Fall 2009

Wavelet regression with long memory infinite moving average errors

Juan Liu
University of New Hampshire, Durham

WAVELET REGRESSION WITH LONG MEMORY INFINITE MOVING AVERAGE ERRORS

BY

JUAN LIU

B.A., Hebei University of Technology, China (2002)

M.S., East China University of Science and Technology, China (2004)

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

in

Mathematics

Sep 2009

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Professor Linyuan Li, for his support and encouragement during my graduate studies at the University of New Hamphsire. Without his guidance, help and kindness, I could not get through some tough times and this thesis could never have been written.

I would like to express my gratitude to Professor Don Hadwin for teaching me mathematics, and for his constant encouragement. I would like to thank Professor Ernst Linder and Professor Phil Ramsey for spending their valuable time with me discussing statistical topics.

I also want to extend my many thanks to Professor Rita Hibschweiler and Professor Mehmet Orhon for their constant support and encouragement. My thanks also go to Professor Eric Grinberg for being such a wonderful chairman.

I am grateful to the Department of Mathematics and Graduate School for providing me such a warm and supportive environment.

TABLE OF CONTENTS

ABSTRACT

Wavelet Regression with Long Memory Infinite Moving Average Errors

by

Juan Liu

University of New Hampshire, Sep 2009

For more than a decade there has been great interest in wavelets and wavelet-based methods. Among the most successful applications of wavelets is nonparametric statistical estimation, following the pioneering work of Donoho and Johnstone (1994,1995) and Donoho et al. (1995). In this thesis, we consider the wavelet-based estimators of the mean regression function with long memory infinite moving average errors, and investigate the rates of convergence of estimators based on thresholding of empirical wavelet coefficients. We show that these estimators achieve nearly optimal minimax convergence rates within a logarithmic term over a large class of non-smooth functions that involve many jump discontinuities, where the number of discontinuities may grow polynomially fast with sample size. Therefore, in the presence of long memory moving average noise, wavelet estimators still achieve nearly optimal convergence rates and demonstrate explicitly the extraordinary local adaptability of this method in handling discontinuities. We illustrate the theory with numerical examples.

A technical result in our development is the establishment of Bernstein-type exponential inequalities for infinite weighted sums of i.i.d. random variables under certain

cumulant or moment assumptions. These large and moderate deviation inequalities may be of independent interest.

Chapter 1

Introduction

The origin of wavelets can be traced back to the beginning of the 20th century; however, wavelets, understood systematically as a way of providing local orthogonal bases, are a recent product of existing theories in various fields and some important research discoveries. The term "wavelet" originates from the work of Morlet et al. (1982), in the context of the analysis of seismetic reflection data. Since then wavelets have led to exciting applications in many areas, such as signal processing, for example Mallat (1989), and image processing, for example Shapior (1993). In the early 1990s, a series of papers by Donoho and Johnstone and their coauthors demonstrated that wavelets are powerful tools in problems of denoising, regression, and density estimation. The subsequent booming wavelet research broadened a large range of statistical problems.

Wavelets provide a framework with some key advantages. Firstly, wavelets are orthonormal basis functions that are localized in both time and frequency, with time-widths adapted to their frequency. This enables their ability to model a signal with high frequency components, such as discontinuities or sharp spikes, in contrast to

more traditional statistical methods for estimating an unknown function. Secondly, fast orthogonal discrete wavelet transformation makes the application of wavelets available. A third advantage is that wavelet coefficients are often sparse and, therefore, representations of functions could be economical. These essential attributes make wavelets an outstanding tool for statistical denoising.

With the introduction of nonlinear wavelet methods in statistics by Donoho and Johnstone (1994, 1995, 1998) and Donoho et al. (1995), the theory and application of wavelet approaches to nonparametric regression has developed rapidly. Many papers have been written on this topic.

1.1 Basic Definitions

We consider nonparametric regression

$$Y_i = g(x_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{1.1.1}$$

where $x_i = i/n \in [0, 1]$, $\varepsilon_1, \cdots, \varepsilon_n$ are observational errors with mean 0 and g is an unknown function to be estimated.

Common assumptions on $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. errors or stationary processes with short-range dependence such as the classic ARMA (Autoregressive Moving Average) processes, of the form:

$$\varepsilon_t = c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2} + \cdots + c_p\varepsilon_{t-p} + w_t + d_1w_{t-1} + \cdots + d_qw_{t-q}, \tag{1.1.2}$$

where ε_t is stationary, $c_1, c_2, \cdots, c_p$ and $d_1, d_2, \cdots, d_q$ are constants, $c_p \neq 0$, $d_q \neq 0$ and w_t is a Gaussian white noise series with mean equal to zero.

Concisely, the ARMA(p,q) model can be written as

$$c(B)\varepsilon_t = d(B)w_t.$$

Define the **autocovariance function ACF** to be as the second moment product

$$r(s, s+j) = r(j) = E[(\varepsilon_s - \mu_s)(\varepsilon_{s+j} - \mu_{s+j})], \tag{1.1.3}$$

and define the **autocorrelation function** to be

$$\rho(s, s+j) = \frac{r(j)}{\sqrt{r(s,s)r(s+j, s+j)}} \tag{1.1.4}$$

See below for two figures (1.1 and 1.2) depicting the autocorrelation functions of an independent and ARMA(1,1) process.

The conventional ARMA process is often referred to as a short memory process. However, in many fields which include economics, geosciences, biology and hydrology, it is unrealistic to assume that the observational errors are independent. Instead, these observational errors exhibit slow decay in correlation which is often referred to as long-range dependence or long memory.

Suppose $\varepsilon_1, \cdots, \varepsilon_n, \cdots$ is a stationary error process with mean 0 and variance 1. Then $\{\varepsilon_i, i \geq 1\}$ is said to have **long-range dependence** or **long memory**, if there exist constants $\alpha \in (0,1)$ and $C_0 > 0$ such that

$$r(j) = E(\varepsilon_1 \varepsilon_{1+j}) \sim C_0 |j|^{-\alpha}, \tag{1.1.5}$$

where $a_j \sim b_j$ means that $a_j/b_j \to 1$ when $j \to \infty$. The literature on long-range dependence is very extensive, see, e.g., Beran (1994), Doukhan, *et al.* (2003) and their combined references.

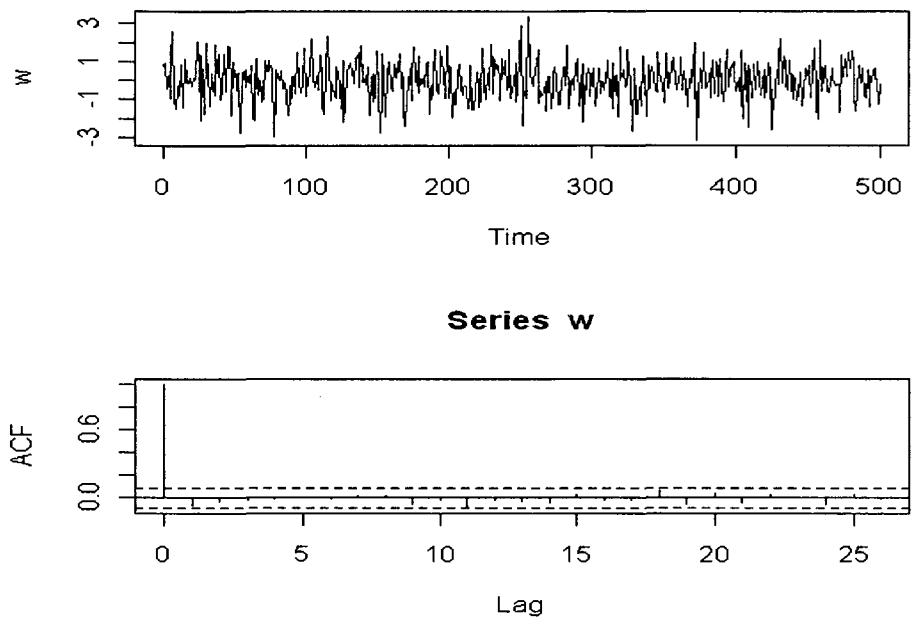

Figure 1.1: *Gaussian white noise series (top) and plot of its sample ACF*

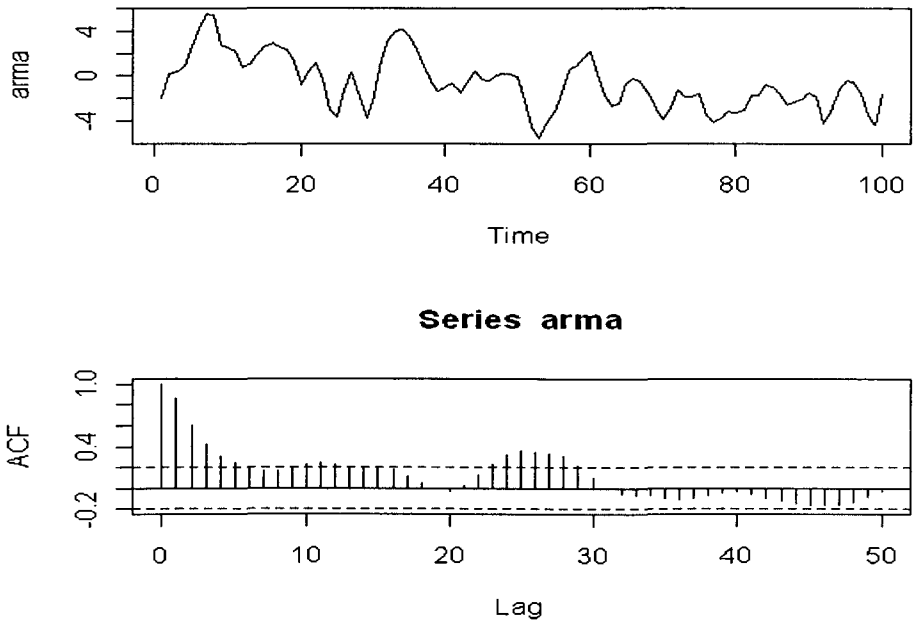

Figure 1.2: *ARMA(1,1) series (top) and plot of its sample ACF*

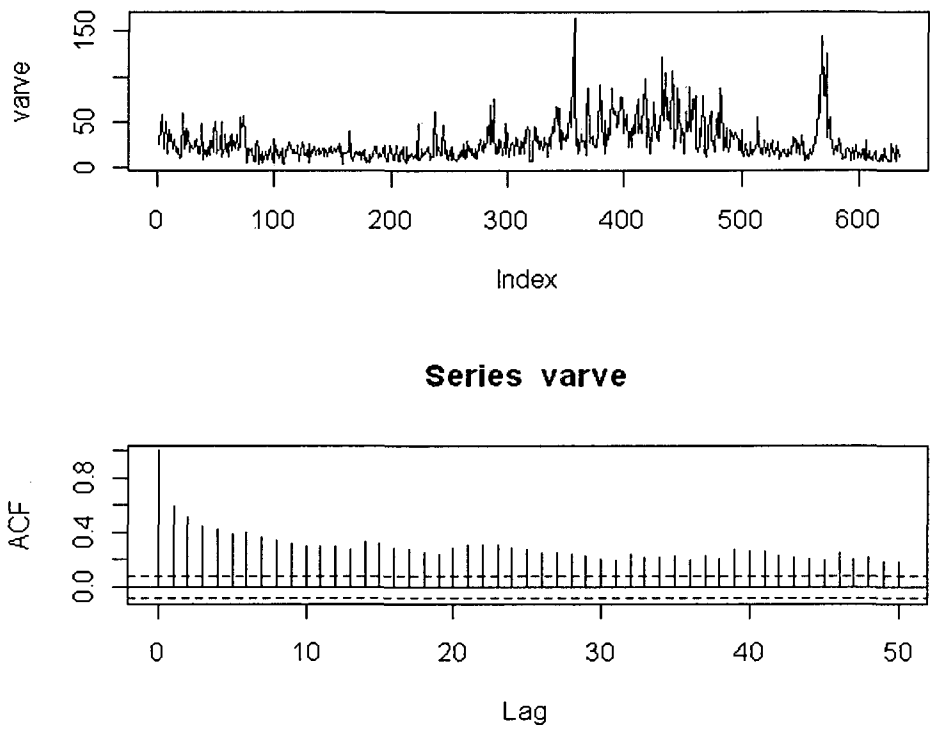

Figure 1.3: *varve data and plot of it's sample ACF*

Estimation for data with long-range dependence is quite different from that for observations with independence or short-range dependence. For example, Hall and Hart (1990) showed that the convergence rates of mean regression function estimators differ from those with independence or short-range dependence assumption. See figure (1.3).

In this thesis we suppose that the errors $\{\varepsilon_i, i \in \mathbb{Z}\}$ constitute a strictly stationary moving average sequence which is defined by

$$\varepsilon_i = \sum_{j \leq i} b_{i-j}\, \zeta_j, \qquad i \in \mathbb{Z}. \tag{1.1.6}$$

Here $\{\zeta_j,\ j \in \mathbb{Z}\}$ is a sequence of i.i.d. random variables with mean zero, variance σ^2, and $b_i,\ i \in \mathbb{Z}_+$ are nonrandom weights such that $\sum_i b_i^2 = \sigma^{-2}$ [This implies

that $E(\varepsilon_i^2) = 1$ for all $i \in \mathbb{Z}]$. Furthermore, we assume that the weights decay hyperbolically, i.e.,

$$b_i \sim C_1 i^{-(1+\alpha)/2}, \qquad 0 < \alpha < 1, \tag{1.1.7}$$

where C_1 is a constant. From equations (1.1.6) and (1.1.7), one can verify that (1.1.5) holds with $C_0 = C_1^2 \sigma^2 \int_0^\infty (u + u^2)^{-(1+\alpha)/2} du$. Hence the random errors $\{\varepsilon_i, i \in \mathbb{Z}\}$ defined in (1.1.6) have long memory, and their distribution may be more general than being Gaussian.

The family of long memory processes defined by (1.1.6) includes the important class of fractional ARIMA processes. For more information on their applications in economics and other sciences, see Baillie (1996). For various theoretical results pertaining to the empirical processes of long memory moving averages, we refer to Ho and Hsing (1996, 1997), Giraitis, *et al.* (1996, 1999), Koul and Surgailis (1997, 2001) and the references therein.

