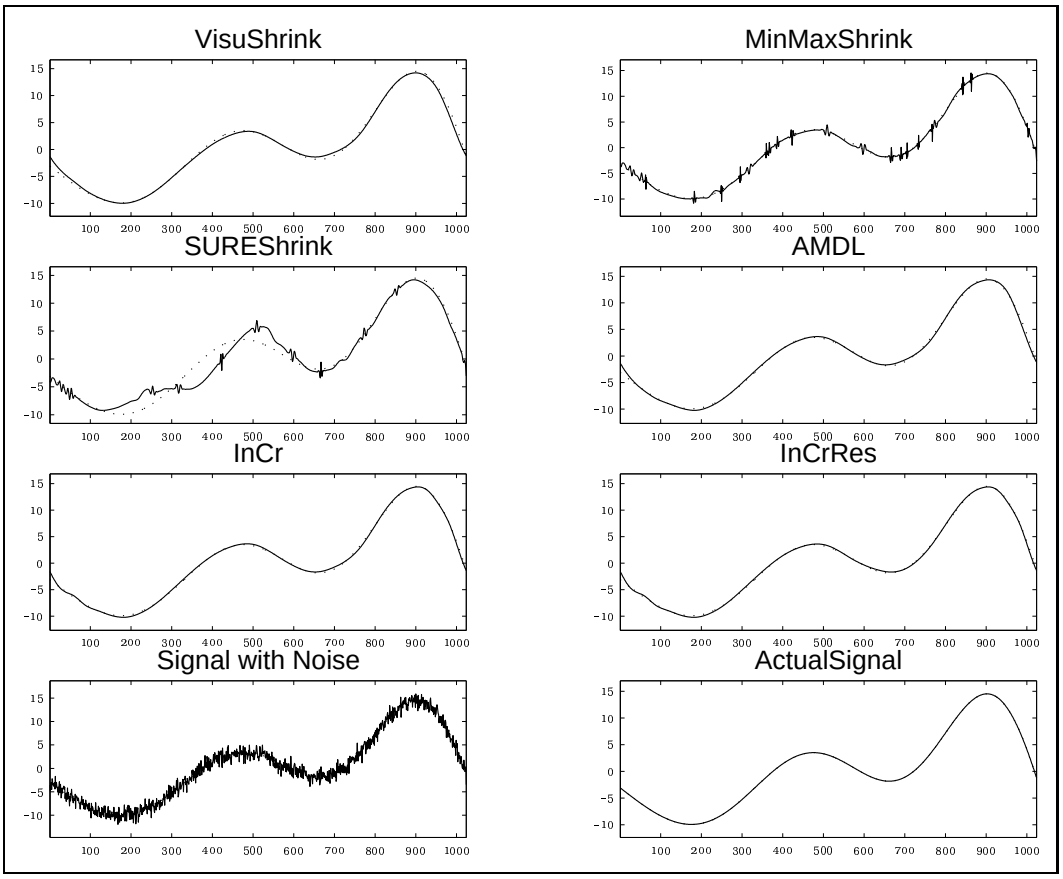

Figure 4.16: Reconstructions of the noisy *polysine* function.

Table 4.12: Results for the noisy *polysine* function.

#	**METHOD**	**k**	**RelErr**	**CR**	**ADML**
1	$VisuShrink$	11	0.060553	99%	4018.2442
2	$RiskShrink$	45	0.057790	96%	4459.2419
3	$SureShrink$	46	0.227447	96%	6498.3188
4	$AMDL$	9	0.042341	99%	3459.7209
5	$InCr(k, h = 1.3)$	11	0.032140	99%	3082.4844
6	$InCr(k, h = 1.3\|f)+$ $InCr(k, h = 0.3\|Res)$	12	0.029183	99%	2954.8678

The problem with de-noising appears to be tightly connected to the data reduction issue, at least in the context of the discrete wavelet transform. In essence, de-noising is achieved by setting to zero wavelet coefficients which represent the noise. The fact that all coefficients are affected by noise makes the problem even more difficult. Naturally, the danger of either eliminating too many coefficients or failing to clear enough coefficients bringing in the noise, is always in waiting. In addition, the nature of the wavelet transform seems to be setting its own limitations on the gains simultaneously in good fit and de-noising, that is if we set the threshold too low we allow noise, if we set it too high, we hurt the quality of the fit. The somewhat better results produced by the last two methods are due to the fact that we can choose the parameter h appropriately to achieve an optimal balance between the values of the relative error and the compression ratio.

4.2 Replicated data

Run-to run (R2R) control is the term related to the integration of classical statistical process control and engineering process control techniques used for modeling and control of processes in the semiconductor industry. The term run-to-run is typically used to describe the case where the control action is made for each batch of wafers at a particular process step. Control actions that take place during a process run are typically made by classical proportional-integral-derivative controllers. As a result, a run-to-run con-

troller is used in a supervisory manner to adjust the process set-points of the automatic controllers so that response variables of interest are maintained within process limits or specifications.

The main reason for utilizing a run-to-run controller is to compensate for time-dependent drifting or disturbance in the process variables as well as modeling error, which can subsequently affect the response variables. For batch sizes as small as one wafer, as in rapid thermal processing (RTP) processing, the importance of accurately modeling and controlling a process to within specifications is especially important. In addition, if an uncontrollable system disturbance occurs (i.e. an equipment fault), it becomes vital to detect such occurrences so that further wafers are not processed and preventive maintenance may be scheduled.

There are usually two aspects to fault detection, particularly, when it concerns processes with a large number of controlled limits and observations. The first is the need to establish, in a form of a template or a baseline prototype, the typical pattern exhibit by the processes which comply with the process specifications. To this prototype then are the new runs to be compared. In this phase, it is natural to think about data reduction so the comparison is quick and informative. Other concerns to be mentioned are regarding the construction of the baseline prototype. On the one hand, it should contain enough detail about the particular pattern of measurements of the manufacturing process, so that it captures and controls all the specified limits and restrictions. On the other hand, we should be cautious against

including an unnecessary level of detailed modeling, so that acceptable variation due to natural process fluctuations and material variability, allows for the particular runs to remain within the acceptable set of criteria. This could be easily translated into exactly the set of reasoning reflected in the characteristics of the proposed $InCr(k, h)$ criterion. As mentioned earlier, it balances precision and economical representation of the signal, and allows for finding an optimal trade off between the two when faced with a particular set of replicated data. Another aspect concerns the detection of difference between the so established prototype and the new runs or replicates. At this stage statistical tests for establishing significant differences ought to be considered.

4.2.1 Real-life examples.

In this section we compare the performance of the methods considered in the earlier comparative study, when applied to real-life data sets. Figure 4.17 displays the plots of the signals for which we are comparing the different methods.

Antenna **data.** With the increasing popularity of wireless communications, demand for antenna equipment used to send and receive communications signals is growing rapidly. Because of the technological sophistication required by new antennae, a high degree of quality is needed during the production process. This data is collected at Nortel production facility in Research Tri-

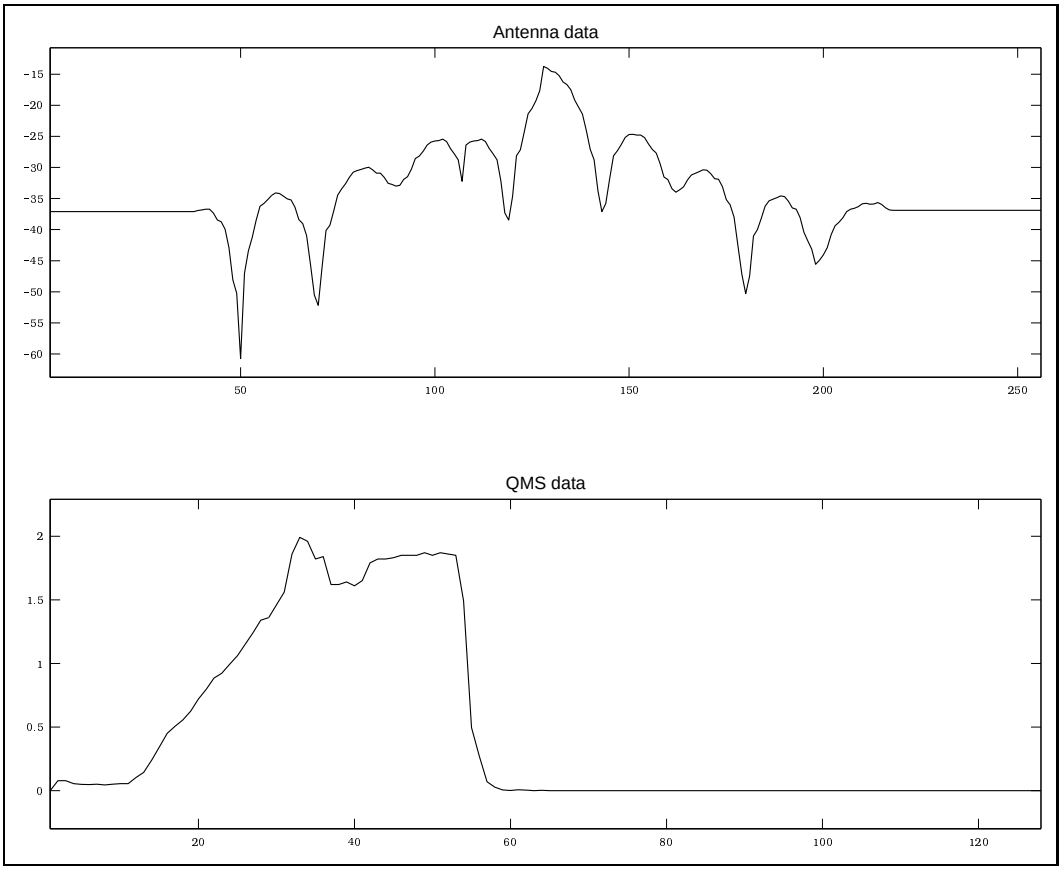

Figure 4.17: Real-life signals.

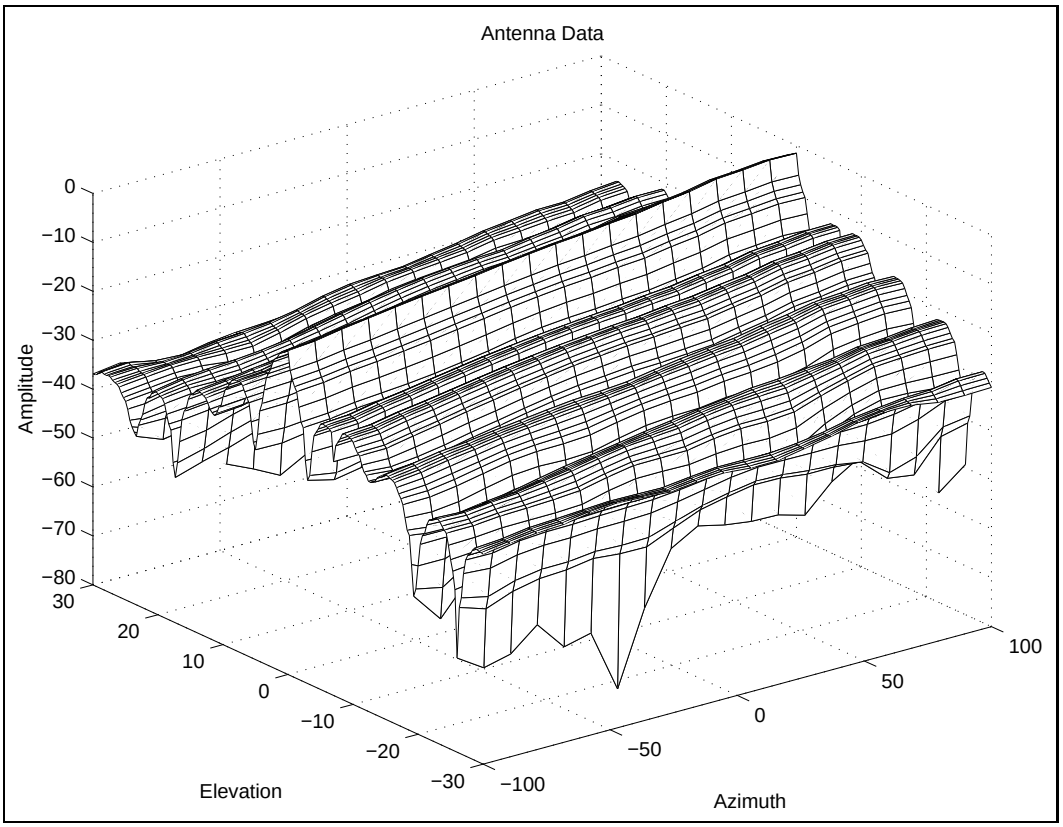

Figure 4.18: Typical antenna signal pattern.

angle Park, with the goal of developing procedures to detect process problems. The Nortel testing equipment received antenna signals at different levels of elevation and degrees of azimuth as presented in Figure 4.18. Antenna data has numerous "peaks" and "valleys", displaying rather irregular patterns, which present difficulties when modeled by the standard statistical procedures, such as regression or splines.

The data available is measured at 181 azimuth values ranging from $-\pi/2$ to $\pi/2$ radians and at 181 elevation values ranging from $-\pi/6$ to $\pi/6$ radians with equal size grid. That is we have 181x181 = 32761 observations.

The antenna quality is evaluated by various regulations regarding the signal pattern. For example, there are certain specification limits on the peaks and on the difference between the peaks and their corresponding valleys. The three main lobes in the center are the most important because they encompass the situations found most frequently in normal usage. The complexity of the data and the further restrictions just mentioned, allow for no simple way to monitor the quality of the antenna production.

Table 4.13 shows the results obtained by the various methods applied to the *antenna* signal and Figure 4.19 displays the *antenna* signal with its wavelet estimates.

All methods, except $VisuShrink$, reduce the data well, using only a part of the coefficients, more than 70% reduction, while capturing the main features of the signal. We can observe from Figure 4.19 that the first method, $VisuShrink$, over-smooths the signal and misses over one of the lobes which

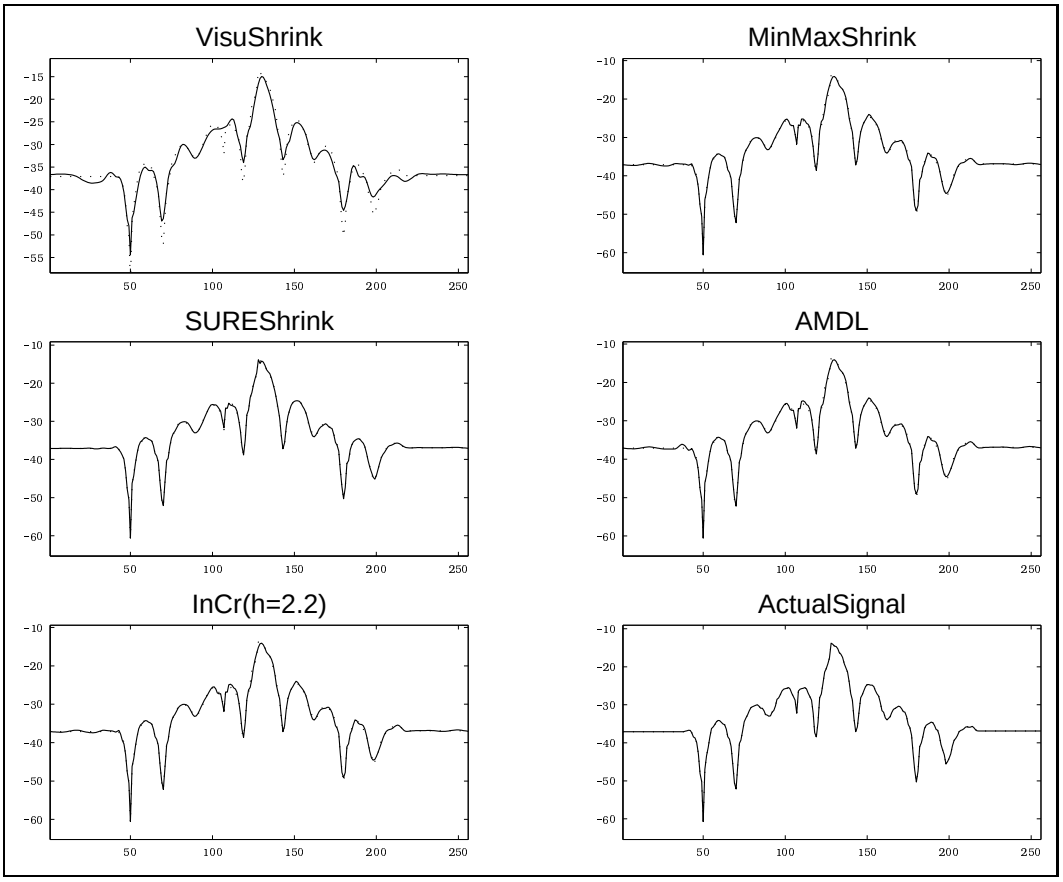

Figure 4.19: Reconstructions of the *antenna* data.(dotted line is the original signal)

Table 4.13: Results for the *antenna* signal.

#	**METHOD**	**k**	**RelErr**	**CR**	**ADML**
1	$VisuShrink$	39	0.041709	85%	1630.2895
2	$RiskShrink$	56	0.011841	78%	1369.2564
3	$SureShrink$	75	0.007764	71%	1441.3503
4	$AMDL$	54	0.012601	79%	1368.2232
5	$InCr(k, h = 2.2)$	55	0.012226	79%	1369.0658

characterizes the nature of the data. It provides the most economical of wavelet estimates but its closeness to the original signal suffers visible discrepancies. This is confirmed from the results in Table 4.13, where the value **RelErr** is the highest among all other entrees. The method *SureShrink*, as its usual performance shows, leans towards higher precision at the cost of including too many coefficients in its wavelet estimate. The value of its **RelErr** is the smallest while the number of coefficients it selects to retain for the wavelet estimate, is the largest. The notions of level of precision and reduction are both undefined and depend on the purpose of the desired estimate and the type of the data analyzed. For example, later on when we talk about a particular implementation of the analysis of the reduced signal, we offer a discussion about the particular set of criteria on what might be viewed as appropriate level of reduction. The last method in the table holds the advantage that it could be fine tuned to the needs of additional processing for further decision making.

QMS **data.** The data represent quadrupole mass spectrometry (QMS) samples of a rapid thermal chemical vapor deposition (RTCVD) process of thin film on a silicon substrate. The RTCVD process deposits thin films on the wafer by a temperature driven surface chemical reaction. As feature size decreases, the operation of the functional parts, like transistors, becomes increasingly susceptible to variations of the other factors in the process. This makes the task of controlling the variability of the process crucial. These

Table 4.14: Results for the QMS data.

#	METHOD	k	RelErr	CR	ADML
1	$VisuShrink$	5	0.478515	96%	333.7591
2	$RiskShrink$	8	0.155629	94%	157.8408
3	$SureShrink$	8	0.155629	94%	157.8408
4	$AMDL$	18	0.037784	86%	1.4320
5	$InCr(k, h = 2.3)$	36	0.012137	72%	-19.2836

particular QMS data was collected from a series of process and diagnostic experiments involving polycrystalline silicon deposition from 10% SiH_4/Ar in a single RTCVD tool ranging in temperature from 625° to 725°. The data consist of 21 nominal RTCVD process runs. For more in-depth description of the data and the process that generates them see Rying et al. [1997].