The main objective of the present thesis is to study the wavelet-based estimator of the regression function g in (1.1.1), where g belongs to a large function class that may have a large number of jump discontinuities, and the number of jump discontinuities diverges polynomially fast with sample size. We investigate the asymptotic convergence rates of the estimators and show that discontinuities of the unknown curve have a negligible effect on the performance of nonlinear wavelet curve estimators.

1.2 Background of The Problem

Wavelet methods in nonparametric curve estimation have become a well-known technique. For a systematic discussion of wavelets and their applications in statistics, see the monograph by Härdle, *et al.* (1998). The major advantage of the wavelet method is its adaptability to the varying degrees of smoothness of the underlying unknown curves. Wavelet estimators typically achieve the optimal convergence rates over exceptionally large function spaces. For reference, see Donoho and Johnstone (1995, 1998), Donoho, *et al.* (1995, 1996) and Hall, *et al.* (1998, 1999). The results of the above papers are based on the assumption that the errors are independent normal variables. For correlated noise, Wang (1996) and Johnstone and Silverman (1997) examine the asymptotic properties of wavelet-based estimators of mean regression function with long memory Gaussian noise. Kovac and Silverman (2000) and von Sachs and Macgibbon (2000) consider a correlated heteroscedastic and/or nonstationary noise sequence. They show that these estimators achieve minimax rates over a wide range of function spaces. In those papers it is assumed that the underlying function belongs to a large smooth function space. Li and Xiao (2007) consider the block thresholding wavelet estimation of a mean regression function when the errors are long memory Gaussian processes.

In this thesis, we assume that the mean regression function g belongs to a large class of functions with discontinuities. The observational errors follow a long memory moving average precess which is primarily non-Gaussian. More specifically, we will

consider two types of assumptions on the random variables $\{\zeta_j,\, j \in \mathbb{Z}\}$ in (1.1.6), which lead to large and moderate deviations of weighted partial sums of the long-range dependent errors $\{\varepsilon_i, i \in \mathbb{Z}\}$. The first assumption on the random variables $\{\zeta_j,\, j \in \mathbb{Z}\}$ is the Statulevičius condition (S_γ): There exist constants $\gamma \geq 0$ and $\Delta > 0$ such that

$$|\Gamma_m(\zeta_j)| \leq \frac{(m!)^{1+\gamma}}{\Delta^{m-2}} \quad \text{for} \quad m = 3, 4, \ldots, \tag{1.2.1}$$

where $\Gamma_m(\zeta_j)$ denotes the cumulant of ζ_j of order m; see Section 2.3 for its definition and some basic properties. Amosova (2002) has shown that, when $\gamma = 0$, the condition (S_γ) is equivalent to the celebrated Cramér condition; and when $\gamma > 0$, it is equivalent to the Linnik condition. Hence, the class of random variables satisfying (S_γ) is very large. Our main result is Theorem 3.1, where we show that the wavelet-based estimators, based on simple thresholding of the empirical wavelet coefficients, attain nearly optimal convergence rates over a large space of non-smooth functions. For proving this result, we will establish a Bernstein-type exponential inequality for a weighted sums of i.i.d. random variables ζ_j (see Lemma 3.7 below), which may be of independent interest.

The second assumption on $\{\zeta_j,\, j \in \mathbb{Z}\}$ is weaker than the condition (S_γ) and it only requires $E(|\zeta_1|^{2+\eta}) < \infty$ for a certain constant $\eta > 0$. We will show that, after adjusting the threshhold appropriately, the main result still holds under the latter moment condition.

The rest of this thesis is organized as follows. In the next chapter, we recall some elements of wavelet transforms, provide nonlinear wavelet-based mean regression

function estimators and state some large and moderate deviation estimates, due to Bentkus and Rudzkis (1980), Petrov (2002) and Frolov (2005), respectively. These results are applicable to weighted partial sums of the random variables $\{\varepsilon_i, i \geq 1\}$. The main results (i.e. Theorem 3.1 and Theorem 3.8) are provided in Chapter 3, together with some discussions. Chapter 4 contains a modest simulation study. The proof of the main results appears in Chapter 5.

In the above, we use C to denote positive and finite constants whose value may change from line to line. Specific constants are denoted by C_0, C_1, C_2, A, B, M and so on. Throughout this thesis, $\mathbb{R}$ is the set of real numbers, $\mathbb{Z}$ is the set of integers, $\mathbb{L}^2(\mathbb{R})$ is the set of square integrable real-valued functions on $\mathbb{R}$ and $\mathbb{L}^\infty(\mathbb{R})$ is the set of bounded integrable functions on $\mathbb{R}$.

Chapter 2

Wavelets and Nonparametric Regression

In this chapter, an overview of background knowledge relevant to subsequent chapters is given.

2.1 Wavelets and Multiresolution Analysis

Firstly an introduction of wavelets and relevant properties is presented. Then the definition of multiresolution analysis is given and we will show how wavelets fit into it.

Review of Wavelets

The definition and a brief introduction of wavelets and how they evolved over time

are provided here, see Vidakovic (1999). More detailed mathematical descriptions of wavelets can be found in Meyer(1992) and Daubechies (1992).

There are a number of ways of defining a wavelet. The first "wavelet basis" was discovered in 1910 when Alfred Haar showed that any continuous function $f(x)$ on $[0,1]$ can be approximated by

$$f_n(x) = \langle \xi_0, f \rangle \xi_0(x) + \langle \xi_1, f \rangle \xi_1(x) + \cdots + \langle \xi_n, f \rangle \xi_n(x),$$

and that, when $n \to \infty$, f_n converges to f uniformly, where $\langle \xi_i, f \rangle$ is the inner product of f and ξ_i. The Haar basis is very simple:

$$\xi_0(x) = I(0 \le x \le 1),$$

$$\xi_1(x) = I(0 \le x \le 1/2) - I(1/2 \le x \le 1),$$

$$\xi_2(x) = \sqrt{2}[I(0 \le x \le 1/4) - \sqrt{2} I(1/4 \le x \le 1/2)]$$

$$\cdots$$

$$\xi_n(x) = 2^{j/2}[I(k2^{-j} \le x \le (k+1/2)2^{-j}) - I((k+1/2)2^{-j} \le x \le (k+1)2^{-j})],$$

$$\cdots$$

where n is uniquely decomposed as $n = 2^j + k$, $j \ge 0$, $0 \le k \le 2^j - 1$, and $I(A)$ is the indicator function of a set A.

The first definition of wavelets can be attributed to Morlet et al. (1982) and Morlet and Grossmann (1984), and it is given in the Fourier domain: A **wavelet** is an $\mathbb{L}^2(\mathbb{R})$ function for which the Fourier transformation $\Psi(\omega)$ satisfies

$$\int_0^\infty |\Psi(t\omega)|^2 \frac{dt}{t} = 1,$$

for almost all ω.

The definition of Morlet and Grossmann is quite broad, and over time, the meaning of the term *wavelet* became narrower. Currently, the term wavelet is usually associated with a function $\psi \in \mathbb{L}^2(\mathbb{R})$ such that the translations and dyadic dilations of ψ,

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), j, k \in \mathbb{Z}$$

constitute an orthonormal basis of $\mathbb{L}^2(\mathbb{R})$.

Later, Meyer (1992, page 66) gave an elaborate definition of mother wavelet ψ which describes most of the properties we want to give to a wavelet function:

Let r be a non-negative integer. A function $\psi(x)$ of a real variable is called a **basic wavelet** of class r if the following properties hold:

(a) if $r = 0$, $\psi(x) \in \mathbb{L}^\infty(\mathbb{R})$; if $r > 1, \psi(x)$ and all its derivatives up to order r belong to $\mathbb{L}^\infty(\mathbb{R})$;

(b) $\psi(x)$ and all its derivatives up to order r decrease rapidly as $x \to \infty$;

(c) $\int_{-\infty}^{\infty} x^k \psi(x) dx = 0$ for $0 \leq k \leq r$;

(d) the collection of functions $2^{j/2}\psi(2^j x - k)$, $j, k \in \mathbb{Z}$ is an orthonormal basis of $\mathbb{L}^2(\mathbb{R})$.

Condition (a) is associated the regularity(See page 15) of the basic wavelet when the wavelet function is compactly supported. The Condition (b) describes the lo-

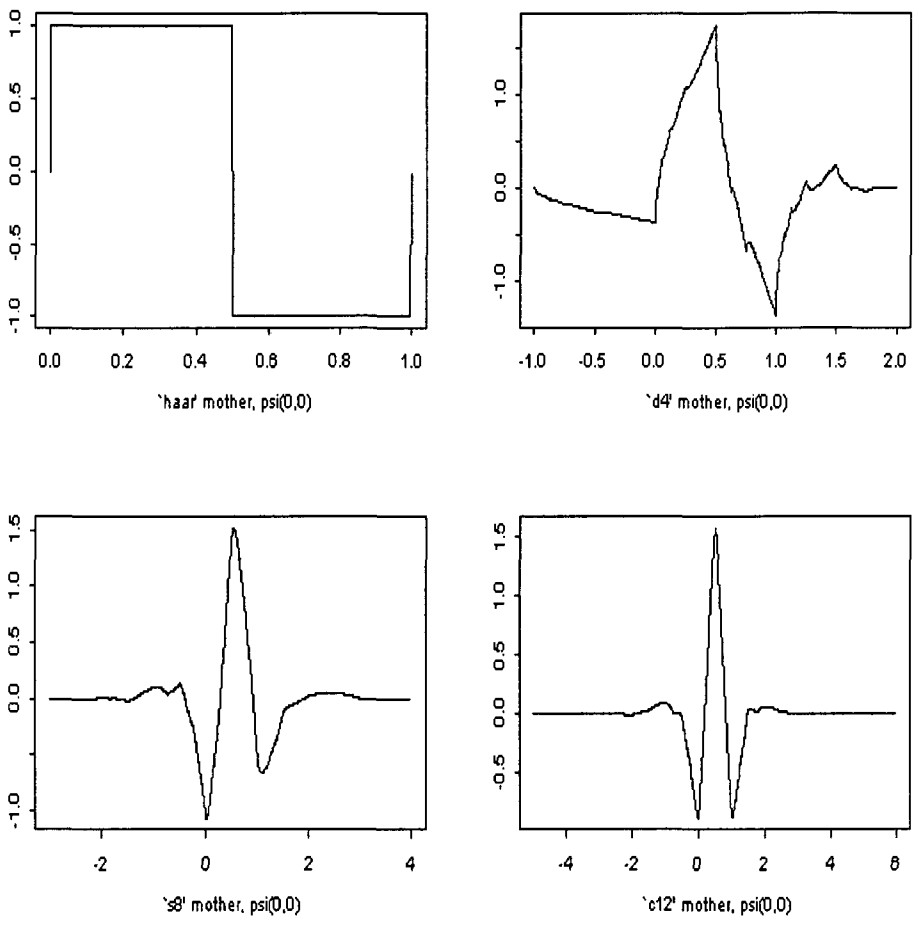

Figure 2.1: *Four different orthogonal mother wavelets "haar", "d4", "s8", and "c12"*

calization property and extends also to the frequency domain. With regard to this property, many wavelets used in practice are compactly supported. Condition (c) specifies the oscillatory character, known as the vanishing moments property(See page 14). Condition (a), (b), and (c) are the characteristics we want to give to a mother wavelet. There are many mother wavelets, e.g. the well-known Haar wavelet, discovered by the mathematician Haar in 1910, Symmlet wavelet, Daubechies wavelet and Coiflet wavelet, all discussed by Daubechies (1992). See Figure 2.1. Although they have different expressions and characteristics, all of them satisfy the above definition.

A wavelet ψ has r-vanishing moments if

$$\int x^k \psi(x) dx = 0, \quad for \quad k = 0, 1, \cdots, \quad r-1.$$

Mallat's Multiresolution Analysis

Mallat's Multiresolution Analysis (MRA) provides a tool to constructively describe different wavelet bases(Mallat, 1989).

A multiresolution analysis (MRA) is a sequence of closed subspaces $V_n, \quad n \in \mathbb{Z}$ in $\mathbb{L}^2(\mathbb{R})$ satisfying the following properties:

(1) $V_j \subset V_{j+1}$,

(2) $f(\cdot) \in V_j \Leftrightarrow f(2\cdot) \in V_{j+1}$,

(3) $f(\cdot) \in V_0 \Leftrightarrow f(\cdot - k) \in V_0$,

(4) $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$,

(5) $\overline{\bigcup_{j \in \mathbb{Z}} V_j} = \mathbb{L}^2(\mathbb{R}), \quad j \in \mathbb{Z}$, i.e., $\{V_j\}_{j \in \mathbb{Z}}$ is dense in $\mathbb{L}^2(\mathbb{R})$.

(6) a scaling function $\phi \in V_0$ has a non-vanishing integral such that the collection $\{\phi(x-k) | k \in \mathbb{Z}\}$ constitutes an orthonormal basis for the space V_0.

Condition (1) implies that the orthogonal complement W_j of V_j in V_{j+1} can be found such that $V_{j+1} = V_j \bigoplus W_j$, where the symbol $\bigoplus$ stands for direct sum. Similarly, $V_j = V_{j-1} \bigoplus W_{j-1}$ and so on. It follows that W_{j-1} is also orthogonal to W_j and all the spaces W_j (unlike the spaces V_j) are mutually orthogonal.

From condition (2) and (3), $\forall j \in \mathbb{Z}$, $\{\phi_{jk}, k \in \mathbb{Z}\}$ constitutes an orthonormal basis for V_j , where

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k).$$

Let P_j be the orthogonal projection operator onto V_j. Condition (4) implies that, when $j \to -\infty$, we lose all the details of f, and $\{P_j f\}$ converges to $\{0\}$ in an $\mathbb{L}^2$ space, which could be expressed as $\lim_{j\to-\infty} P_j f = 0$, where convergence of $P_j f$ in an $\mathbb{L}^2$ space means that $\lim_{j\to-\infty} \int_{\mathbb{R}} |P_j f(x)|^2 dx = 0$. The other end, in the same sense, ensures that the signal approximation converges to the original signal in condition (5): $\lim_{j\to\infty} P_j f = f$. The approximation $P_j f$ of a function f at resolution level j is given by

$$P_j f(x) = \sum_{k\in\mathbb{Z}} \alpha_{jk}\phi_{jk}(x),$$

where coefficients

$$\alpha_{jk} = \int_{-\infty}^{\infty} f(x)\phi_{jk}(x)\, dx.$$

Condition (6) gives a definition of MRA and ϕ is called r-regular, if $\phi \in C^r$, where C^r is the set of functions having derivatives up to order r, and ϕ and every derivative up to order r can be chosen in such a way that for every integer $m \geq 0$, there exists a constant C_m satisfying

$$|\phi^{(j)}(x)| \leq \frac{C_m}{(1+|x|)^m} \quad for \quad j = 0, 1, \cdots, r.$$

The following graph (Figure 2.2) shows the projection of the original signal into different orthogonal spaces.