Table 4.14 shows the results obtained by the various methods applied to the QMS data and Figure 4.20 displays the QMS data with its wavelet estimates.

This signal shows particularly well how sometimes the $VisuShrink$, $Riskshrink$, and $SureShrink$ are unreasonably aggressive in eliminating a great deal of wavelet coefficients. This inconsistency of the performance, lessens the value of being an automatic selection. Our method selects this time the highest number of coefficients used for the reconstruction, and the reason for that is that we need to capture the specific way in which the data changes at the elevated part of the signal. We include this many of the wavelet coefficients because we want to be able to recover the basic pattern of the original signal without jeopardizing the level of data reduction. This

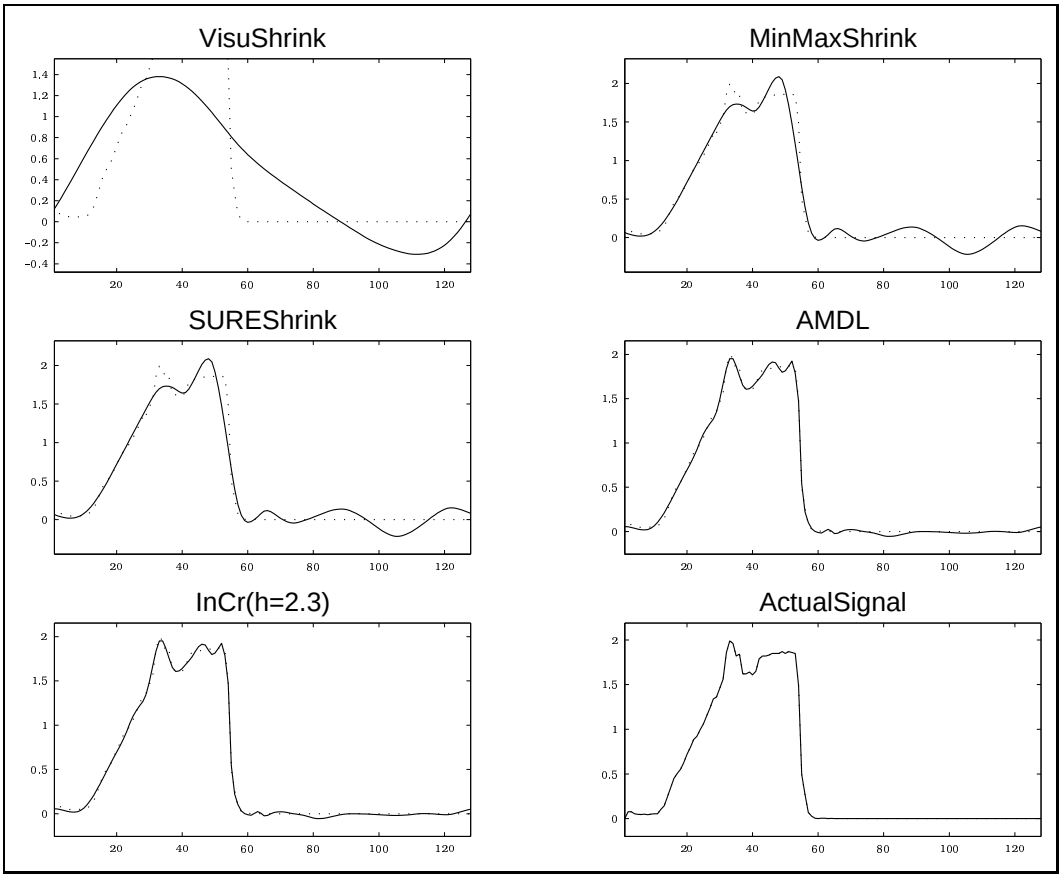

Figure 4.20: Reconstructions of the QMS data.(dotted line is the original signal)

level of data reduction allows for further adjustments to accommodate the needs of a secondary data processing and model selection.

We, next, propose an algorithm which allows for a quick check, after each run, whether it conforms to the accepted general pattern, or it deviates over and above the process limits established for this particular case.

4.2.2 An extension.

Here we propose an algorithm for fault detection applicable to replicated data which displays a clear and distinct pattern.

Step 1. Select from the replicated data, a number of runs based on which we will create the template. Understandably, this is a crucial point and it depends greatly on the control limit requirements for the particular process.

Step 2. For the runs chosen at Step 1, find their wavelet estimates based on a the $InCr(k, h)$ criterion. This step depends on the particular restrictions for the number of coefficients selected for retention, as well as the level of desired precision of the wavelet estimate.

Step 3. At this step the template positions are located. From the selected in Step 1 replicates, we identify the positions of the non-zero wavelet coefficients found at Step 2, and include all of them for the template positions.

Step 4. For the fault detection step, we find the wavelet estimate of the new replicate, and keep only these coefficients from this wavelet estimate which fall in the positions of the template established at Step 3.

Step 5. Next, we find the value of the test statistic

$$\frac{\|\mathbf{w} - \hat{\mathbf{w}}(K)\|^2}{\hat{\epsilon}^2}$$

and compare it to the quantiles of a χ^2_{N-K} distribution. Here $\mathbf{w}$ is the initial wavelet estimate obtained from the DWT, and $\hat{\mathbf{w}}(K)$ is the modified wavelet estimate in which only the elements with positions identified by the template are retained. If the statistic is significant we label the process as faulty.

A justification for the algorithm. First, we assume that each run can be modeled as:

$$Y_i(t_j) = f(t_j) + \sigma z_i(t_j), \quad t_j = \frac{j}{N}\ , j = 1, 2, \ldots N\ ;\ i = 1, 2, \ldots, r\ , \tag{4.1}$$

where $\{z_i(t_j)\}_{i=1}^{i=r}$ are iid, standard normal random variables, r is the number of runs in the data. This assumptions are equivalent to a model expressed in terms of the coefficients from the DWT. That is we believe that the cor-

responding discrete wavelet coefficients come from:

$$w_i(t_j) = \theta_j + \epsilon z_i(t_j), \quad t_j = \frac{j}{N}\,, j = 1, 2, \ldots, N\,;\; i = 1, 2, \ldots, r\,, \quad (4.2)$$

where $\{z_i(t_j)\}_{i=1}^{i=r}$ are iid, $N(0,1)$. From here we consider a null hypothesis which states that

$$H_0 : \theta_{(1)}, \ldots, \theta_{(K)} \quad \text{are not zero, and} \quad \theta_{(K+1)}, \ldots, \theta_{(N)} \quad \text{are zero.}$$

Here $\theta_{(1)}, \ldots, \theta_{(N)}$ are the wavelet coefficients set in a descending order, based on their absolute values. That is, the assumption is that there is a certain number of true wavelet coefficients which are significantly different from 0, and suppose these are the coefficients that are identified in the template. This explains why we are considering the largest in absolute values wavelet coefficients to form the template. Under the assumption of the null hypothesis, the remaining part of the wavelet transform of the signal will be white noise. From here we conclude that the squared approximation error $\|\mathbf{w}(k) - \hat{\mathbf{w}}(K)\|^2/\epsilon^2$ will be, under H_0, simply a sum of $(N-K)$ independent χ_1^2 random variables, and we can use it as a statistic for the test against the alternative hypothesis that more than the first K true wavelet coefficients are not zero.

We use the algorithm proposed above to detect faulty processes.

Antenna **data** For the use of the algorithm we sampled the available data at every 3th azimuth point at the grid, taking 20 samples total. First, we set the parameter value of $h = 1.9$. We made the selection of the value of the parameter h based on the visual closeness of the reconstructed signal to the original. We then look at the signals in the sample of replicates. We notice runs which represent extremes close to the accepted limits, as well as, runs which appear typical. We chose a number of these replicates to help with the identification of the template positions. In this fashion we selected Run 1, Run 3, Run 8, Run 12, Run 17, and Run 19. We then calculated the wavelet estimate for each of the selected signals. To identify the template positions, we selected all positions of the non-zero elements in all of the wavelet estimates of the selected runs. The number of coefficients thus chosen for the template was 54.

In the next step we calculated the discrete wavelet transform (DWT) for each signal in the sample. From the wavelet transform we obtained a new wavelet estimate following two steps: first, we identified the wavelet coefficients which lied in the template positions and left them unchanged; second, set to zero all the wavelet coefficients which were not located at the positions identified by the template positions. This way we formed the final wavelet estimate for each replicate. Figure 4.21 displays four of the replicates together with their approximations based on the wavelet shrinkage chosen as explained above. After that we calculated the estimates for the mean square error (MSE) $\|\mathbf{w} - \hat{\mathbf{w}}(K)\|^2$.

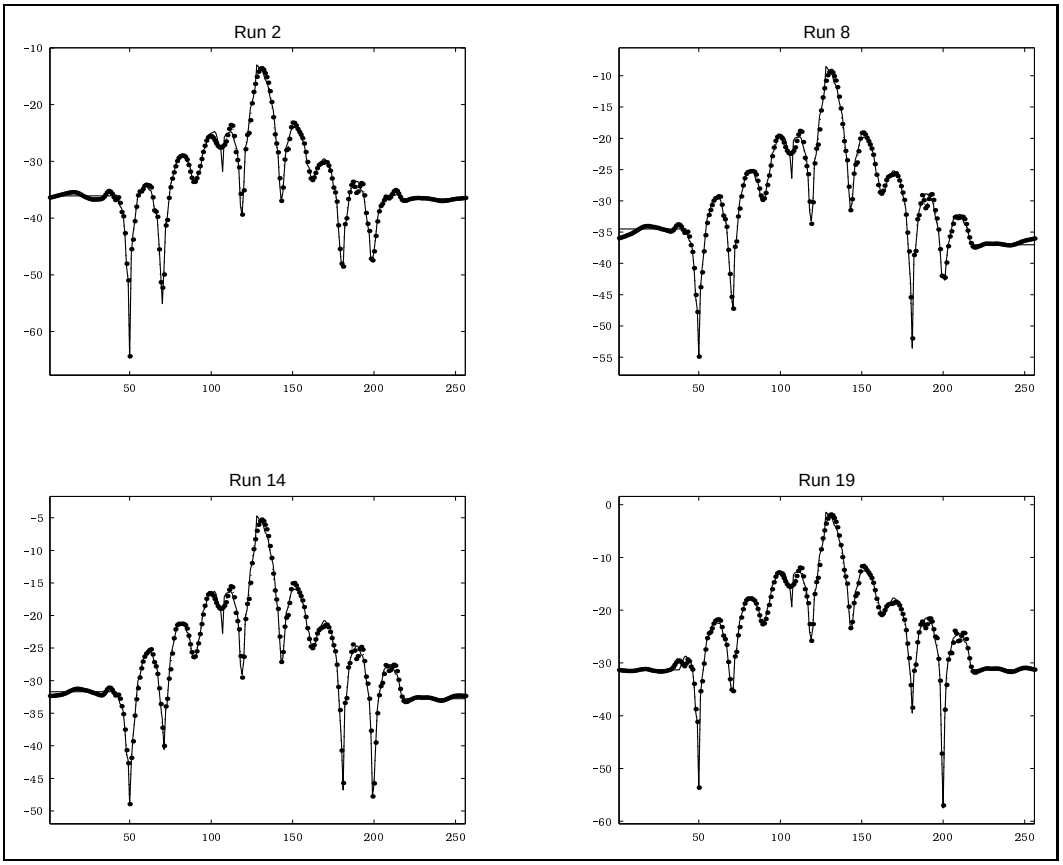

Figure 4.21: Reconstructed *antenna* signals. (line – data points, dots – wavelet approximation)

Table 4.15: Test results for the four runs, *antenna* data.

#	**Run**	**Test Statistics**	**P-value**
1	Run 1	141.4849	0.9996
2	Run 7	161.0523	0.9846
3	Run 13	146.8164	0.9987
4	Run 19	180.0797	0.8642

Before calculating the test statistics, we had to find an estimate for the variance of the noise. For that purpose we first, found the standard deviation of the vectors of wavelet coefficients which occupied the same position across all replicates. Thus we obtained a vector of standard deviations of length the length of the original signal. Then we took the median of the vector of standard deviations, across all positions. The value obtained for the estimate of the variance was $\hat{\epsilon}^2 = 0.7504$. Table 4.15 displays the values of the test statistic $MSE/\hat{\epsilon}^2$ for the four runs, together with the P-values. (The 0.05 cut-off point from χ^2_{202} equals 170.1143.) The interpretation is that all four replicates fall well into the accepted process limits.

QMS **data** We followed the same procedure as with the *antenna* data. First, we decided on a value of the parameter h for the use of the $InCr$ method. The selected value was $h = 1.6$. By inspecting the available data and observing the sets which are extreme or typical we chose Run 3, Run 6, Run 7, Run 11, Run 20 and Run 21 to establish the template positions. Next we found the wavelet shrinkage based on $InCr(k, h)$ for $h = 1.6$ and selected the positions where the non-zero wavelet coefficients resided. The number of coefficients for the template turned out to be $K = 28$. Figure 4.22 displays four of the replicates together with their approximations based on the wavelet shrinkage chosen as explained above. Next, we found an estimate of the variance of the wavelet coefficients in the data set. Once again we first, calculate the estimate of the variance across the replicates and then take the

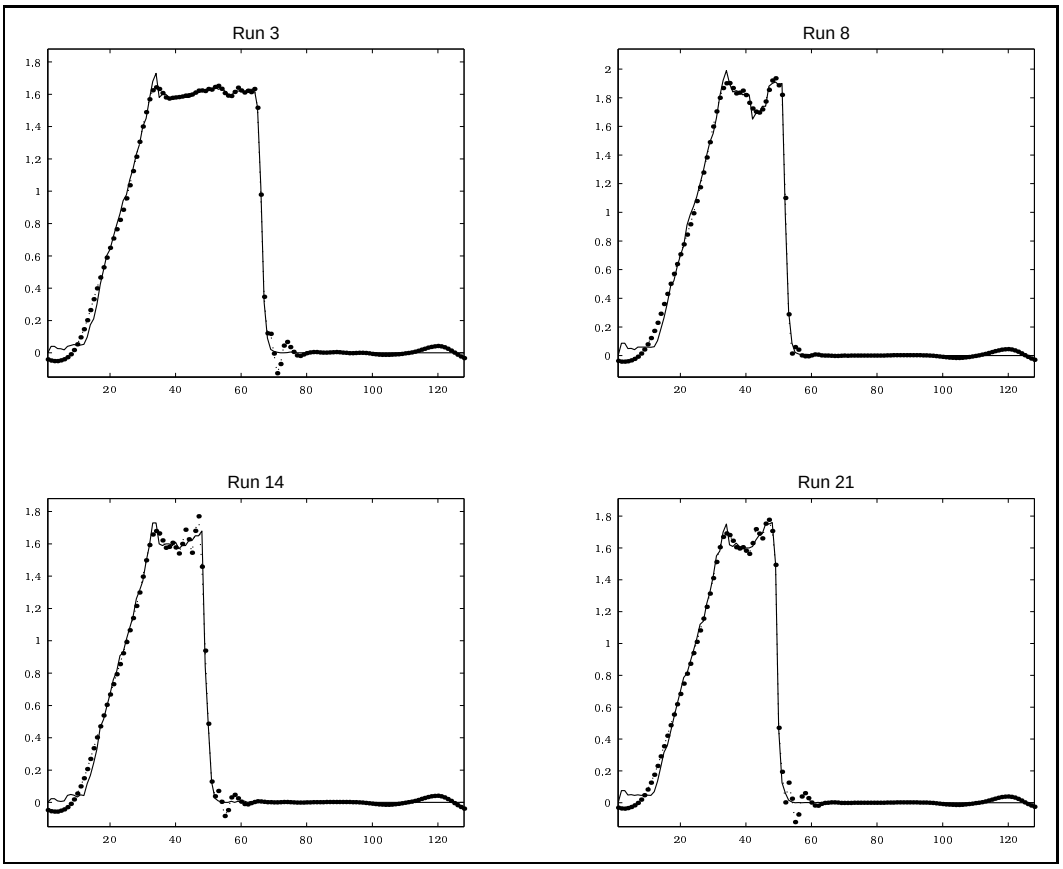

Figure 4.22: QMS data from the template. (line – data points, dots – wavelet approximation)

Table 4.16: Test results for the runs in the QMS data.

#	**Run**	**Test Statistics**	**P-value**
1	Run 3	48.3836	1.0000
2	Run 8	58.4733	0.9997
3	Run 14	62.8649	0.9986
4	Run 21	48.4530	1.0000

median across the observations. Here we operated on the entire set, but left the coarsest resolution level out. The estimate for the variance was thus obtained to be $\hat{\epsilon}^2 = 0.0036$. Table 4.16 displays the values of the test statistic and the corresponding $P-values$. (The 0.05 cut-off point from a χ^2_{100} equals 124.3421.) Once again all the runs are within the process limits.

Overall comments. The advantages for such an algorithm come from the fact that the wavelet basis is good at capturing irregular patterns, on the one hand, and the ability to automatically check the state of a running process, on the other hand. As it is, the algorithm requires, human intervention and decisions only at the first part, while selecting the template positions. Yet this step could be automated as well, when a particular type of processes are concerned. It could be noticed from the tables of the $P-values$ that as the signals get closer to the extreme values, the corresponding $P-values$ decrease as well; that is the test statistic can serve also as a good measure of the distance of the particular run from the assigned control limits.

Chapter 5

Concluding Remarks

5.1 Summary of the thesis results

In this thesis we study the topic of data reduction and model selection for nonparametric regression models based on the discrete wavelet transform (DWT). We consider separately the case when the data reflect a deterministic situation and the case when the signal comes from a stochastic process.

In the deterministic case we assume that the underlying function is Lipschitz α, that is, it possesses certain regularity properties. Based on that assumption we derive an estimate for the approximation error of the wavelet shrinkage estimate obtained by retaining the largest in absolute value K wavelet coefficients. This result mirrors another result found in the literature, which establishes similar bound for the approximation error, under the assumption that the true signal belongs to a certain type of Besov spaces.

The proof of our result uses a lemma which extends a known result establishing the rate of decrease of the coefficients of the wavelet transform, when they come from a function which is Lipschitz α, by asserting that the same bounds hold for the ordered in absolute values wavelet coefficients, as well.

Next we propose a new model selection criterion fashioned after the measure of the average minimum description length (AMDL), on the one hand, and the well known information criteria like Akaike's or C_p, on the other hand. We use this model selection criterion, $(InCr)$, as the expression whose optimization will lead us to identifying a wavelet estimate of desirable properties. The new model selection criterion contains two terms; the first is proportional to the approximation error of the selected wavelet estimate; the second is made of a suitably chosen function. The idea is that the sum of the two terms is dominated by the change in the first term while the largest in magnitude wavelet coefficients are being added to the model and the approximation error descends rapidly. Once the decrease of the approximation error reaches a plateau, that is adding more wavelet coefficients improves the approximation only ever so slightly, the sum of the two terms is now being dominated by the behavior of the second term, that is the value of the increasing function as a function of k, the the number of coefficients already in the model. The function in the second term was chosen so that the minimum of the $InCr$ is achieved around the point where the decrease of the approximation error becomes unnoticeable.