Define functions

$$S_J(t) = \sum_k s_{J,k}\phi_{J,k}(t)$$

and

$$D_j(t) = \sum_k d_{j,k}\psi_{j,k}(t)$$

to be the smooth signal and the detail signals respectively. The orthogonal wavelets series approximation to a continuous signal $f(t)$ is expressed in terms of these signals:

$$f(t) \approx S_J(t) + D_J(t) + D_{J-1}(t) + \cdots + D_1(t).$$

The terms in this approximating sum constitute a decomposition of the signal into orthogonal signal components $S_J(t)$, $D_J(t)$,$D_{J-1}(t)$, $\cdots$, $D_1(t)$ at different scales. Because the terms at different scales represent components of the signal $f(t)$ at different resolutions, the approximation is called a multiresolution decomposition (MRD).

The fine scale features-the high frequency oscillations at the beginning of the signal, are captured mainly by the fine scale detail components D_1 and D_2. The coarse scale components D_6 and S_6 correspond to lower frequency oscillations towards the end of the series.

Wavelet Transform

In the later part of this section, we show how a fast wavelet transform can be derived from the multiresolution analysis properties.

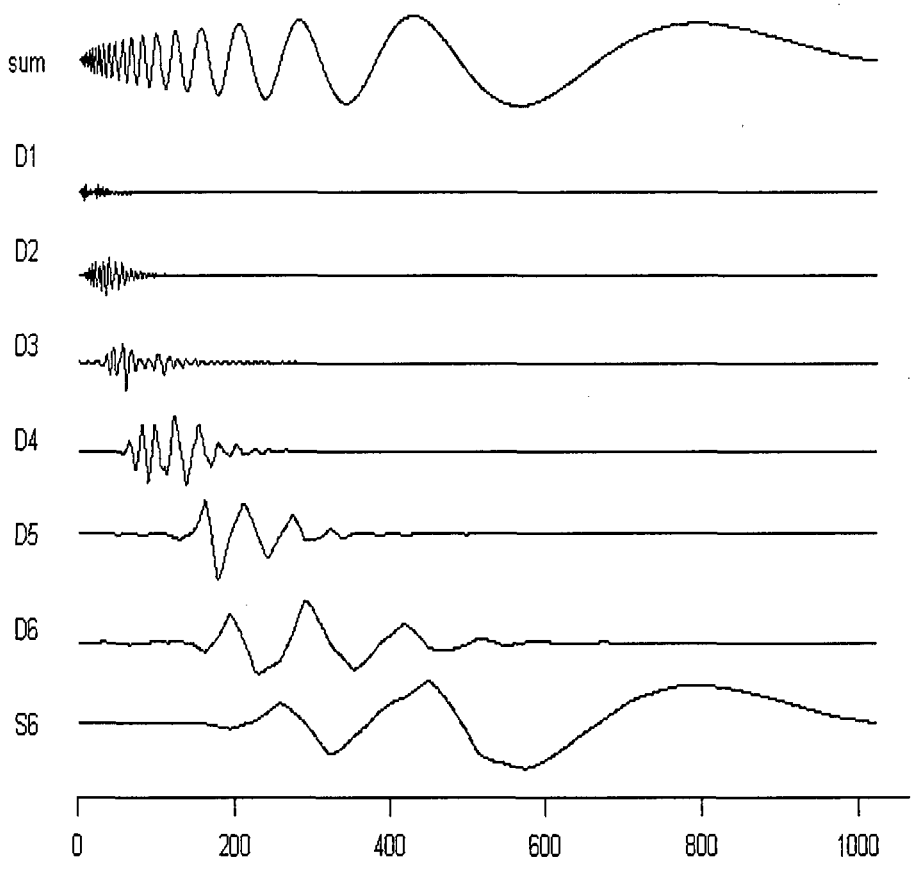

Figure 2.2: *Multiresolution decomposition of the doppler signal*

Wavelet Representation

From the definition of MRA and property $V_j \subset V_{j+1}$, there exists an orthogonal complement W_j of V_j such that $V_{j+1} = V_j \oplus W_j$ with $V_j \perp W_j$. Therefore for some $j_0 \in \mathbb{Z}$, there is a series of mutually orthogonal subspace W_j, $j \in \mathbb{Z}$, such that $V_j = V_{j_0} \oplus \bigoplus_{k=j_0}^{j-1} W_k$ for $j > j_0$. $\mathbb{L}^2(\mathbb{R})$ can be decomposed into mutually orthogonal subspaces, i.e., $\bigoplus_{j\in\mathbb{Z}} W_j = \mathbb{L}^2(\mathbb{R})$.

A scaling function $\phi \in V_{j_0}$ with a non-vanishing integral exists such that the collection $\{\phi(x-k)|k \in \mathbb{Z}\}$ constitutes an orthonormal basis for V_{j_0}. Now we consider the generation of an orthonormal wavelet basis for functions $f \in \mathbb{L}^2(\mathbb{R})$. For some $j_0 \in \mathbb{Z}$, $\{\phi_{j_0k}, \psi_{jk} : j, k \in \mathbb{Z}, j \geq j_0\}$ forms an orthonormal basis for $\mathbb{L}^2(\mathbb{R})$.

Therefore, a function f in $\mathbb{L}_2(\mathbb{R})$ can be represented as

$$f(x) = \sum_{k\in\mathbb{Z}} \alpha_{j_0k}\phi_{j_0k}(x) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(x), \tag{2.1.1}$$

Where the coefficients are given by

$$\alpha_{j_0k} = \int f(x)\phi_{j_0k}(x)\,dx, \quad \beta_{jk} = \int f(x)\psi_{jk}(x)\,dx$$

The orthogonality properties of ϕ and ψ imply:

$$\begin{aligned} &\int_{-\infty}^{\infty} \phi_{j_0k_1}(x)\phi_{j_0k_2}(x)dx = \delta_{k_1k_2}, \quad \int_{-\infty}^{\infty} \psi_{j_1k_1}(x)\psi_{j_2k_2}(x)dx = \delta_{j_1j_2}\delta_{k_1k_2}, \\ &\int_{-\infty}^{\infty} \phi_{j_0k_1}(x)\psi_{jk_2}(x)dx = 0, \quad \forall j_0 \leq j, \end{aligned} \tag{2.1.2}$$

where δ_{jk} denotes the Kronecker delta, i.e., $\delta_{jk} = 1$, if $j = k$; and $\delta_{jk} = 0$, otherwise. For more information on wavelets see Daubechies (1992).

The first term on the right hand side of (2.1.1) is the projection $P_{j_0}f$ of f at resolution level j_0. Using $V_{j+1} = V_j \oplus W_j$, and since $\{\psi_{jk} : k \in \mathbb{Z}\}$ is a basis for W_j, $\sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} \beta_{jk}\psi_{jk}(x)$ is the difference between $P_j f$ and the finer resolution approximation $P_{j+1}f$. So for each value of j, the second term in (2.1.1) adds another level of detail into the representation.

Due to the vanishing moments property, if f is smooth, the wavelet representation is very economical because there will be few wavelet coefficients β_{jk} that are noticeably different from 0. Also, because wavelets are localized in time and scale, a discontinuity, or other high frequency feature, in f will only result in large wavelet coefficients for values of k corresponding to the location of the feature. Therefore, many functions can be adequately represented by a small number of wavelet coefficients. This property explains the application of wavelets to data compression and is also important in statistical applications.

Discrete Wavelet Transformation

Given real life data in a statistical setting, we are typically concerned with discrete samples, rather than continuous functions, since data are observed at a finite number of discrete time points in practice. Therefore, a discrete wavelet transform (DWT) is born.

First consider some properties of ϕ. Since $\phi \in V_0 \subset V_1$, there exist an h_n such

that $\phi(x) = \sum_n h_n \phi_{1,n}(x)$, where

$$h_n =< \phi, \phi_{1,n} >= \int \phi(x)\phi_{1,n}(x)dx.$$

Therefore, for all $j, k \in \mathbb{Z}$,

$$\phi_{j-1,k}(x) = \sum_{l\in\mathbb{Z}} h_{l-2k}\phi_{j,l}(x),$$

and $\sum_{l\in\mathbb{Z}} |h_l|^2 = 1$. Similarly, we have

$$\psi_{j-1,k}(x) = \sum_{l\in\mathbb{Z}} g_{l-2k}\phi_{j,l}(x).$$

Mallet showed that one possible choice is that $g_n = (-1)^n h_{1-n}$.

The recursive relationship below between the scaling function and wavelet coefficients at successive levels can be obtained from the previous equations.

$$\alpha_{j-1,k} = \int f(x)[\sum_l h_{l-2k}\phi_{jl}(x)]dx$$

$$= \sum_l h_{l-2k}[\int f(x)\phi_{jl}(x)dx] = \sum_l h_{l-2k}\alpha_{j,l},$$

and with the same reasoning

$$\beta_{j-1,k} = \sum_l g_{l-2k}\alpha_{j,l}.$$

This recursive relationship is another important property of wavelet transform held between scaling function coefficients and wavelet coefficients at successive levels. This property is related to the pyramid algorithm, a fast algorithm to calculate the coefficients provided by Mallat (1989).

Consider a vector of function values $f = (f(t_1), \cdots, f(t_n))^T$ at equally spaced points t_i, and let n be an interger power of 2, say 2^{J+1}. A function can be constructed

at level $J+1$ as follows:

$$\hat{f}_{J+1}(x) = \sum \alpha_{J+1,k}\phi_{J+1,k}(x),$$

where $\alpha_{J+1,k} = f(t_k)$. The function $\hat{f}_{J+1}(x)$ is an element of V_{J+1} and can be projected onto spaces V_J and W_J, giving

$$\hat{f}_{J+1}(x) = (P_{V_J}\hat{f}_{J+1}(x)) + (P_{W_J}\hat{f}_{J+1}(x)) = \sum_l \alpha_{J,l}\phi_{J,l}(x) + \sum_l \beta_{J,l}\psi_{J,l}(x).$$

The corresponding scaling coefficients in level J are

$$\alpha_{J,l} = \langle \hat{f}_{J+1}, \phi_{J,l}\rangle = \sqrt{2}\langle \hat{f}_{J+1}, \sum_k h_{k-2l}\phi_{J+1,k}\rangle = \sqrt{2}\sum_k h_{k-2l}\alpha_{J+1,k}.$$

Similaryly, the wavelet coefficients are

$$\beta_{J,l} = \sqrt{2}\sum_k g_{k-2l}\beta_{J+1,k}.$$

Applying this procedure recursively, we can find the coefficients α_{j_0k} and β_{jk}, for $j_0 \leq j \leq J$.

Note that at each level of the reconstruction, finer scale coefficients are obtained from coarser ones as illustrated by

$$\begin{aligned}
&\sum_k \alpha_{j-1,k}\phi_{j-1,k}(x) + \sum_k \beta_{j-1,k}\psi_{j-1,k} \\
&= (Proj_{V_{j-1}}\hat{f})(x) + (Proj_{W_{j-1}}\hat{f})(x) \\
&= (Proj_{V_j}\hat{f})(x) = \sum_k \alpha_{j,k}\phi_{j,k}(x),
\end{aligned}$$

where

$$\alpha_{j,k} = < \phi_{j,k}, Proj_{V_j}\hat{f} >$$

$$= \sum_l \alpha_{j-1,l} < \phi_{j,k}, \phi_{j-1,l} > + \sum_l \beta_{j-1,l} < \phi_{j,k}, \psi_{j-1,l} > .$$

This gives Mallet's Pyramid Algorithm.

See Figure 2.3 below for an illustration of how the DWT works for a dopplor signal (the first one), and the wavelet coefficients in different levels are shown. Observe the following properties of the DWT coefficients:

1. Typically, the wavelet coefficients at coarse scales are larger than the wavelet coefficients at fine scales. This is a cosequence of the smoothness of the doppler signal.

2. The smooth coefficients $s_{6,k}$ correspond to the smooth at scale 2^6, mainly capturing the low frequency oscillations in the latter portion of the signal.

3. The detail coefficients $d_{6,k}$, $d_{5,k}$, $\cdots$, $d_{1,k}$ represent progressively finer "correlations" to the smooth trend, capturing the higher frequency oscillations in the beginning of the signal.