After that we establish a number of properties for the wavelet estimate

obtained by the optimization of the new model selection criterion. First, we show it is equivalent to a hard thresholding procedure. As a consequence we can assert that the $InCr$-wavelet estimate, possesses all the properties of the hard-thresholding wavelet shrinkage, found in the literature. For example, it converges point-wise to the underlying function as the number of observations increases. Next we derive a bound for the number of coefficients selected for inclusion in the $InCr$-wavelet estimate, which shows that when the underlying function is smoother, the number of coefficients included in the model increases slower compared to the case when the function has sharper edges and cusps, as the number of coefficients increases. On the ground that the $InCr$-wavelet estimate is equivalent to an estimate obtained through a hard thresholding procedure, we find a bound for the rate of increase of the value of this threshold, as the number of observations increases.

We consider separately the stochastic case because it imposes a different type of difficulties in approaching the problem. For example, in the deterministic case the more coefficients we include in the reconstruction the more accurate it becomes. When the observations of the data set are contaminated by noise, this is no longer the case; since the high frequency of the signal are severely affected by the noise, as we add more coefficients to the model there comes a point at which the accumulated noise starts influencing the wavelet estimate, that is it begins deviating from the true signal significantly. If we take the mean squared error (MSE) as a measure of the proximity of the estimate to its true function, then we would like to find the point which

minimizes the MSE.

In this context we establish a bound for the MSE for the wavelet estimate obtained by selecting the K largest in absolute value wavelet coefficients. This bound reflects the fact that there is a specific number of wavelet coefficients for which the MSE achieves a minimum. To derive this bound we need an estimate for the largest order statistic from a sample of iid χ_1^2 random variables. We found such estimate in the form of an integral. We use this result to deduce two corollaries. In the first we take an estimate of the integral found in the literature, and in the second we use the fact that when the underlying signal is Lipschitz α we have an estimate for the vector of true wavelet coefficients. Another result concerns the bias of the estimate of the MSE. As mentioned above the nature of the problem in the stochastic case is defined by the fact that we don't know the actual MSE and we use estimates of it instead. One such estimate is the squared norm of the difference between the DWT and the wavelet shrinkage. We obtain a bound for the bias of this estimate.

The latter result allows us to introduce a corrected model selection criterion which is more appropriate for the stochastic case. Once again we are guided by the desire to find the number of coefficients which minimize the MSE. We modify the model selection criterion so that the first term is close to the value it would have, had there been no noise added. Thus we keep the idea of identifying the point at which the improvement of the estimate slows down noticeably as we keep adding new coefficients to the model.

Next we consider some properties of the wavelet estimate selected by the optimization of the corrected model selection criterion. One of the advantages the $InCr$ possesses is that it has a single minimum. We derive an estimate for the expected value of the wavelet coefficient, whose index in the vector of ordered squared wavelet coefficients reflects the value at which this minimum is achieved. As a consequence we establish the fact that for an appropriately selected value of the parameter h, a parameter involved in the definition of the $InCr$, we can find through our method a wavelet estimate which has smaller MSE than the hard thresholding wavelet shrinkage. Finally, we establish an estimate for the value of this parameter h which allows us to find the wavelet estimate which has approximately the same number of coefficients as the number which minimizes the MSE. All these results are found in Chapter 3.

In Chapter 4 we present a comparative study. We take a set of synthesized signals, established as benchmarks in the literature on this topic, and apply to them six methods: $VisuShrink$, $RiskShrink$, $SureShrink$, $AMDL$ and two methods proposed by us which use $InCr$. We consider two separate cases: first, when the signals are considered coming from observations of a deterministic function, and second, when white noise is added to them. In the latter case, we take the noise-to-signal ratio to be 7, the same as other authors have chosen. We establish certain measures for the performance of these methods, like the number of coefficients they select for inclusion, the relative approximation error of the wavelet estimates, and the level of reduction of the initial data sets. Based on these criteria and supplemen-

tal plots and graphs, we give our interpretation on the type of deductions and conclusions that can be drawn. We consider similar comparative study when all the methods listed above are applied to two data sets collected in real-life manufacturing processes. The first data set comes from a Notel production facility, a company located in the Research Triangle Park, North Carolina. This represent a cross-section antenna signals measured at different levels of elevation and degrees of azimuth. The second data set represents a quadrupole mass spectrometry (QMS) sample of a rapid thermal chemical vapor deposition (RTCVD) process of thin film on a silicon substrate. Here we calculate the corresponding measures of the performance of all methods, and conclude that they confirm the expectation that the method based on $InCr$ produces better results.

At the end of Chapter 4 we consider the case of replicated data. This is the situation when we focus on a particular manufacturing process and would like to establish a method for fault detection and process control. First, we propose an algorithm which through the use of the $InCr$ allows for run-to-run control of the particular manufacturing process. The algorithm involves the selection of a template positions, to which the DWT of the replicated data are compared. These template positions are established through the use of the $InCr$-wavelet estimates of a chosen set of these replicates. We finish the chapter with examples for the use and the application of the algorithm. We apply it to the two real-life data sets described above.

5.2 Research Direction

A number of questions still remains unanswered. We will address here few of them.

5.2.1 Varying variance.

It is customary to assume that the random component, that is the noise in the signal is normally distributed, with 0 mean and constant variance. Not rarely this assumption is not fulfilled in reality. It is therefore interesting to investigate the cases when the noise is not normal, or when it comes from different types of distributions, like t- distribution, exponential, normal mixture, etc.. Another way in which this assumption can be wrong is the constancy of the variance across all observations. We are going to discuss the latter case.

In the case when the variance changes throughout the extension of the signal, it might be desirable to develop a method for partitioning of the signal into pieces where the variance is close to a constant. A possible tool to achieve this might be found in the introduction of the *modulus maxima lines* by Mallat [1998] and studied by Jaffard [1991]. By studying the decay of the continuous wavelet transform, that is the modulus maxima lines, one is able to locate all points of sharp signal transitions and singularity. Further more the wavelet modulus maxima define a complete and stable signal representation, that is, algorithms are found which recover signal approximations from

their wavelet maxima. Figure 5.1 displays the maxima lines found for the particular signal and their application.

Thus the plan could be to attempt to use similar techniques to identify the points at which the variance of the noisy signal changes magnitude. Figure 5.2 has the plot of a noisy signal which exhibits great change in the variance and gives a good example for signals which will be candidates for the application of such methods.

5.2.2 Future research.

Minimax property Let us point out that Corollary 6 states that there exist values of the parameter h in the corrected model selection criterion $InCr(k, h)$ for which it performance better than the thresholding method from the point of view of the MSE. Based on this observation, it seems likely that the wavelet shrinkage obtained through the model selection criterion possesses a version of the mimimaxity property, like asymptotically minimax, or adaptively minimax and it might be worthwhile to research this direction.

Replicated data A possible way to improve on the proposed algorithm for fault detection, might be by considering the breaking down of the process to each resolution level. That is instead of taking the largest in absolute value wavelet coefficients overall, we can focus on each resolution level separately, and devise a separate method for approximation specific to the resolution levels. This is a method considered in forming the $SureShrink$ wavelet

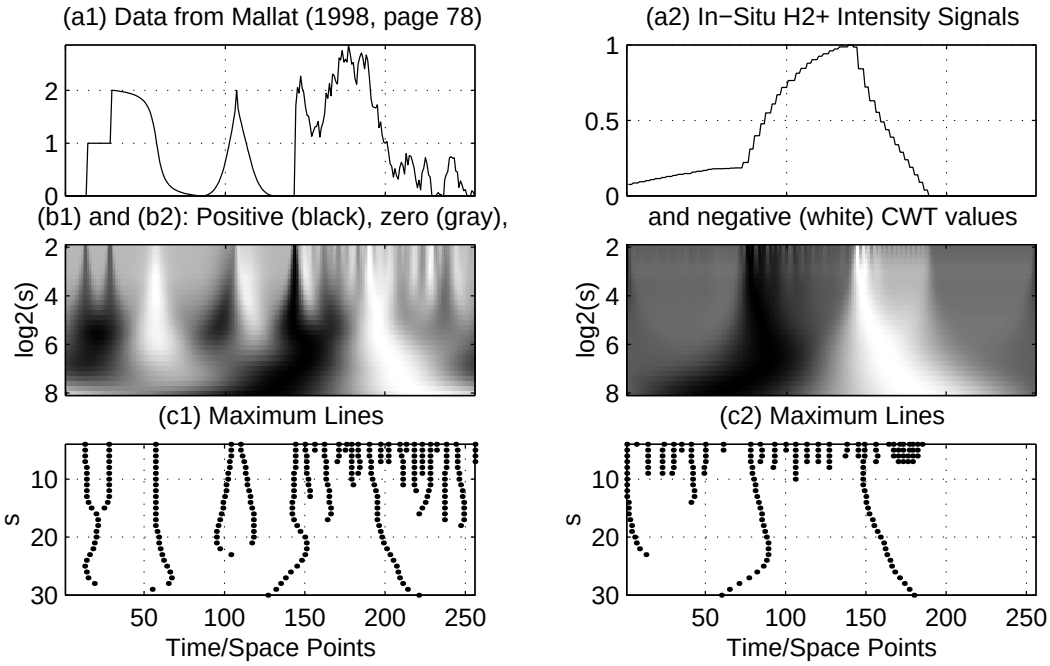

Figure 5.1: (a) Original signal. (b) Wavelet transform $Wf(u,s)$;the vertical and horizontal axes represent $log_2(s)$ and u, respectively; black, grey and white points correspond to positive, zero and negative wavelet coefficients, respectively; singularities create large amplitude coefficients in their cone of influence. (c) Modulus maxima of $Wf(u,s)$;each black point indicates the position of a modulus maximum in the wavelet transform shown in (b).

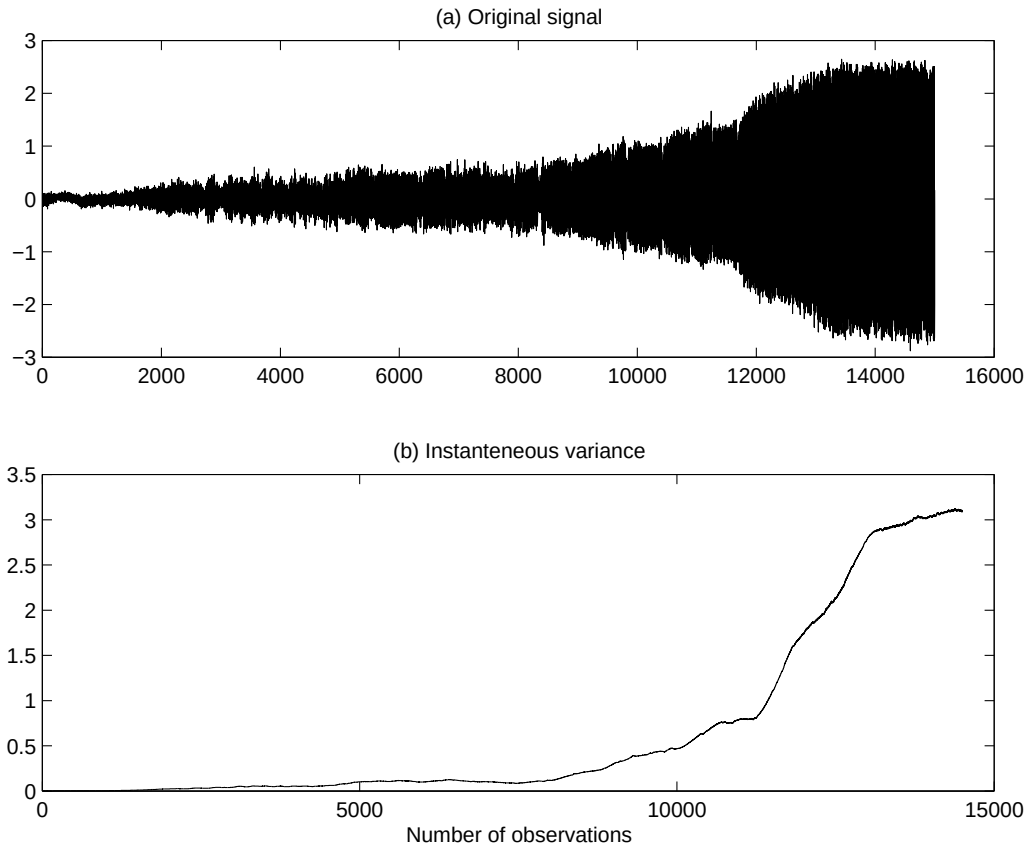

Figure 5.2: A stationary signal with increasing variance.

estimate, as well as, other wavelet shrinkage methods proposed recently.

Data compression and encoding Another immediate application of the methods proposed here could be found in the area of data compression and encoding which is a fast developing area of research in computer science. The subtlety here is the realization that in addition to the method of shrinkage one need to encode the information necessary to replicate the process, that is the values of all involved parameters, number of coefficients retained for reconstruction, etc..

Last comments Wavelet theory has provided statisticians with powerful new techniques for non-parametric inference by combining recent advances in approximation theory with insights gained from applied signal analysis. In this thesis we focus our attention on two aspects of this new development, namely, the automated model selection and data reduction and de-noising. Their application is found in a wide variety of real-life situations and comprise a major part in the study of signal processing. Despite the numerous proposed methods and approaches, an improvement in this direction is still needed and this is where this thesis makes its contribution. Our proposed method for automated model selection is relatively simple to implement, efficient to use, and displays a number of desirable properties. The comparison study suggests that our methods of data reduction and de-noising performs well in relation to the existing methods, while keeping within the

same level of computational complexity. In the world of data, where larger and larger masses of data are being collected daily, requiring timely processing for extracting valuable information, our proposed work is a step up in the development of the overall methodology.

Appendix A

Appendix: Proofs of the results

A.1 Deterministic Case

A.1.1 Proofs of the general results.

Proof of Lemma 1.

Proof: From the notation we know that the coefficient $w^2_{(i)}$ is located at the j -th resolution level and m -th position within it, that is $i = 2^j + m$. We will express this by saying that the index i corresponds to the pair (j, m). Let (j_1, m_1) be the pair that corresponds to the index of the element $w^2_{(i)}$ in the original, unordered vector of coefficients, that is the element $w_{(i)}$ lies in the m_1 position of j_1 resolution level. From Theorem 3 we know that

$$w^2_{(i)} = w^2_{j_1,m_1} \leq \frac{A}{2^{j_1(2\alpha+1)}} .$$

If $j \leq j_1$ then this implies

$$w_{(i)}^2 \leq \frac{A}{2^{j_1(2\alpha+1)}} \leq \frac{A}{2^{j(2\alpha+1)}} .$$

Let's consider the case when $j_1 < j$. That means that the element $w_{(i)}^2$ is in a resolution level which is finer than the resolution level of the original element w_{j_1,m_1}^2. This could happen only if another element, denote it by w_{j_2,m_2}^2, has moved to a resolution level which is coarser than the resolution level j of $w_{(i)}^2$. That is $j < j_2$. Once again by Theorem 3 we obtain that

$$w_{j_2,m_2}^2 \leq \frac{A}{2^{j_2(2\alpha+1)}} .$$

Because the sequence of wavelet coefficients is ordered in a descending order, the fact that w_{j_2,m_2}^2 lies in a coarser level than $w_{(i)}^2$ implies that:

$$w_{(i)}^2 \leq w_{j_2,m_2}^2 \leq \frac{A}{2^{j_2(2\alpha+1)}} .$$

Taking in consideration that $j < j_2$ gives the desired inequality.

□

Proof of Theorem 10.

Proof: We will actually show that

$$\|f - \hat{f}_K\|^2 = \|\mathbf{w} - \hat{\mathbf{w}}(K)\|^2 \leq \frac{A}{K^{2\alpha}} .$$

Let $l = \lfloor \log_2(K) \rfloor$, that is $2^l \leq K < 2^{l+1}$. Using the result of Lemma 1 we have

$$\begin{aligned}
\|\mathbf{w} - \hat{\mathbf{w}}(K)\|^2 &= \sum_{i>K} w_{(i)}^2 \\
&\leq A \left[\frac{2^l}{2^{l(2\alpha+1)}} + \frac{2^{l+1}}{2^{(l+1)(2\alpha+1)}} + \cdots + \frac{2^{l+j}}{2^{(l+j)(2\alpha+1)}} + \cdots \right] \\
&= \frac{A}{(2^{2\alpha})^l} \frac{1}{1 - \frac{1}{2^{2\alpha}}} = \frac{A}{(2^{2\alpha})^l} \frac{2^{2\alpha}}{2^{2\alpha} - 1} \\
&= \frac{A_1}{(2^{2\alpha})^{)}(l+1)} \leq \frac{A_1}{K^{2\alpha}} \, .
\end{aligned}$$

□

A.1.2 Proofs for the new model selection criterion.

Proof of Lemma 2.

Proof: It is enough to point out that in the expression for $InCr(k,h)$

$$InCr(k,h) = \frac{\|\mathbf{w}_N - \hat{\mathbf{w}}(k)\|^2}{\|\mathbf{w}_N\|^2} + \frac{(k+2)ln(k+2) - (k+2)}{h\, N^h}$$

the first part is a decreasing, piece-wise constant function of k while the second part is an increasing function of k.

□

Proof of Proposition 1.

Proof: The hard threshold function selects all wavelet coefficients w_i such that $|w_i| > \lambda$ where the threshold value $\lambda \in (0, \infty)$. If all wavelet coefficients

are assumed different, then this will be equivalent to choosing the model with k largest in absolute value coefficients for any $k = 1, 2, \ldots, N$. If some of the coefficients are identical then they are either both included in the selection or simultaneously excluded. All coefficients of value 0.0 are not considered.

For simplicity, let's assume first, that all wavelet coefficients are different in value and not zero. Need to show that for an appropriate choice of h, the minimum of $InCr(k,h)$ can be achieved at any arbitrary chosen $k = 1, 2, \ldots, N$.

(I) Let us start with the case when $1 < k < N$. Since $InCr(k,h)$ starts out as a decreasing function of k, has a single minimum and then begins increasing (Lemma 2), this minimum will occur at a value k if

$$\begin{aligned} InCr(k-1,h) &> InCr(k,h) \\ InCr(k+1,h) &> InCr(k,h)\,. \end{aligned} \tag{A.1}$$

Equivalently,

$$\begin{aligned} InCr(k,h) - InCr(k-1,h) &< 0 \\ InCr(k+1,h) - InCr(k,h) &> 0\,. \end{aligned}$$

This leads to the following two inequalities:

$$\frac{-w_{(k)}^2}{\|\mathbf{w}_N\|^2} + y(k,h) - y(k-1,h) < 0$$
$$\frac{-w_{(k+1)}^2}{\|\mathbf{w}_N\|^2} + y(k+1,h) - y(k,h) > 0\,.$$

where $y(k,h) = \frac{(k+2)ln(k+2)-(k+2)}{h\,N^h}$. By the mean value theorem we have:

$$\begin{aligned} y(k,h) - y(k-1,h) &= y'(a,h), \quad a \in (k-1,k)\,, \\ y(k+1,h) - y(k,h) &= y'(b,h), \quad b \in (k,k+1)\,. \end{aligned}$$

From here the inequalities (A.1) are equivalent to:

$$y'(a,h) < \frac{w_{(k)}^2}{\|\mathbf{w}_N\|^2} = c_1$$
$$y'(b,h) < \frac{w_{(k+1)}^2}{\|\mathbf{w}_N\|^2} = c_2\,.$$

Now $y'(k,h) = \frac{ln(k+2)}{h\,N^h}$ is an increasing function of k, that is $y'(a,h) < y'(b,h)$ for all values of h since $a < b$. In addition, for a fixed k the function $y'(k,h)$ decreases from ∞ to 0.0 as $h \in (0,\infty)$. Finally, because $w_{(k)}^2 > w_{(k+1)}^2$ this implies that $1 \geq c_1 > c_2 > 0$. Then the

inequalities (A.1) will follow if the following holds:

$$y'(a,h) < c_2 < c_1$$
$$y'(b,h) > c_2\,.$$

Let h_1, h_2 be such that $y'(a,h_1) = c_2\,, y'(b,h_2) = c_2$. Then we know that $h_1 < h_2$ and for $h_1 < h < h_2$ the inequalities (A.1) hold true.