4. The coefficients are sparse in the sense that many coefficients are very small or nearly zero.

Wavelet Analysis vs. Fourier Analysis

The fast Fourier transform (FFT) and the discrete wavelet transform (DWT) share some similarities: both of them are linear operations, and the mathematical properties of the matrices involved in the transforms are also similar. The inverse

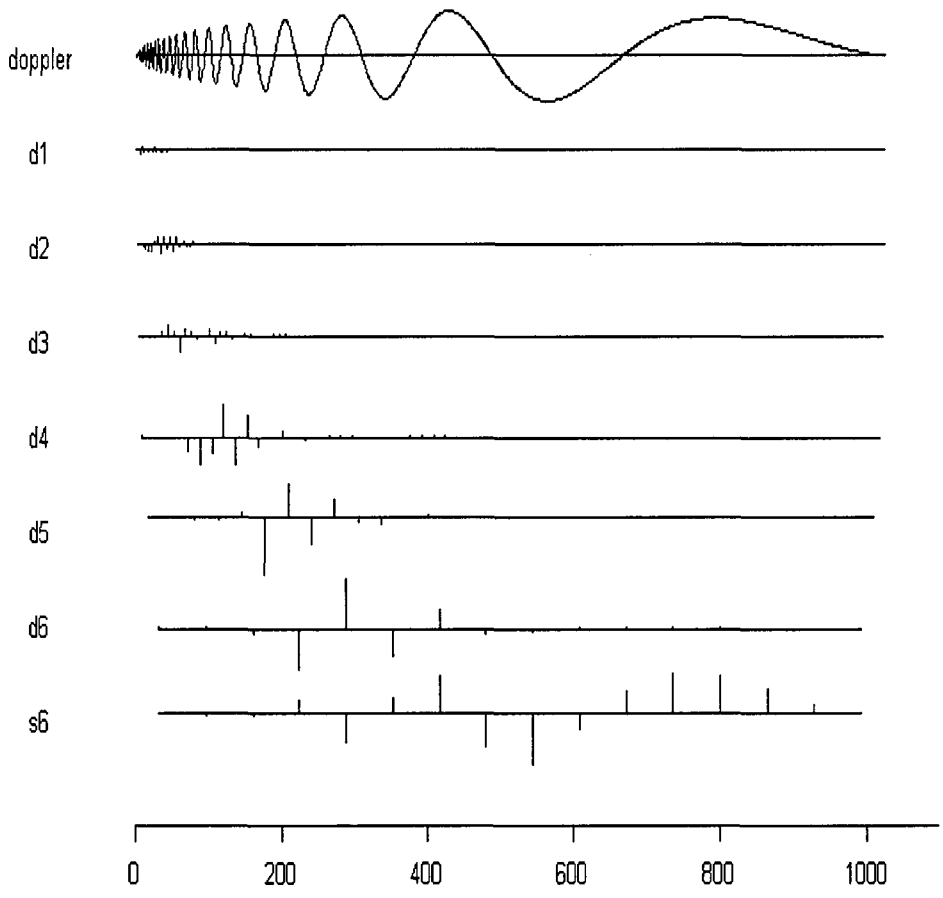

Figure 2.3: *DWT of the doppler signal using default s8 wavelet*

transform matrix of both FFT and DWT is the transpose of the original. Another similarity is that the basis functions of both transforms are localized in frequency.

However, it is the difference between the two that strikes us and makes DWT stand out from FFT. The most interesting difference between these two kinds of transforms is that individual wavelet functions are localized in space, while the Fourier sine and cosine functions are not. The localization feature in both frequency scale via dilations and space via translations makes wavelets very useful and more trustworthy in many cases. For example, one major advantage of wavelet methods is their high adaptability and their ability to capture discontinuities and singularities. Another consequent advantage is the sparseness of wavelets coefficients when functions and operators are transformed into the wavelet domain. This sparseness results in a number of useful applications, such as removing noise from data, and will be discussed later.

Figure 2.4 is a graph comparing wavelet basis and Fourier basis.

2.2 Nonparametric Regression and Wavelet Shrinkage

In nonparametric regression problems, we want to estimate an unknown signal $f(t)$ from some data y_i that contain noise. For example, suppose we are given n noisy samples of a function f:

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

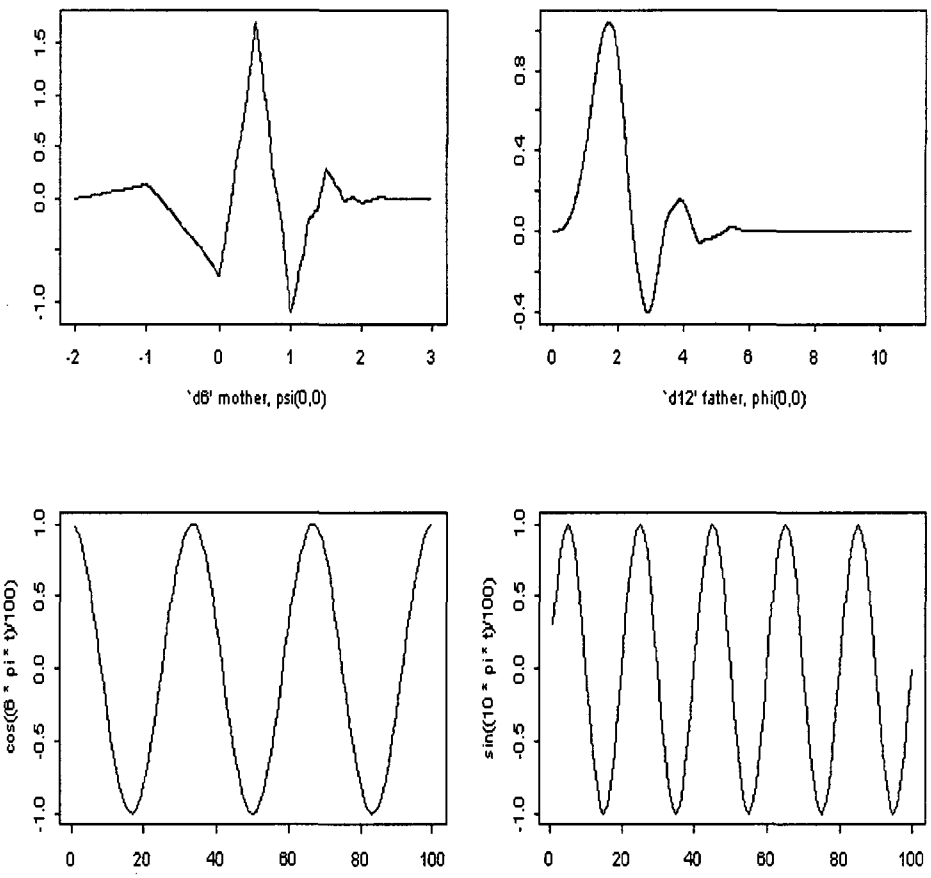

Figure 2.4: *wavelet basis vs. Fourier basis*

Our goal is to estimate f with the least mean square error:

$$R_n(\hat{f}, f) = E\|\hat{f} - f\|_2^2 = E\int_0^1 (\hat{f}(t) - f(t))^2 dt.$$

The usual parametric regression requires knowing a particular model for f. In nonparametric regression, we make minimal assumptions about the exact nature of f. We only know a priori that f belongs to a certain class $\mathcal{F}$ of smooth functions, but nothing more. Some of the common estimators include those based on kernel functions, smoothing splines and orthogonal series. Each one has its own strength and weakness. A typical drawback to these nonparametric techniques is that they could fail unless strong smoothness assumptions are satisfied everywhere.

Wavelet-based methods are developing in statistics in areas such as regression, density and function estimation, modeling and forecasting in time series analysis, and spatial statistics. One of the most successful applications of wavelets is in nonparametric statistical estimation. Donoho and Johnstone showed that by shrinking wavelets coefficients, wavelets estimators for nonparametric regression problems had statistical optimality properties, with $\hat{f}$ attaining the minimax risk

$$\mathcal{R}(n, \mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} R_n(\hat{f}, f).$$

Wavelet Shrinkage and Thresholding Procedure

It was pointed out by many researchers that linear methods are not efficient when signals have considerable time-inhomogeneity such as varying degrees of smoothness.

Non-linear estimators can improve the efficiency and achieve better rates. The key advantages of wavelet estimators can be fully explored only when considering non-linear wavelet estimators. The non-linearity comes from shrinking or thresholding the empirical coefficients $\hat{\beta}_{jk}$, while the scaling function coefficients $\hat{\alpha}_{j_0k}$ are kept untouched. The coefficients $\hat{\beta}_{jk}$, $j = j_0, \cdots J$, $k = 0, \cdots 2^j - 1$ and $\hat{\alpha}_{j_0k}$, come from the DWT of the noisy data. Wavelet shrinkage and thresholding approaches were first introduced by Donoho and Johnstone (1994). The goal in this situation is to recover a signal in the presence of noise with non-random design point x_i taken to be $x_i = i/n$.

Donoho and Johnstone (1994, 1995, 1998) have developed an impressive theory and methodology for nonparametric regression based on the principle of wavelet shrinkage. To be more detailed, wavelet shrinkage refers to estimates obtained by:

i) Applying the discrete wavelet transform (DWT) to observations $y_i, \quad i = 1, 2, \cdots, n$, to obtain a sequence of wavelet coefficients $d_i, \quad i = 1, 2, \cdots, n$.

ii) Using threshold methodology to shrink the wavelet coefficients $d_i, \quad i = 1, 2, \cdots, n$.

iii) Applying the inverse discrete wavelet transform to thresholded coefficients to recover the estimator of the function f.

Figure 2.5, 2.6 and 2.7 give an graphical illustration of how DWT works.

Steps i) and iii) are straightforward to implement, once the wavelet basis functions have been chosen. Some fast and efficient algorithms are available for performing the calculations. Step ii), the aim is to de-noise the empirical wavelet coefficients. There have been a number of approaches for a proper threshold, including the following.