(II) Let us consider now the case when $k = 1$. The minimum of $InCr(k,h)$ will occur at $k = 1$ if

$$InCr(1,h) < InCr(2,h)\,.$$

As above this is equivalent to:

$$y'(a,h) > \frac{w_{(2)}^2}{\|\mathbf{w}_N\|^2}$$

for $a \in (1,2)$. This will hold true for any h small enough, because $\lim_{h\to 0} y'(a,h) = \infty$.

(III) Finally, the case $k = N$, that is $InCr(k,h)$ assumes a minimum at $k = N$. This will be true if

$$InCr(N-1,h) > InCr(N,h)\,.$$

Once again this is equivalent to

$$y'(b,h) < \frac{w_{(N)}^2}{\|\mathbf{w}_N\|^2}$$

for $b \in (N-1, N)$. And this will hold for any h large enough, since $\lim_{h\to\infty} y'(b,h) = 0$.

As indicated earlier, the hard threshold procedure doesn't distinguish between wavelet coefficients which happen to be of equal value and doesn't consider coefficients of value 0.0. $InCr$-estimator has the same behavior.

□

Proof of Theorem 11.

Proof: First, we will show that

$$InCr(k,h) - InCr(k-1,h) \geq -\frac{Ac(N)}{k^{2\alpha+1}} + \frac{k^{s(N)}}{h\,N^h}\,, \tag{A.2}$$

where $\lim_{N\to\infty} c(N) = 1$. From Lemma 1 we know that $w_{(k)}^2 \leq \frac{A}{2^{(l-1)(2\alpha+1)}}$ where $l = \lceil \log_2(k) \rceil$, that is $2^{(l-1)} \leq k < 2^l$. From here we obtain

$$w_{(k)}^2 \leq \frac{A\,2^{(2\alpha+1)}}{l\,(2\alpha+1)} \leq \frac{A_1}{k^{2\alpha+1}}\,.$$

Then

$$\frac{-w_{(k)}^2}{\|\mathbf{w}_N\|^2} > -\frac{A_2 c(N)}{k^{2\alpha+1}}\,, \tag{A.3}$$

where $\lim_{N\to\infty} c(N) = \lim_{N\to\infty} \frac{\|\mathbf{w}\|^2}{\|\mathbf{w}_N\|^2} = 1$. Taking in consideration that

$$\begin{aligned} & InCr(k,h) - InCr(k-1,h) \\ & \quad = \frac{-w_{(k)}^2}{\|\mathbf{w}_N\|^2} + y(k,h) - y(k-1,h) \\ & \quad = \frac{-w_{(k)}^2}{\|\mathbf{w}_N\|^2} + y'(a,h) \end{aligned}$$

where $a \in (k-1,k)$, and the fact that

$$y'(a,h) > y'(k-1,h) = \frac{ln(k+1)}{h\,N^h} \leq \frac{k^{s(N)}}{h\,N^h}$$

where $s(N) = \frac{ln(ln(N+1))}{ln(N)}$, allows us to conclude (A.2).

If k_0 is that value for which the right hand side of (A.2) equals to 0.0, then $K \leq k_0$, where K is the value at which $InCr(k,h)$ achieves its minimum (N is fixed). This is true because

$$InCr(k_0,h) - InCr(k_0-1,h) \geq 0$$

implies that the value k_0 is larger than the values at which $InCr(k,h)$ achieves its minimum. Finally,

$$k_0 = A_2 c(N)\, N^{h/(s(N)+2\alpha+1)}$$

is the root of the RHS of (A.2) and this proves the first part of Theorem 11

since we have shown that

$$K \leq k_0 = A_2 c(N)\, N^{h/(s(N)+2\alpha+1)}\,,$$

where $\lim_{N\to\infty} c(N) = 1$.

To show that the second part of Theorem 11 holds we start by noticing that $InCr(K+1,h) > InCr(K,h)$ since at K the function $InCr(k,h)$ has a minimum. Furthermore to select the model with the K largest in absolute value coefficients, it is enough to choose the value of the hard threshold $T_{InCr} = w^2_{(K+1)}$. Thus we can conclude that

$$\frac{-w^2_{(K+1)}}{\|\mathbf{w}_N\|^2} + y'(b,h) > 0$$

where $b \in (K, K+1)$. From here

$$\begin{aligned} & T_{InCr} \\ & \quad < y'(b,h)\|\mathbf{w}_N\|^2 \\ & \quad \leq y'(K+1,h)\|\mathbf{w}_N\|^2 = \|\mathbf{w}_N\|^2 \frac{ln(K-1)}{h\, N^h} \\ & \quad \leq \|\mathbf{w}_N\|^2 A_2 \frac{h}{s(N)+2\alpha+1} \frac{ln(N)}{h\, N^h} \\ & \quad \leq A_3 \frac{ln(N)}{(s(N)+2\alpha+1)N^h}\,, \end{aligned}$$

since $\|\mathbf{w}_N\|^2 \leq \|\mathbf{w}\|^2$ and the latter norm is a constant independent of N.

□

A.2 Proofs for the stochastic case

A.2.1 Proof of General Results.

Proof of Lemma 3.

Proof: Let $f_Y(t)$ be the density function of the random variable Y, where $Y = |X|$, for X coming from a standard normal, $N(0,1)$ distribution. If $Y_1, Y_2, \ldots, Y_n$ are iid random variables, coming from a distribution with density function $f_Y(t)$, then $Y_1^2, Y_2^2, \ldots, Y_n^2$ are iid χ_1^2 and the mean of the largest order statistic from a χ_1^2 can be found as $E\left[Y_{(n)}^2\right]$. First, note that

$$f_Y(t) = \frac{2}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\,.$$

The distribution function of Y is $F_Y(y) = 2(1 - \Phi(y))$, where $\Phi(y)$ is the distribution function of the standard normal distribution. Let us denote by $U = Y_{(n)}$ largest order statistics. If then $f_U(u)$ denotes the density function of U, we have

$$f_U(u) = \frac{n2^n}{\sqrt{2\pi}}\, e^{-\frac{u^2}{2}}\, [1 - \Phi(u)]^{n-1} = n2^n\, [1 - \Phi(u)]^{n-1}\, \Phi'(u)\,. \qquad \text{(A.4)}$$

From here

$$
\begin{aligned}
& E[U^2] \\
&= n2^n \int_0^\infty u^2 \, [1-\Phi(u)]^{n-1} \, \Phi'(u) \, du \\
&= n2^n \int_0^\infty u^2 \, [1-\Phi(u)]^{n-1} \, d\,\Phi(u) \\
&= -n2^n \int_0^\infty u^2 \, d\frac{[1-\Phi(u)]^n}{n} \\
&= -2^n u^2 \frac{[1-\Phi(u)]^n}{n}\Big|_0^\infty + 2^{n+1} \int_0^\infty u \, [1-\Phi(u)]^n \, du \, .
\end{aligned}
$$

□

Proof of Theorem 12.

Proof: Let $k = K$ for the duration of the proof. From the orhogonality of the DWT we know that the norms are preserved that is:

$$E\|\mathbf{f} - \hat{\mathbf{f}}(K)\|^2 = E\|\theta - \hat{\mathbf{w}}(K)\|^2 \, .$$

From here

$$
\begin{aligned}
& E\|\theta - \hat{\mathbf{w}}(k)\|^2 \\
&= E\sum_{i=1}^{N}\left(\theta_i - w_i I(w_i^2 \geq w_{(k)}^2)\right)^2 \\
&= \sum_{i=1}^{N} E\left((\theta_i - w_i)^2 \mid w_i^2 \geq w_{(k)}^2\right) P(w_i^2 \geq w_{(k)}^2) + \sum_{i=1}^{N} \theta_i^2 P(w_i^2 < w_{(k)}^2) \\
&= \sum_{i=1}^{N} E\left((\epsilon z_i)^2 \mid w_i^2 \geq w_{(k)}^2\right) P(w_i^2 \geq w_{(k)}^2) + \sum_{i=1}^{N} \theta_i^2 P(w_i^2 < w_{(k)}^2)\,.
\end{aligned}
$$

If we recall that $w_{(1)}^2 > w_{(2)}^2 > \cdots > w_{(k)}^2$ and use the Holder inequality we can rewrite this as:

$$
\begin{aligned}
& E\|\theta - \hat{\mathbf{w}}(k)\|^2 \\
&\leq \epsilon^2 \sum_{i=1}^{N} E\left(z_N^2 \mid w_i^2 \geq w_{(k)}^2\right) P(w_i^2 \geq w_{(k)}^2) \\
&+ \left(\sum_{i=1}^{N} \theta_i^4\right)^{1/2} (N-k)^{1/2} \\
&\leq \epsilon^2 k E(z_{(N)}^2) + \left(\sum_{i=1}^{N} \theta_i^4\right)^{1/2} (N-k)^{1/2}\,.
\end{aligned}
$$

Here we use the Jensen inequality for a concave function, that is $E(\sqrt{Y}) \leq$

$\sqrt{E(Y)}$. It follows that

$$E\|\theta - \hat{\mathbf{w}}(k)\|^2$$
$$\leq \quad E(z_{(N)}^2)\epsilon^2 K + \left(\sum_{i=1}^{N} \theta_i^4\right)^{1/2} (N-K)^{1/2} \; .$$

To obtain the desired inequality, we take into account the result from Lemma 3 and the fact that $\epsilon^2 = \sigma^2/N$.

□

Proofs of Lemma 4 and Lemma 5.

Proof: The proof of Lemma 4 follows from Theorem 3. To establish Lemma 5 we use the result from Leadbetter et al. [1983] that if $\{z_i\}$ are iid $N(0,1)$ then

$$Pr\left\{\|\mathbf{z}\| \leq \sqrt{2lnN}\right\} \to 1 \; , \quad N \to \infty \; .$$

This allows us to assert that with high probability $z_{(N)}^2 < 2ln\, N$ and from here $E(z_{(N)}^2) < 2ln\, N$ which establishes the statement.

□

Proof of Theorem 13.

Proof: We proceed as in the the proof of Theorem 12.

$$
\begin{aligned}
&\left| E\|\mathbf{w}_N - \hat{\mathbf{w}}(k)\|^2 - E\|\theta - \hat{\mathbf{w}}(k)\|^2 \right| \\
&= \left| E\sum_{i=1}^{N} \left(w_i - w_i I(w_i^2 \geq w_{(k)}^2)\right)^2 - E\sum_{i=1}^{N} \left(\theta_i - w_i I(w_i^2 \geq w_{(k)}^2)\right)^2 \right| \\
&= \Bigg| \sum_{i=1}^{N} E\left((w_i - w_i)^2 \mid w_i^2 \geq w_{(k)}^2\right) P(w_i^2 \geq w_{(k)}^2) \\
&+ \sum_{i=1}^{N} E(w_i^2 \mid w_i^2 < w_{(k)}^2) P(w_i^2 < w_{(k)}^2) \\
&- \sum_{i=1}^{N} E\left((\theta_i - w_i)^2 \mid w_i^2 \geq w_{(k)}^2\right) P(w_i^2 \geq w_{(k)}^2) \\
&- \sum_{i=1}^{N} \theta_i^2 P(w_i^2 < w_{(k)}^2) \Bigg| \\
&\leq \sum_{i=1}^{N} \left[E(w_i^2 \mid w_i^2 < w_{(k)}^2) - \theta_i^2\right] P(w_i^2 < w_{(k)}^2) + \epsilon^2 k E(z_{(N)}^2) \\
&\leq \sum_{i=1}^{N} \left[E(w_i^2) - \theta_i^2\right] P(w_i^2 < w_{(k)}^2) + \epsilon^2 k E(z_{(N)}^2) \\
&= \epsilon^2(N-k) + \epsilon^2 k G(N)\,.
\end{aligned}
$$

This establishes the result.

□

A.2.2 Proofs for the corrected model selection criterion.

Proof of Lemma 4.

Proof: The proof of Lemma 4 follows from the observation that in the expression (3.6) for $InCr(k,h)$ the first part is a decreasing, piece-wise constant function of k when the condition $w_{(k)}^2 > 2\epsilon^2$ holds, while the second part is an increasing function of k.

□

Proof of Proposition 2.

Proof: We start by observing that the minimum will be achieved at k for which:

$$I(k,h) - I(k-1,h) + y'(a,h) = 0$$

where $y'(k,h) = \frac{ln(k+2)}{h\,N^h}$ and $a \in (k-1,k)$. This is equivalent to:

$$\frac{-w_{(k)}^2 + 2\epsilon^2}{\|w_N\|^2 - N\epsilon^2} + \frac{ln(a+2)}{hN^h} = 0\,.$$

When we solve for $w_{(k)}^2$ we obtain:

$$w_{(k)}^2 = \frac{ln(a+2)}{hN^h}(\|w_N\|^2 - N\epsilon^2) + 2\epsilon^2\,.$$

The proof is complete after we take expectation on both sides.

□

Proof of Corrolary 6.

Proof: As pointed out above there is an optimal number of coefficients for which the mean square error is minimal. Proposition 2 shows that by choosing the value of h appropriately we can ahieve this minimum and therefore make the inequality in Corrolary 6 true.

□

Proof of Proposition 3.

Proof: Here we proceed the same way as in the proof of Proposition 2. Thus from

$$I(k+1,h) - I(k,h) + y'(a,h) = 0$$

we derive:

$$\frac{-w_{(k)}^2 + 2\epsilon^2}{\|w_N\|^2 - N\epsilon^2} + \frac{ln(a+2)}{hN^h} = 0\,.$$

From here we obtain that the minimum is achieved for the value of h which satisfies the following equation:

$$hN^h = \frac{\|w_N\|^2 - N\epsilon^2}{w_{(k)}^2 - 2\epsilon^2} ln(a+2)\,,$$

where $a \in (k-1,k)$. When we substitute this value for hN^h back in the formula (3.6) and use the fact that $(ln(k+2)-1))/ln(a+2) \approx 1$, we reach the result stated in the proposition.

□

Bibliography

F. Abramovich and Y. Benjamini. Thresholding of wavelet coefficietns as multiple hypotheses testing procedure. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, pages 5 – 14. Springer-Verlag, New York, 1995.

F. Abramovich and T. Sapatinas. Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference inwavelet Based Models*, New York, 1999. Lecture Notes in Statistics, Springer-Verlag.

H. Akaike. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22:203–217, 1970.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Czàki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiadò.

A. Antoniadis, I. Gijbels, and G. Gregoire. Model selection using wavelet decomposition and applications. *Biometrika*, 84:751–763, 1997.

J. B. Buckheit and D. L. Donoho. Wavelab and reproductive research. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, number 103 in Lecture Notes in Statistics, pages 55–82. Springer-Verlag New York, Inc., 1995.

C. K. Chui. *An Introduction to Wavelets.* Academic Press, San Diego, 1992.

R. R. Coifman and M. W. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions of Information Theory*, 38:713–718, 1992.

I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics*, 41:909–996, 1988.

R. A. DeVore, B. Jawerth, and V. Popov. Compression of wavelet decompositions. *American Journal of Mathematics*, 1992.

D. Donoho. De-noising via soft-thresholding. *IEEE Transactions Information Theory*, 41:613–627, 1995.

D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptotia? (with discussion). *Journal of the Royal Statistical Society*, 57:301–337, 1995.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 1998.

J. Fan. Test of significance based on wavelet thresholding and *Neyman's* truncation. *Journal of the American Statistical Association*, 91:674–688, 1996.

J. Fan, P. Hall, M. Martin, and P. Patil. Adaptation to high spatial inhomogeneity based on wavelets and on local linear smoothing. Technical Report CMA-SR1893, Centre for Mathematics and Its Applications, Australian National University, Canberra, 1993.

A. Grossman and J. Morlet. Decompositions of hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Analysis*, 15(4):723–736, July 1984.

S. Jaffard. Pointwise smoothness, two-microlocalization and wavelet coefficients. *Journal of Applied and Computational Harmonic Analysis*, 1991.

H. Krim and I. C. Schick. Minimax description length for signal denoising and optimized representation. *IEEE Transactions on Information Theory*, 45(3):1–12, 1999.

M. R. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Process.* Springer-Verlag, New York, 1983.

G. H. Lee. *A Statistical Wavelet Approach to Model Selection and Data Driven Neyman Smooth Tests.* PhD thesis, Texas A&M University, College Station, TX, May 1997.

J. C. Lu, W. Zhou, D. Chen, J. M. Hughes-Oliver, and S. K. Ghosh. Process equipment fault detection and classification based on reduced-size data constructed from structured wavelet models. *Technometrics (paper in revision).*

S. G. Mallat. Multiresolutional approximation and wavelet orthonormal bases of $l^2(r)$. *Transactions American Mathematical Society*, 315:69–87, 1989.

S. G. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

Y. Meyer. *Ondelettes et Operateurs I: Ondelettes.* Herman, 1990.

G. P. Nason. Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society*, 58(Series B):463–479, 1996.

R. Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

T. Ogden. *Wavelet Thresholding in Nonparametric Regression with Change-point Applications.* PhD thesis, Texas A&M University, College Station, TX, May 1994.

J. Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 1984.

E. A. Rying, R. S. Gyurcsik, J. C. Lu, G. Bilbro, G. Parsons, and F. Y. Sorell. Wavelet analysis of mass spectrometry signals for transient even detection and run-to-run process control. In M. Meyyappan, D. Economou, and

S. Butler, editors, *Process Control, Diagnostics, and Modeling in Semiconductor Manufacturing*, volume 97-9 of *Electrochemical Society Proceedings*, 1997.

N. Saito. Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. In E.Foufoula-Georgiou and P.Kumar, editors, *Wavelets in Geophysics*, pages 299–324. Academic Press, New York, 1994.

C. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

T. Tao. On the almost everywhere convergence of wavelet summation methods. Technical report, Department of Mathematics, Princeton University, 1991.

S. Verdù and H. V. Poor. On minimax robustness: a general appraoch and applications. *IEEE Transactions on Information Theory*, 30:328–340, 1984.

B. Vidakovic. *Statistical Modeling by Wavelets.* John Wiley & Sons, New York, 1999.

M. V. Wickerhauser. *Adapted Wavelet Analysis: From Theory to Software.* AK Peters, 1994.