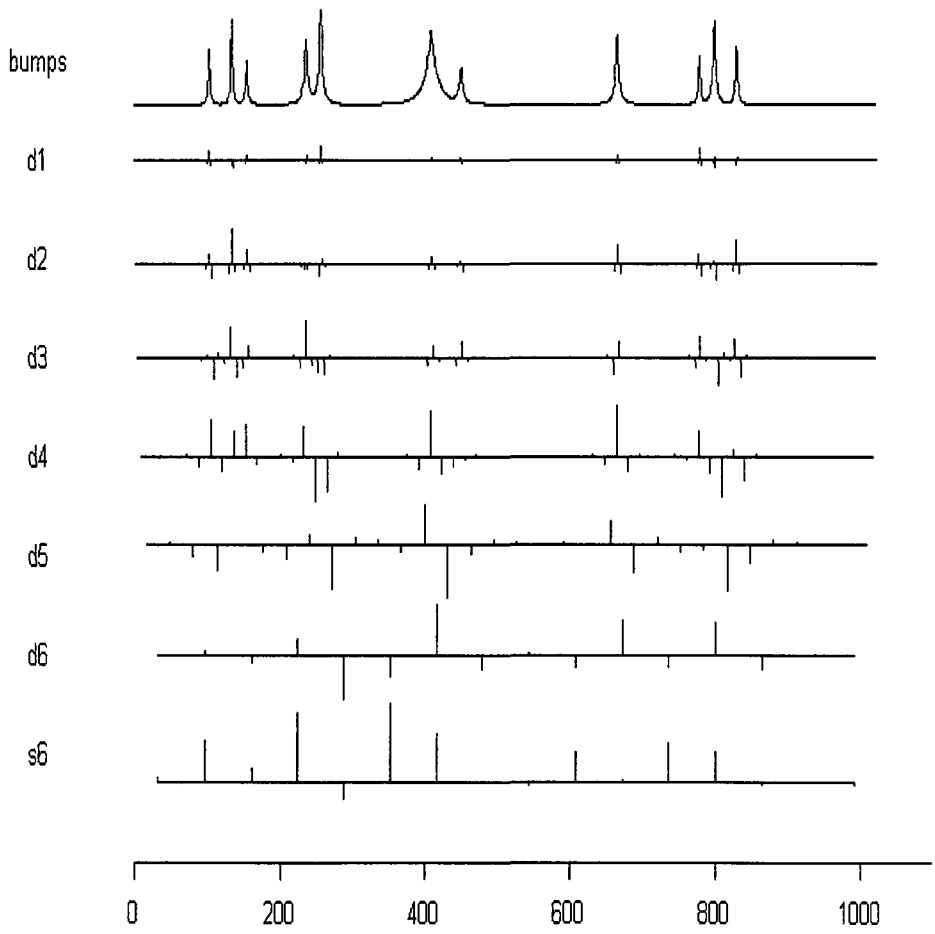

Figure 2.5: *The bumps signal and its DWT*

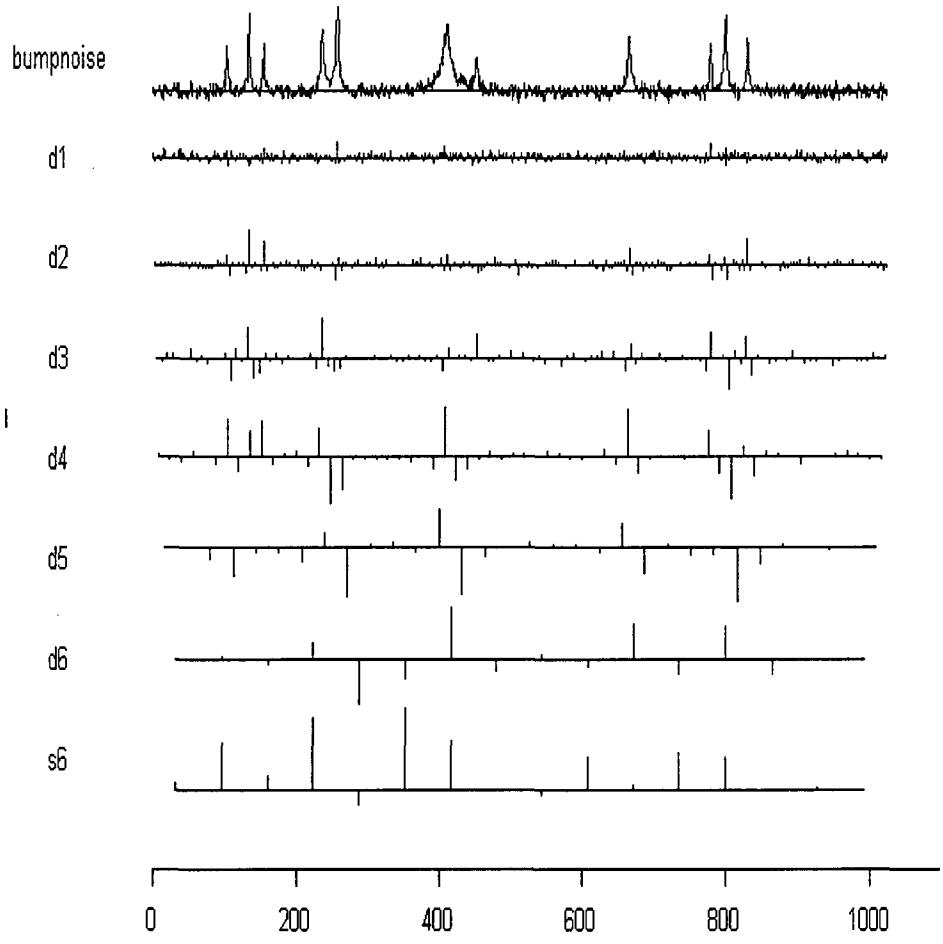

Figure 2.6: *DWT of the noise.bumps signal*

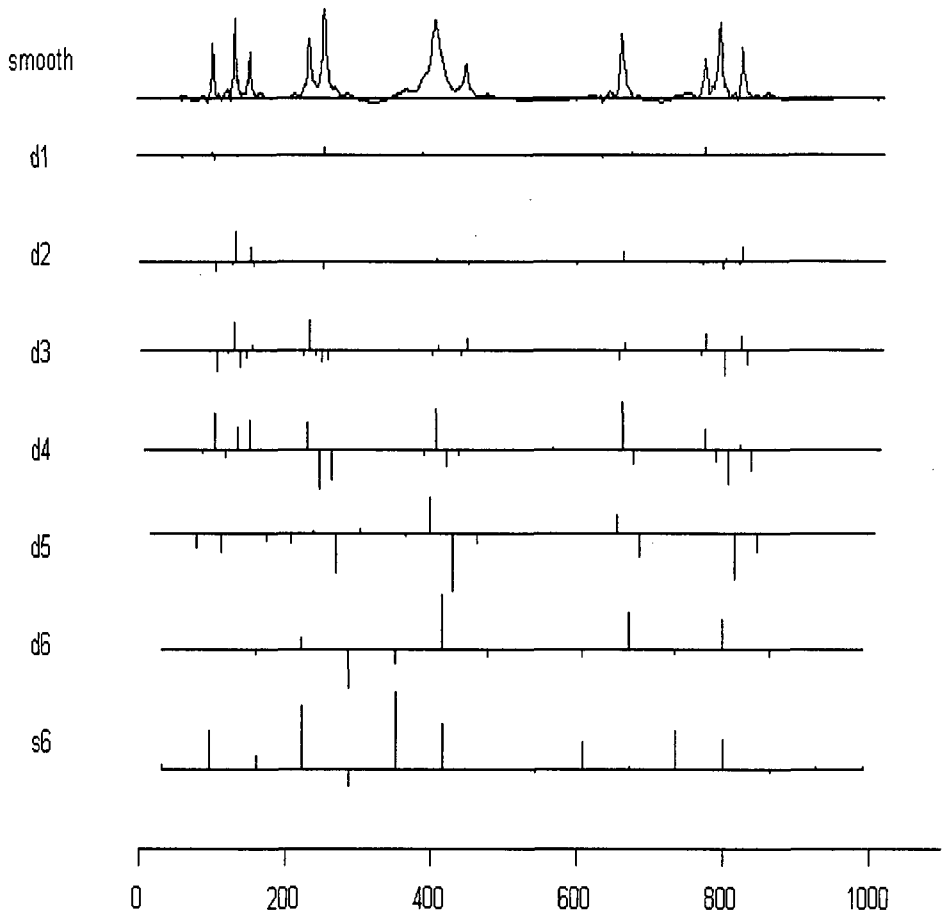

Figure 2.7: *Apply the waveshrink function to the noisy.bumps signal and plot the DWT of the estimated signal*

- The classic thresholding scheme, the hard and soft thresholding methods, discussed in detail by Donoho and Johnstone (1994, 1995) and Donoho et al. (1995)

- The cross-validation scheme, see Nason (1996) and also Hall and Penev (2001).

- The frequentist block thresholding scheme. See Hall et al. (1998, 1999), Cai (1999, 2002) and Cai and Silverman (2001).

- The empirical Bayes (EB) methods, see Chipman et al. (1997)and Clyde and George (2000).

Antoniadis et al. (2001) gives a very comprehensive summary of the above methods.

Classical thresholding methods and Choices of Threshold

Donoho and Johnstone (1994, 1995) suggested two types of thresholding methods, hard and soft thresholding, based on the following assumptions: ϵ_i are independent Gaussian noise, then the wavelet coefficients are also contaminated with independent Gaussian noise. So in this case, the empirical wavelet coefficients can be written as

$$\hat{\beta}_{jk} = \beta_{jk} + \epsilon_{jk}$$

and $\hat{\beta}_{jk}$ is distributed as

$$\hat{\beta}_{jk} \sim N(\beta_{jk}, \sigma^2)$$

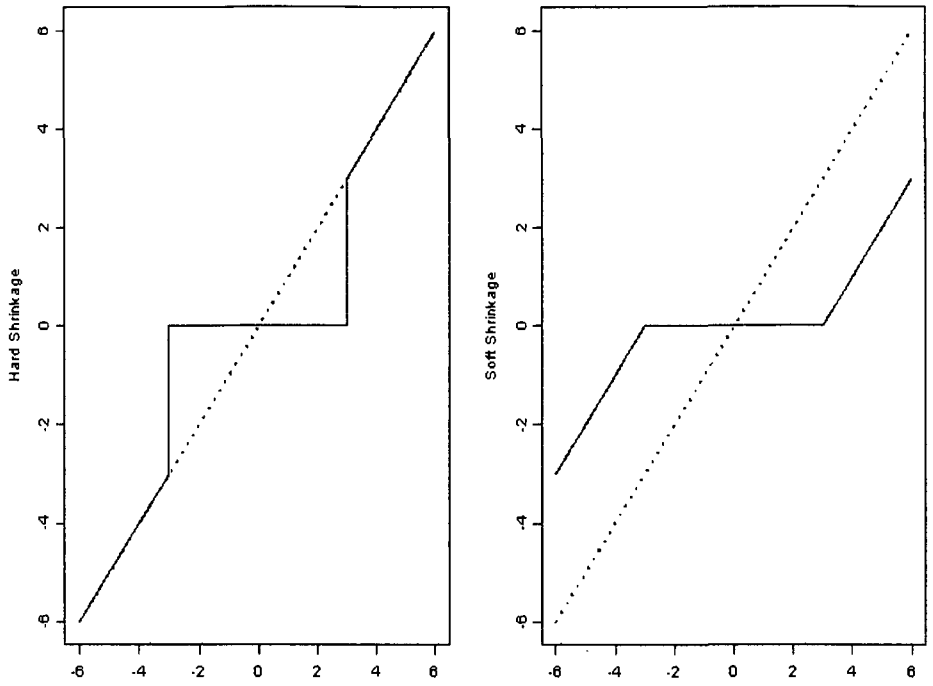

Figure 2.8: *The shrinkage function with threshold $c = 3$ (solid line) applied to a linear function (dashed line). Left: "Hard shrinkage". Right: "soft" shrinkage.*

Hard thresholding sets all the wavelet coefficients to be 0 if their absolute values are below a certain threshold $\lambda > 0$:

$$\delta^h(\hat{\beta}_{jk}, \lambda) = \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > \lambda).$$

Soft thresholding shrinks the wavelet coefficients that are larger than the threshold by λ:

$$\delta^s(\hat{\beta}_{jk}, \lambda) = sgn(\hat{\beta}_{jk})(|\hat{\beta}_{jk}| - \lambda) I(|\hat{\beta}_{jk}| > \lambda).$$

Hard and soft thresholdings are illustrated in the Figure 2.8.

After studying the performance of these thresholding methods, Dohono and John-

stone (1994, 1995) concluded that the resulting function estimate is asymptotically minimax (see section 2.2) for a wide variety of loss functions and functions f belonging to a wide range of smoothness classes. More importantly, they show that the wavelet estimator is nearly optimal for a wide variety of objectives.

Choice of Threshold

Clearly, an appropriate choice of a threshold value λ is fundamental to the effectiveness of the procedure described in the previous page. Too large a threshold might cut off important parts of the true function underlying the data, whereas too small a threshold may excessively retain noise in the reconstruction.

1. Universal threshold

Donoho and Johnstone (1994) proposed the universal threshold:

$$\lambda = \sigma\sqrt{2log(n)}.$$

When sigma is unknown, it may be replaced a robust estimate $\hat{\sigma}$, such as the median absolute deviation (MAD) of the wavelet coefficients at the finest level $J = log(N) - 1$ divided by 0.6745 and can be expressed as

$$\hat{\sigma} = MAD\{\beta_{Jk}, k = 1, \cdots, 2^J\}/0.6475.$$

Despite its simplicity, it can be shown that hard- and soft-thresholding rules with

universal threshold can asymptotically approach the "oracular" risk. Donoho and Johnstone (1994) showed that if $\{\epsilon_i\}_{i=1}^n$ is a white noise sequence with variance 1,

$$P(max(\epsilon_i) > \sqrt{2logn}) \to 0, as\, n \to \infty.$$

This means that in addition to its good asymptotic minimax properties, the universal threshold removes noise with high probability contributing to the visual quality of reconstructed signals.

However, the universal threshold depends on the data only through σ (or its esitmate). In fact, for large samples, it may be shown that the universal threshold will remove with high probability all the noise in the reconstruction, but part of the real underlying function might also be lost. As a result, the universal threshold tends to over-smooth data in practice.

2. SureShrink threshold

Donoho and Johnstone (1995) introduced a procedure, SureShrink, based on minimizing the Stein unbiased risk estimate (Sure). This threshold is implemented in an adaptive denoising procedure. The adaptation in SureShrink is achieved by specifying thresholds level-wise. The theoretical background for the threshold selection is in the following results:

Let $d_i \sim^{i.i.d.} \mathcal{N}(\theta_i, 1), i = 1, \cdots, k$. Let $\hat{\underline{\theta}}$ be an estimator of $\underline{\theta} = (\theta_1, \cdots, \theta_k)$. If the function $\mathbf{g} = \{g_i\}_{i=1}^k$, in the representation $\hat{\underline{\theta}}(\underline{d}) = \underline{d} + \mathbf{g}(\underline{d})$, is weakly differentiable, then

$$E^{\theta}\|\hat{\underline{\theta}} - \underline{\theta}\|^2 = k + E^{\theta}\{\|\mathbf{g}(\underline{d})\|^2 + 2\nabla\mathbf{g}(\underline{d})\}, \tag{2.2.1}$$

where $\mathbf{g} = \sum_{i=1}^{k}\{\frac{\partial}{\partial d_i}g_i\}$.

It is interesting that the estimator $\hat{\underline{\theta}}$ in 2.2.1 can be nearly arbitrary; for instance, it can be biased, non-linear, and so on. The application of 2.2.1 to soft threshold gives

$$SURE(\underline{d}, \lambda) = k - 2\sum_{i=1}^{k}\mathbf{I}(|d_i| \leq \lambda) + \sum_{i=1}^{k}(|d_i| \wedge \lambda)^2,$$

as an unbiased estimator of risk, i.e.,

$$E\|\delta^s(\underline{d}, \lambda) - \underline{\theta}\|^2 = ESURE(\underline{d}, \lambda).$$

When k is large, the law of large numbers(LLN) argument states that $SURE$ is close to its expectation, motivating the following threshold selection:

$$\lambda^{sure} = arg \min_{0 \leq \lambda \leq \lambda^U} SURE(\underline{d}, \lambda).$$

This procedure is very simple to implement since at each level, there are only k such values, and the algorithm to calculate λ^{sure} is fast. It has been shown that SureShrink is smoothness-adaptive: if the unknown function contains jumps, the reconstruction does also; if the unknown function has a smooth piece, the construction is as smooth as what the mother wavelet allows. In addition, this shrinkage can be tuned to be asymptotically minimax over a wide range of smoothness classes.

3. Cross-Validation

Cross-validation is a classical statistical procedure used in different statistical settings. For example, in density estimation or in spline smoothing, cross-validation provides an automatic procedure for choosing the bandwidth or a smoothing parameter. Some general reference are Burman (1989), Silverman (1986) and Green and Silverman(1994). Nason (1996) applied cross-validation to the problem of threshold selection. His method utilized the standard paradigm: minimize the prediction error generated by comparing a prediction, based on part of the data, to the remainder of the data,

$$M(\lambda) = E \int \{\hat{f}_\lambda(x) - f(x)\}^2 dx.$$

We give a brief overview of Nason's two-folded cross-validation procedure. It works by leaving out half the data points, can be used to select a threshold for a wavelet shrinkage estimator based on $n = 2^{J+1}$ points. Let $y_1, y_2, \cdots, y_n$ be the observations. Firstly, take all the evenly indexed data points $\{y_{2j}\}$, $j = 1, \cdots, n/2$, to form a wavelet threshold estimate $\hat{f}^E_\lambda$ using a particular threshold while the remaining points are used to estimate the Mean Integrated Squared Error (MISE) at that threshold. Let $\bar{f}^E_{\lambda,j}$ be an interpolated estimator, defined as

$$\bar{f}^E_{\lambda,j} = \frac{1}{2}(\hat{f}^E_{\lambda,j+1} + \hat{f}^E_{\lambda,j}),$$

where $\hat{f}^E_{\lambda,n/2+1} = \hat{f}^E_{\lambda,1}$ is assumed. The counterpart of the odd-indexed points is

computed to give the interpolant $\bar{f}^{O}_{\lambda,j}$ and the cross-validatory estimate of the mean square error is

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

Nason(1996) showed that one can almost find a unique minimizer of $\hat{M}(\lambda)$ and compared the performance of the cross-validatory threshold to the Donoho-Johnstone universal and SureShrink methods. He also reported that in the case of heavy-tailed noise the described method did not perform well. Wang(1996) addresses the problem in which the noise is correlated and exhibits long-range dependence.

4. Block Thresholding Estimators

Most of the standard wavelet methods achieve adaptivity through term-by-term thresholding of the empirical wavelet coefficients, which either kill or retain an individual coefficient based on its magnitude. Frequentist block thresholding methods shrink wavelet coefficients in groups rather than individually, assuming that information on neighboring coefficients has influence on the treatment of particular coefficients. Simultaneous decisions are made to retain or to discard all the coefficients within a block.

Motivated by the need for spatial adaptivity, Hall et al. (1998,1999) first suggested grouping wavelet coefficients into blocks, modelling them blockwise and exploiting the information that coefficients convey about the size of their nearby neighbour.

Performance of the block thresholding method can be superior to that of its term-by-term counterparts. Block thresholding schemes are shown to reduce the bias and to react more rapidly to sudden frequency changes in the signal. However, it was demonstrated that some block thresholded estimators are more sensitive with respect to the selection of threshold.

2.3 Large and Moderate Deviation Estimates

Large Deviation Estimates

For the mean of n independent and i.i.d. random variables, a deviation, λ_n, is called ordinary, if $\sqrt{n}\lambda_n = O(1)$, excessive, if $n\lambda_n^2 \to \infty$, large, if $\lambda_n = O(1)$, and moderate, if $\lambda_n = c\sqrt{logn/n}$.

Numerous results exist on large and moderate deviations for sums of independent or weakly dependent random variables. However, there are few results on large deviation for moving average sequence $\{\varepsilon_i, i \in \mathbb{Z}\}$ defined in (1.1.6) with long range dependence. We refer to Ghosh and Samorodnitsky (2008) for an excellent overview of the subject and the most recent results under the Cramér condition. Their results show that, among other things, long range dependence changes the large deviations dramatically.

In the following, we state an exponential inequality due to Bentkus and Rudzkis (1980) under the condition (S_γ). To this end, we first recall the definition of cumulant

and its basic properties.

Let ξ be a random variable with characteristic function $f_\xi(t) = E\exp(it\xi)$ and $E|\xi|^m < \infty$. The cumulant of ξ of order m, denoted by $\Gamma_m(\xi)$, is defined by

$$\Gamma_m(\xi) = \frac{1}{i^m}\frac{d^m}{dt^m}\Big(\log f_\xi(t)\Big)\Big|_{t=0}, \tag{2.3.1}$$

where log denote the principal value of the logarithm so that $\log f_\xi(0) = 0$. Note that, under the above assumptions, all cumulants of order not exceeding m exist and

$$\log f_\xi(t) = \sum_{j=1}^{m} \frac{\Gamma_j(\xi)}{j!}(it)^j + o(|t|^m) \qquad \text{as } t \to 0.$$

Cumulants are in general more tractable than moments. For example, if $\xi_1, \ldots, \xi_n$ are independent random variables and if $S_n = \xi_1 + \cdots + \xi_n$, then (2.3.1) implies

$$\Gamma_m(S_n) = \sum_{j=1}^{n} \Gamma_m(\xi_j). \tag{2.3.2}$$

Moreover, if $\eta = a\xi$, where $a \in \mathbb{R}$ is a constant, then $\Gamma_m(\eta) = a^m\Gamma_m(\xi)$. We refer to Petrov (1975) and Saulis and Statulevičius (2000) for further information on cumulants and their applications to limit theory.

The large tail probability estimates of ξ can be described by using information on the cumulants $\Gamma_m(\xi)$. We will make use of the following result due to Bentkus and Rudzkis (1980) [see also Lemma 1.7 and Corollary 1.1 in Saulis and Statulevičius (2000)].

Lemma 2.1. *Let ξ be a random variable with mean 0. If there exist constants $\gamma \geq 0$, $H > 0$ and $\widetilde{\Delta} > 0$ such that*

$$|\Gamma_m(\xi)| \leq \Big(\frac{m!}{2}\Big)^{1+\gamma}\frac{H}{\widetilde{\Delta}^{m-2}}, \qquad \textit{for all } m = 2, 3, \ldots, \tag{2.3.3}$$

then for all $x > 0$,

$$P(|\xi| \geq x) \leq \begin{cases} \exp\Big(-\frac{x^2}{4H}\Big), & \text{if } 0 \leq x \leq (H^{1+\gamma}\widetilde{\Delta})^{1/(1+\gamma)}, \\ \exp\Big(-\frac{1}{4}(x\widetilde{\Delta})^{1/(1+\gamma)}\Big), & \text{if } x \geq (H^{1+\gamma}\widetilde{\Delta})^{1/(1+\gamma)}. \end{cases} \tag{2.3.4}$$

Condition (2.3.3) can be regarded as a generalized Statulevičius condition. It is more general than the celebrated Cramér and Linnik conditions. Recall that a random variable ξ is said to satisfy the Cramér condition if there exists a positive constant a such that

$$E\exp\big(a|\xi|\big) < \infty. \tag{2.3.5}$$

See Petrov (1975, p. 54) for other equivalent formulations of the Cramér condition and its various applications.

A random variable ξ is said to satisfy the Linnik condition if there exist positive constants δ and C_ν such that

$$E\exp\Big(\delta\,|\xi|^{4\nu/(2\nu+1)}\Big) < C_\nu \quad \text{for all } \nu \in \big(0, \frac{1}{2}\big). \tag{2.3.6}$$

Clearly, the Linnik condition is weaker than the Cramér condition. Amosova (2002) has proved that (i) If $\gamma = 0$, then the Statulevičius condition (S_γ) coincides with the Cramér condition; (ii) if $\gamma > 0$, then (S_γ) coincides with the Linnik condition. See Amosova (2002) for the precise relations among the constants γ, Δ, δ and ν in these conditions.

It is also worthwhile to mention the following result of Rudzkis, Saulis and Statulevičius (1978) [see also Lemma 1.8 in Saulis and Statulevičius (2000)]: Let ξ be a random variable that satisfies the following conditions: $E(\xi) = 0$, $E(\xi^2) = \sigma^2$ and

there exist constants $\gamma \geq 0$ and $K > 0$ such that

$$|E(\xi^m)| \leq (m!)^{1+\gamma} K^{m-2}\sigma^2, \qquad m = 3, 4, \ldots. \tag{2.3.7}$$

Then ξ satisfies condition (2.3.3) with $H = 2^{1+\gamma}\sigma^2$ and $\widetilde{\Delta} = \left[2(K \vee \sigma)\right]^{-1}$.

Condition (2.3.7) is a generalization of the classical Bernstein condition: $|E(\xi^m)| \leq \frac{1}{2} m!\, K^{m-2}\sigma^2$ for all $m = 3, 4, \ldots$, which has been used by many authors. For examples, see Petrov (1975, p.55), Johnstone (1999, p.64), Picard and Tribouley (2000, p.301), Zhang and Wong (2003, p.164), among others.

Moderate Deviation Estimates

Moderate deviation results for independent or weakly dependent random variables have been established by Rubin and Sethuraman (1965), Amosova (1972, 1982), Petrov (1975, 2002), Frolov (1998, 2005), Saulis and Statulevičius (2000), Wu and Zhao (2008). The last two articles contain very nice overviews on the topics together with an extensive list on the related references.

We start by recalling a result of Petrov (2002). Let $X_1, \ldots, X_i, \ldots$ be independent random variables such that $E(X_i) = 0$ and $E(|X_i|^{2+\eta}) < \infty$ $(i \geq 1)$ for some $\eta \in (0, 1]$. For any integer $K \geq 1$, denote

$$B_K = \sum_{i=1}^{K} E(X_i^2), \qquad L_K = \frac{1}{B_K^{1+\eta/2}} \sum_{i=1}^{K} E\big(|X_i|^{2+\eta}\big). \tag{2.3.8}$$

In the terminology in Petrov (2002), L_K is called the generalized Lyapunov fraction.

For every $x \in \mathbb{R}$, define

$$F_K(x) = P\left(B_K^{-1/2} \sum_{i=1}^{K} X_i < x \right).$$

Petrov (2002) proved that, if $L_K \to 0$ as $K \to \infty$, then for any constant $C_3 \in (0,1)$ one has

$$\lim_{K\to\infty} \frac{1 - F_K(x)}{1 - \Phi(x)} = \lim_{K\to\infty} \frac{F_K(-x)}{\Phi(-x)} = 1 \tag{2.3.9}$$

uniformly for all $x \in [0, (2C_3 \ln(1/L_K))^{1/2}]$. In the above, $\Phi(x)$ is the distribution function of a standard normal random variable. Frolov (2005) improved the above result under more general conditions.

Chapter 3

Main Results

We consider the nonparametric regression model (1.1.1) with long memory random errors $\{\varepsilon_i\}$ satisfying (1.1.6), (1.1.7) and (1.2.1). The following theorem shows that the wavelet-based estimators defined as in (3.1.4), based on simple thresholding of the empirical wavelet coefficients, attain nearly optimal convergence rates over a large class of functions with discontinuities, where the number of discontinuities is allowed to diverge polynomially fast with sample size. These results show that the discontinuities of the unknown curve have a negligible effect on the performance of nonlinear wavelet curve estimators.

3.1 Function Spaces Considered and Proposed Wavelet Estimators

Common Function Spaces

In accordance with many papers in the wavelet literature, we investigate wavelet-based estimators' asymptotic rates of convergence over a large range of Besov function classes $B^{\sigma}_{p,q}$, $\sigma > 0$, $1 \leq p, q \leq \infty$, which is a very rich class of function space. The parameter σ is an index of regularity or smoothness and parameters p and q are used to specify the type of norm. They include, in particular, the well-known Sobolev space H^m, Hölder spaces C^{σ} of smooth functions, ($B^m_{2,2}$ and $B^{\sigma}_{\infty,\infty}$ respectively), as well as function classes of significant spatial inhomogeneity such as the Bump Algebra and Bounded Variations Classes. For a more detailed study we refer to Triebel (1992).

For a given $r - regular$ mother wavelet ψ with $r > s$, define the sequence norm of the wavelet coefficients of a function $f \in B^s_{p,q}$ by

$$|f|_{B^s_{p,q}} = \left(\sum_k |\alpha_{j_0,k}|^p \right)^{1/p} + \left\{ \sum_{j=j_0}^{\infty} \left[2^{j\sigma} \left(\sum_k |\beta_{j,k}|^p \right)^{1/p} \right]^q \right\}^{1/q},$$

where $\sigma = s + 1/2 - 1/p$. Meyer (1992) showed that the Besov function norm $\| f \|_{B^s_{p,q}}$ is equivalent to the sequence norm $|f|_{B^s_{p,q}}$ of the wavelet coefficients of f. Therefore, we will use the sequence norm to calculate the Besov norm $\| f \|_{B^s_{p,q}}$ in the sequel.

For any constant $M > 0$, define the standard Besov function space $B^s_{p,q}(M)$ by

$$B^s_{p,q}(M) = \left\{ g \in B^s_{p,q} : \| g \|_{B^s_{p,q}} \leq M, 1 \leq p, q \leq \infty, supp g \subseteq [0,1] \right\}.$$

Proposed Wavelet Estimators with Associated Function Spaces

In the regression model (1.1.1), the mean function g is supported on a fixed unit

interval $[0,1]$, thus we assume that ϕ and ψ are compactly supported on $[0,1]$. We also assume that both ϕ and ψ satisfy a uniform Hölder condition of exponent $1/2$, i.e.,

$$|\psi(x) - \psi(y)| \leq C|x-y|^{1/2}, \qquad \text{for all } x,\ y \in [0,1]. \tag{3.1.1}$$

Daubechies (1992, Chap.6) provides examples of wavelets satisfying these conditions.

For a given *r-regular* mother wavelet ψ with $r > \sigma$, the wavelet expansion of $g(x)$ is

$$g(x) = \sum_{k\in\mathbb{Z}} \alpha_{j_0k}\phi_{j_0k}(x) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \beta_{jk}\psi_{jk}(x), \quad x \in [0,1], \tag{3.1.2}$$

where

$$\alpha_{j_0k} = \int_0^1 g(x)\phi_{j_0k}(x)\,dx \qquad \beta_{jk} = \int_0^1 g(x)\psi_{jk}(x)\,dx,$$

and the series in (3.1.2) converges in $L^p([0,1])$.

Let

$$\mathcal{G}^{\sigma}_{\infty,\infty}(M,\,A) = \left\{g :\, g \in B^{\sigma}_{\infty,\infty},\, \|g\|_{B^{\sigma}_{\infty,\infty}} \leq M,\, \|g\|_{\infty} \leq A,\, \operatorname{supp} g \subseteq [0,\,1]\right\},$$

and let $P_{d\tau A}$ be the set of piecewise polynomials $g_2(x)$ of degree $d \leq r-1$, with support contained in $[0,\,1]$, such that the number of discontinuities is no more than τ (for detail, see Theorem 3.1 below in Chapter 3) and the supremum norms of g_2 and g_2' are no more than A. The spaces of mean regression functions we consider in this paper are defined by

$$V_{d\tau A}\{\mathcal{G}^{\sigma}_{\infty,\infty}(M,\,A)\} = \left\{g :\ g = g_1 + g_2;\ g_1 \in \mathcal{G}^{\sigma}_{\infty,\infty}(M,\,A),\ g_2 \in P_{d\tau A}\right\}. \tag{3.1.3}$$

i.e., $V_{d\tau A}\{\mathcal{G}^{\sigma}_{\infty,\infty}(M,\,A)\}$ is a function space in which each element is a mixture of a regular function g_1 from the Besov space $B^{\sigma}_{\infty,\infty}$ with a function g_2 that may pose

discontinuities.

In the statement below, the notation $2^{j(n)} \simeq h(n)$ means that $j(n)$ is chosen to satisfy the inequalities $2^{j(n)} \leq h(n) < 2^{j(n)+1}$.

Our proposed nonlinear wavelet estimator of $g(x)$ is

$$\hat{g}(x) = \sum_{k\in\mathbb{Z}} \hat{\alpha}_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{j_1}\sum_{k\in\mathbb{Z}} \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > \delta_j)\psi_{jk}(x), \tag{3.1.4}$$

where

$$\hat{\alpha}_{j_0k} = \frac{1}{n}\sum_{i=1}^{n} Y_i\phi_{j_0k}(x_i), \qquad \hat{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{jk}(x_i), \tag{3.1.5}$$

$I(E)$ is the indicator random variable of the event E and the smoothing parameters j_0, j_1 are chosen to satisfy $2^{j_0} \simeq \log_2 n$ and $2^{j_1} \simeq n^{1-\pi}$ for some constant $\pi > 0$ (We will choose $\pi < 0.75(2r+1)^{-1}$ in our main theorem below. Also for notational convenience, we will suppress the subscript n for j_0 and j_1).

In (3.1.4), the threshold δ_j is level j dependent. We will choose

$$\delta_j^2 = 2^{3+\gamma} C_2 n^{-\alpha} 2^{-j(1-\alpha)} \ln n$$

if the condition (S_γ) is assumed; and $\delta_j^2 = C_2\pi\eta(1-\alpha)\, n^{-\alpha}2^{-j(1-\alpha)} \ln n$ under the moment condition $E(|\zeta_1|^{2+\eta}) < \infty$. In the above γ is the constant in (1.2.1), α is the long memory parameter in (1.1.5) and $C_2 = C_0 \iint |x-y|^{-\alpha}\psi(x)\psi(y)\, dxdy$.

3.2 Main Theorems

Theorem 3.1. *Suppose the wavelet ψ is r-regular. Our wavelet estimator $\hat{g}$ is defined as in (3.1.4) with $\pi < 0.75(2r+1)^{-1}$ and $\delta_j^2 = 2^{3+\gamma} C_2 n^{-\alpha} 2^{-j(1-\alpha)} \ln n$. Let τ_n be any*

sequence of positive numbers that satisfy $\tau_n = O(n^{\theta+0.25\alpha(2r+1)^{-1}})$, *where* $\theta > 0$ *is a small constant such that*

$$\theta + 0.25\alpha(2r+1)^{-1} < \min\{1/2, \alpha/(2r+\alpha)\}.$$

Then, for any constants $A, M \in (0,\infty)$ *and* $\sigma \in [1/2, r)$, *there exists a constant* $C > 0$ *such that*

$$\sup_{d<r,\, \tau\le\tau_n} \sup_{g\in V_{d\tau A}\{\mathcal{G}^{\sigma}_{\infty,\infty}(M,A)\}} E\int_0^1 \left[\hat{g}(x) - g(x)\right]^2 dx \le C\, n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n.$$

Remark 3.2. The above wavelet estimators $\hat{g}$ do not depend on the unknown parameters σ and d. However, because of the long-range dependence nature, our thresholds δ_j $(= \lambda\sigma_j)$ must be level-dependent and our estimators depend on the unknown long memory parameter α. Wang (1996, p.480) and Johnstone and Silverman (1997, p.340) provided simple methods to estimate the long memory parameter α. So, in practice, one needs to estimate the long memory parameter before applying the wavelet method. In this paper, we treat it as known. Our thresholds $\delta_j = \lambda\sigma_j = \sqrt{2^{3+\gamma}\ln n}\,\sigma_j$ (for details, see Lemma 3.7 below) are similar to the standard term-by-term hard threshold $\delta = \sqrt{2\ln n}\,\sigma$ in the Gaussian case. However, because of the long memory and non-Gaussian errors here, one needs a bigger constant $2^{3+\gamma}$ instead of 2. One alternative to deal with the unknown parameter α is to use the robust median absolute deviation estimates to replace σ_j. The performance of the estimators with this replacement is under investigation by the authors for future research.

Remark 3.3. Minimax theory indicates that the best convergence rate over the func-

tion space $\mathcal{G}^{\sigma}_{\infty,\infty}(M,\ A)$ is at most $n^{-2\sigma\alpha/(2\sigma+\alpha)}$. Since $\mathcal{G}^{\sigma}_{\infty,\infty}(M,\ A) \subseteq V_{d\tau A}\{\mathcal{G}^{\sigma}_{\infty,\infty}(M,\ A)\}$, the above estimators achieve optimal convergence rates up to a logarithmic term, without knowing the smoothness parameter. From Wang (1996, p470), the traditional linear estimators which include kernel estimators can not achieve the rates stated in Theorem 3.1. Hence our non-linear wavelet estimators achieve nearly optimal convergence rates over a large function space.

Remark 3.4. Wang (1996) and Johnstone and Silverman (1997) considered wavelet estimators of regression functions in the wavelet domain or based on the so-called "sequence space model" with Gaussian error. For details, see Johnstone and Silverman (1997). Based on the asymptotic equivalence between "sequence space model" and "sampled data model" (1.1.1), they derived the minimax optimal convergence rates of wavelet estimators in the wavelet domain. Here, we consider the wavelet estimator in the time domain or directly based on the "sampled data model" (1.1.1) as in Hall, *et al.* (1999).

3.3 Key Lemmas

When studying moderate deviation of the moving average sequence defined in (1.1.6), the conditions $L_K \to 0$ as $K \to \infty$ is not satisfied. [In fact L_K in (5.2.4) will be bounded from below by a positive constant (depending on n) and $B_K \to 1$ as $K \to \infty$.] In order to prove an analogue of Lemma 3.7 below under the moment condition $E(|\zeta_1|^{2+\eta}) < \infty$ for some constant $\eta > 0$, we will make use of the following tail probability estimate, which is essentially implied by the proof of Theorem 1.1 in

Frolov (2005) [see his Remark 1.2].

Lemma 3.5. *Let $\{X_i, i \geq 1\}$ be a sequence of independent random variables such that $E(X_i) = 0$ and $E(|X_i|^{2+\eta}) < \infty$ $(i \geq 1)$ for some $\eta > 0$. Let B_K and L_K be defined as in (2.3.8). If for some constant $C_4 > 0$ such that*

$$\Big(2\ln(1/L_K)\Big)^{2+\frac{5\eta}{2}} L_K \leq C_4 \tag{3.3.1}$$

for all $K \geq 1$, then there exists a finite constant C_5 such that

$$P\Big(B_K^{-1/2}\Big|\sum_{i=1}^{K} X_i\Big| > x\Big) \leq C_5\big(1 - \Phi(x)\big) \tag{3.3.2}$$

for all $x \in [1, (2\ln(1/L_K))^{1/2}]$.

Proof: It is sufficient to prove that

$$1 - F_K(x) \leq C_5\big(1 - \Phi(x)\big) \tag{3.3.3}$$

for all $x \in [1, (2\ln(1/L_K))^{1/2}]$. The method of proving (3.3.3) is similar to that of Theorem 1.1 in Frolov (2005). The most important differences between the conditions in Lemma 3.5 above and Theorem 1.1 in Frolov (2005) are that we do not assume $L_K \to 0$ as $K \to \infty$ nor the condition (1.2) in Frolov (2005). These later conditions are essential for proving $1 - F_K(x) \sim 1 - \Phi(x)$ as $K \to \infty$, but are not necessary for deriving the upper bound in (3.3.3).

Since the complete proof of Theorem 1.1 in Frolov (2005) is rather long and it seems unnecessary to reproduce it here, we will only provide a sketch of the proof of (3.3.3) and point out the modifications that we need to make.

For any $x \in [1, (2\ln(1/L_K))^{1/2}]$ and a fixed constant $\rho \in (0, 1/2)$, set $\ell_K = \rho x\sqrt{B_K}$. As in Frolov (2005), we define the truncated random variables $Y_{K,i} = X_i I\big(X_i \le \ell_K\big)$ for $1 \le i \le K$ and put $T_K = \sum_{i=1}^K Y_{K,i}$. Note that

$$P\Big(\sum_{i=1}^K X_i > x\sqrt{B_K}\Big) \le P\Big(T_K > x\sqrt{B_K}\Big) + \sum_{i=1}^K P(X_i > \ell_K). \tag{3.3.4}$$

For the last term on the right-hand side of (3.3.4), we have

$$\sum_{i=1}^K P(X_i > \ell_K) \le \ell_K^{-(2+\eta)} \sum_{i=1}^K E\big(|X_i|^{2+\eta}\big) = (\rho x)^{-(2+\eta)} L_K. \tag{3.3.5}$$

Since the function $f(x) = x^{-(1+\eta)}e^{x^2/2}$ is increasing for $x > (1+\eta)^{1/2}$, we can argue as in Frolov (2005, pp. 1794–1795) to show that

$$(\rho x)^{-(2+\eta)} L_K\, x e^{x^2/2} = \rho^{-(2+\eta)} L_K\, x^{-(1+\eta)} e^{x^2/2} \le C$$

for all $x \in [1, (2\ln(1/L_K))^{1/2}]$, where C is a finite constant depending on ρ and η only. Consequently,

$$\sum_{i=1}^K P(X_i > \ell_K) \le C\big(1 - \Phi(x)\big) \tag{3.3.6}$$

for all $x \in [1, (2\ln(1/L_K))^{1/2}]$.

In order to bound the first term on the right-hand side of (3.3.4), similarly to Frolov (2005), we introduce independent random variables $\{\overline{Y}_{K,i}, i = 1, 2, \ldots, K\}$ with distributions functions

$$P(\overline{Y}_{K,i} \le z) = \frac{1}{\varphi_{K,i}} \int_{-\infty}^{z} e^{xy/\sqrt{B_K}}\, dP(Y_{K,i} \le y), \tag{3.3.7}$$

where $\varphi_{K,i} = E\big(e^{xY_{K,i}/\sqrt{B_K}}\big)$. Let $\overline{T}_K = \sum_{i=1}^K \overline{Y}_{K,i}$,

$$\overline{M}_K = E(\overline{T}_K) \qquad \text{and} \qquad \overline{B}_K = \mathrm{Var}(\overline{T}_K).$$

Instead of condition (1.2) in Frolov (2005), we show that our condition (3.3.1) implies that

$$\frac{x^4}{B_K}\sum_{i=1}^{K} E\left[X_i^2 I\left(|X_i| \geq \frac{\sqrt{B_K}}{x^5}\right)\right] \leq C_4 \tag{3.3.8}$$

for all $x \in [1, (2\ln(1/L_K))^{1/2}]$. In fact, by using Hölder's inequality we see that the left-hand side of (3.3.8) is at most

$$\begin{aligned}
&\frac{x^4}{B_K}\sum_{i=1}^{K}\left[E\left(X_i^{2+\eta}\right)\right]^{\frac{2}{2+\eta}}\left[P\left(|X_i| \geq \frac{\sqrt{B_K}}{x^5}\right)\right]^{\frac{\eta}{2+\eta}} \\
&\leq \frac{x^4}{B_K}\sum_{i=1}^{K}\left[E\left(X_i^{2+\eta}\right)\right]^{\frac{2}{2+\eta}}\left[\frac{x^{5(2+\eta)}\,E\left(X_i^{2+\eta}\right)}{B_K^{(2+\eta)/2}}\right]^{\frac{\eta}{2+\eta}} \\
&= x^{4+5\eta}\,L_K.
\end{aligned} \tag{3.3.9}$$

Hence (3.3.8) holds thanks to (3.3.1).

Note that, under (3.3.8), the proof of Theorem 1.1 in Frolov (2005), with $o(\cdot)$ being replaced by $O(\cdot)$, continues to work. In particular, it follows from (3.12), (3.13) and (3.16) in Frolov (2005) that

$$\begin{aligned}
P\left(T_K \geq x\sqrt{B_K}\right) &\leq \left(\prod_{i=1}^{K}\varphi_{K,i}\right) e^{-x\overline{M}_K/B_K}\int_{\frac{x\sqrt{B_K}-\overline{M}_K}{\sqrt{\overline{B}_K}}}^{\infty} e^{-xy\sqrt{\overline{B}_K/B_N}}\,dG_K(y) \\
&\leq C\,x^{-1}e^{-x^2/2}
\end{aligned} \tag{3.3.10}$$

for all $x \in [1, (2\ln(1/L_K))^{1/2}]$. In the above, $G_K(y) = P\left(\overline{T}_K < y\sqrt{\overline{B}_K} + \overline{M}_K\right)$. Combining (3.3.6) and (3.3.10) yields (3.3.3). This finishes the proof of Lemma 3.5.

Lemma 3.6. *Suppose that the wavelets ϕ and ψ satisfy the uniform Hölder condition (3.1.1) and let $A, M \in (0,\infty)$ and $\sigma \in [1/2, r)$ be constants. Then for all j_0 and j, we have the following results about the approximation between empirical wavelet coefficients and the true wavelet coefficient(see 5.1.1).*

$$\sup_k |a_{j_0k} - \alpha_{j_0k}| = O\left(n^{-1/2} + \tau n^{-1}\right), \tag{3.3.11}$$

$$\sup_k |b_{jk} - \beta_{jk}| = O\big(n^{-1/2} + \tau n^{-1}\big) \tag{3.3.12}$$

hold uniformly for all mean regression functions g as in (3.1.3).

Lemma 3.7. *Under the assumptions of Theorem 3.1, there exists a positive constant C such that*

$$P\Big(\Big|\hat{\beta}_{jk} - b_{jk}\Big| > \delta_j\Big) \le C\, n^{-1}, \quad \forall j \in [j_0, j_1] \ \text{ and } \ k = 0, 1, \cdots, 2^j - 1. \tag{3.3.13}$$

3.4 A Further Extension

So far we have assumed that the innovation process $\{\zeta_j, j \in \mathbb{Z}\}$ satisfies the Statulevičius condition (S_γ) given by (1.2.1). This condition can be weakened if one is willing to change the threshhold δ_j accordingly.

The following result shows that the conclusion of Theorem 3.1 still holds under the condition $E(|\zeta_1|^{2+\eta}) < \infty$ for some constant $\eta > 0$.

Theorem 3.8. *Suppose that the wavelet ψ is r-regular. The wavelet estimator $\hat{g}$ is defined as in (3.1.4) with $\pi < 0.75(2r+1)^{-1}$ and $\delta_j^2 = C_2\pi\eta(1-\alpha)n^{-\alpha}2^{-j(1-\alpha)} \ln n$. We assume that*

$$\eta\pi(1-\alpha) \ge 2. \tag{3.4.1}$$

Let τ_n be any sequence of positive numbers such that for all $\theta > 0$, $\tau_n = O(n^{\theta + 0.25\alpha(2r+1)^{-1}})$, where $\theta > 0$ is a small constant such that

$$\theta + 0.25\alpha(2r+1)^{-1} < \min\{1/2, \alpha/(2r+\alpha)\}.$$

Then for any constants $A, M \in (0, \infty)$ *and* $\sigma \in [1/2, r)$ *there exists a constant* $C > 0$ *such that*

$$\sup_{d<r,\, \tau \le \tau_n} \sup_{g \in V_{d\tau A}\{\mathcal{G}^{\sigma}_{\infty,\infty}(M,A)\}} E \int_0^1 \left[\hat{g}(x) - g(x)\right]^2 dx \le C\, n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n.$$

Chapter 4

Simulation

To investigate the performance of the proposed wavelet estimator, we present a modest simulation study. We generate Y_i's data according to regression model $Y_i = g(x_i) + \varepsilon_i$, where $x_i = i/n$, $i = 1, 2, \cdots, n$, and n is the sample size. Regression function $g(x)$ is a piece-wise HeaviSine function:

$$g(x) = \begin{cases} \cos(4\pi x) - 3, & \text{if} \quad 0 \leq x < 0.3; \\ \cos(4\pi x) + 1, & \text{if} \quad 0.3 \leq x < 0.7; \\ \cos(4\pi x) + 5, & \text{if} \quad 0.7 \leq x \leq 1. \end{cases}$$

It can be seen that $g \in V_{d\tau A}\{\mathcal{G}^{\sigma}_{\infty,\infty}(M,\ A)\}$ with $d = 0$, $\tau = 3$ and $\sigma = A = 1$.

We use S-Plus function *arima.fracdiff.sim* to generate random errors ε_i, which are a Gaussian *FARIMA(0,d,0)* series with fractional difference parameters $0 < d < 0.5$. From Beran (1994), we have $d = (1-\alpha)/2$ or $\alpha = 1-2d$, where α is our long memory parameter. In order to investigate the effect of the long memory parameter α on the performance of our estimator, in this simulation study, we consider parameter

d with values of $0.05, 0.10, 0.15, ..., 0.45$, which are equivalent to α with values of $0.9, 0.8, 0.7, ..., 0.1$. We consider four different sample sizes: $n = 128,\ 256,\ 512$ and 1024. For numerical comparisons we consider the average norm (AverNorm) of the estimators at the sample points

$$AveNorm = \frac{1}{N}\sum_{l=1}^{N}\left[\sum_{i=1}^{n}\left(\hat{g}_l(x_i) - g(x_i)\right)^2\right]^{1/2},$$

where $\hat{g}_l$ is the estimate of g in l-th replication and N is the total number of replications. Since different wavelets yield very similar results, we only use Daubechies's compactly support wavelet *Symmlet* 8. Note that for Gaussian errors ε_i, we can use level dependent thresholds $\delta_j = \sqrt{2\ln n}\,\hat{\sigma}_j$, where $j = 1, 2, ..., \log_2 n - 1$ and $\hat{\sigma}_j$ is an estimate of scale of noise σ_j from empirical wavelet coefficients in level j using the median of absolute deviation from the median from level to level. The simulation results for different sample sizes n and different long memory parameters α are summarized in Table 4.1 and Figure 1. Based on these finite simulation studies, we see that our empirical results (Average Norms) are consistent with our theoretic results, i.e., Average Norm is a decreasing function of α for all different sample sizes.

Table 4.1: Average Norm from $N = 500$ replications.

	α								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
128	19.923	14.884	12.829	11.915	11.066	10.320	10.054	9.764	9.506
256	27.342	20.533	17.457	15.574	14.476	13.480	12.647	12.380	11.612
512	35.788	26.137	21.357	18.151	16.079	14.508	13.128	12.404	11.495
1024	51.243	35.121	28.020	23.655	20.166	17.927	15.946	14.527	13.147

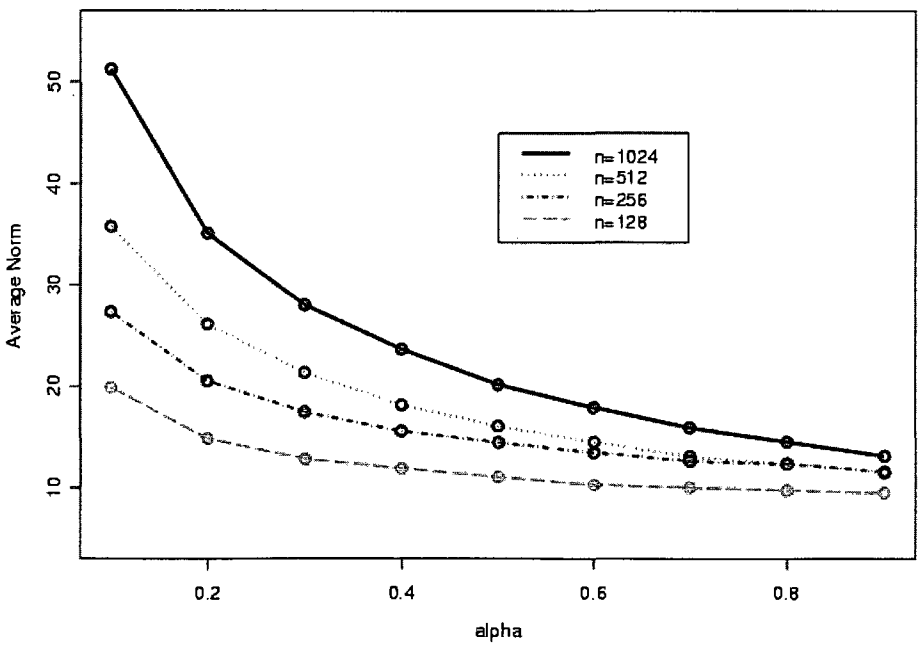

Figure 4.1: The Average Norms of the estimators with different alpha values and sample sizes

Chapter 5

Proofs of Main Theorems

5.1 Proof of Theorem 3.1

The overall proof of Theorem 3.1 is motivated by the arguments of Donoho, *et al.* (1996) and Hall, *et al.* (1998, 1999) for the independent data case. But moving from independent data to long range dependent data, especially non-Gaussian random errors, involves a significant change in complexity. For nonparametric regression model with Gaussian random errors or for density estimation with i.i.d. random variables, one can apply respectively the Gaussian isoperimetric inequality or the standard Bernstein inequality to obtain an exponential bound. However, these techniques are not readily applicable to infinite moving average processes with long memory. The key technical ingredient in our proof is to apply the large deviation estimate (Lemma 2.1) to establish an exponential inequality for a sequence of infinite weighted sums of i.i.d. random variables (For details, see Lemma 3.7).

Proof of Theorem 3.1: The proof of Theorem 3.1 can be broken into several parts. Observing that the orthogonality (2.1.2) of ϕ and ψ implies

$$E\int_0^1 \left[\hat{g}(x) - g(x)\right]^2 dx =: I_1 + I_2 + I_3 + I_4,$$

where

$$I_1 = \sum_k E\left(\hat{\alpha}_{j_0k} - \alpha_{j_0k}\right)^2, \qquad I_2 = \sum_{j=j_0}^{j_\sigma}\sum_k E\left(\hat{\theta}_{jk} - \beta_{jk}\right)^2,$$

$$I_3 = \sum_{j=j_\sigma+1}^{j_1}\sum_k E\left(\hat{\theta}_{jk} - \beta_{jk}\right)^2, \quad I_4 = \sum_{j=j_1+1}^{\infty}\sum_k \beta_{jk}^2.$$

Here $\hat{\theta}_{jk} = \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > \delta_j)$ and $j_\sigma = j_\sigma(n)$ such that $2^{j_\sigma} \simeq \left(n^{-1}\log_2 n\right)^{-\alpha/(2\sigma+\alpha)}$. In order to prove Theorem 3.1, it suffices to show that $I_i \leq Cn^{-2\sigma\alpha/(2\sigma+\alpha)}\log_2 n$, $i = 1, \cdots, 4$, for all d, τ, σ, A, M. These inequalities are shown in Lemmas 5.4 to 5.7, respectively.

We start by collecting the lemmas. Denote

$$\begin{aligned} a_{j_0k} &:= E(\hat{\alpha}_{j_0k}) = \frac{1}{n}\sum_{i=1}^n g(x_i)\phi_{j_0k}(x_i), \\ b_{jk} &:= E(\hat{\beta}_{jk}) = \frac{1}{n}\sum_{i=1}^n g(x_i)\psi_{jk}(x_i). \end{aligned} \tag{5.1.1}$$

Since we consider nonparametric regression with discontinuities on the sample data model, unlike the density estimation problem as in Hall, *et al.* (1998), one more step of approximation between empirical wavelet coefficients and true wavelet coefficients is needed. Lemma 3.6 which estimates the discrepancy between them will be used for proving the other lemmas.

Proof of Lemma 3.6: We only prove (3.3.11). The proof of (3.3.12) is similar and is omitted.

Let $p = 2^{j_0}$, we may write

$$a_{j_0k} = \frac{p^{1/2}}{n} \sum_{i=1}^{n} g\Big(\frac{i}{n}\Big)\phi\Big(\frac{pi}{n} - k\Big). \tag{5.1.2}$$

For fixed n, p and k, we note that

$$0 \le \frac{pi}{n} - k \le 1 \quad \text{if and only if} \quad \frac{nk}{p} \le i \le \frac{n(k+1)}{p}.$$

Let $m_k = \lfloor \frac{nk}{p} \rfloor$, where $\lfloor x \rfloor$ denotes the smallest integer that is at least x. Since ϕ has its support in $[0,1]$, the summation in (5.1.2) runs from m_k to m_{k+1}. However, for simplicity of the notation, we will not distinguish between $\lfloor x \rfloor$ and x. Let $i = m_k + \ell$ in (5.1.2), we have

$$\begin{aligned} a_{j_0k} &= \frac{p^{1/2}}{n} \sum_{\ell=0}^{n/p-1} g(\frac{\ell}{n} + \frac{k}{p})\phi(\frac{p\ell}{n}) \qquad (\text{let } \ t_\ell = \frac{p\ell}{n}) \\ &= \frac{1}{p^{1/2}} \sum_{\ell=0}^{n/p-1} g(\frac{t_\ell + k}{p})\phi(t_\ell)\,\frac{p}{n}. \end{aligned} \tag{5.1.3}$$

Similarly, by a simple change of variables, we have

$$\begin{aligned} \alpha_{j_0k} &= p^{1/2} \int_{k/p}^{(k+1)/p} g(x)\,\phi(px-k)\,dx \qquad (\text{let } \ t = px - k) \\ &= \frac{1}{p^{1/2}} \int_0^1 g(\frac{t+k}{p})\,\phi(t)\,dt. \end{aligned} \tag{5.1.4}$$

Combining (5.1.3) and (5.1.4), we have

$$\begin{aligned} a_{j_0k} - \alpha_{j_0k} &= \frac{1}{p^{1/2}} \sum_{\ell=0}^{n/p-1} \int_{\frac{p\ell}{n}}^{\frac{p(\ell+1)}{n}} \Big[g\Big(\frac{t_\ell + k}{p}\Big)\phi(t_\ell) - g\Big(\frac{t+k}{p}\Big)\,\phi(t)\Big]\,dt \\ &= J_1 + J_2, \end{aligned} \tag{5.1.5}$$

where

$$J_1 = \frac{1}{p^{1/2}} \sum_{\ell=0}^{n/p-1} \int_{\frac{p\ell}{n}}^{\frac{p(\ell+1)}{n}} \Big[g\Big(\frac{t_\ell + k}{p}\Big) - g\Big(\frac{t+k}{p}\Big)\Big]\phi(t_\ell)\,dt$$

and

$$J_2 = \frac{1}{p^{1/2}} \sum_{\ell=0}^{n/p-1} \int_{\frac{p\ell}{n}}^{\frac{p(\ell+1)}{n}} g\Big(\frac{t+k}{p}\Big)\Big[\phi(t_\ell) - \phi(t)\Big]\, dt.$$

Let us consider the term J_1 first. Since $g = g_1 + g_2$ with $g_1 \in \mathcal{G}^{\sigma}_{\infty,\infty}(M, A)$ and $g_2 \in P_{d\tau A}$, we can write $J_1 = J_{1,1} + J_{1,2}$, where

$$J_{1,j} = \frac{1}{p^{1/2}} \sum_{\ell=0}^{n/p-1} \int_{\frac{p\ell}{n}}^{\frac{p(\ell+1)}{n}} \Big[g_j\Big(\frac{t_\ell + k}{p}\Big) - g_j\Big(\frac{t+k}{p}\Big)\Big]\phi(t_\ell)\, dt, \qquad j = 1, 2.$$

Since $g_1 \in \mathcal{G}^{\sigma}_{\infty,\infty}(M, A)$, $\sigma \geq 1/2$ and ϕ is bounded on $[0, 1]$, we have

$$\big|J_{1,1}\big| \leq \frac{1}{p^{1/2}} \sum_{\ell=0}^{n/p-1} \int_{\frac{p\ell}{n}}^{\frac{p(\ell+1)}{n}} C\Big(\frac{|t - t_\ell|}{p}\Big)^{\sigma} dt \leq \frac{1}{p^{1/2}} \cdot C\Big(\frac{1}{n}\Big)^{\sigma} \leq Cn^{-1/2}. \qquad (5.1.6)$$

Since $g_2 \in P_{d\tau A}$, it is piecewise polynomial and has at most τ discontinuities. Moreover, g_2 is bounded on $[0, 1]$ and is Lipschitz on every open subinterval of $[0, 1]$ where g_2 is continuous. For simplicity, we will assume that each interval $(\frac{p\ell}{n}, \frac{p(\ell+1)}{n})$ contains at most one discontinuity of the function $g_2\big(\frac{\cdot+k}{p}\big)$. This reduction, which brings some convenience for presenting our proof, is not essential and the same argument remains true if an interval contains more discontinuities.

If $(\frac{p\ell}{n}, \frac{p(\ell+1)}{n})$ contains no discontinuity of $g_2\big(\frac{\cdot+k}{p}\big)$, then by the Lipschitz condition we have

$$\int_{\frac{p\ell}{n}}^{\frac{p(\ell+1)}{n}} \Big|g_2\Big(\frac{t_\ell + k}{p}\Big) - g_2\Big(\frac{t+k}{p}\Big)\Big|\,\big|\phi(t_\ell)\big|\, dt \leq C\,\frac{p}{n^2}. \qquad (5.1.7)$$

If $(\frac{p\ell}{n}, \frac{p(\ell+1)}{n})$ contains one discontinuity, say t_0, of $g_2\big(\frac{\cdot+k}{p}\big)$, then we will split the integral in (5.1.7) over $(\frac{p\ell}{n}, t_0)$ and $(t_0, \frac{p(\ell+1)}{n})$. Since the values of the integrals remain the same if we modify the values of the function $g_2\big(\frac{\cdot+k}{p}\big)$ at the end-points of the intervals, we may assume that $g_2\big(\frac{\cdot+k}{p}\big)$ are polynomials on the closed intervals $[\frac{p\ell}{n}, t_0]$ and $[t_0, \frac{p(\ell+1)}{n}]$. Hence the triangle inequality and Lipschitz condition imply

that the integral in (5.1.7) is bounded above by a constant multiple of

$$\int_{\frac{p\ell}{n}}^{t_0} \Big|g_2\Big(\frac{t_\ell+k}{p}\Big) - g_2\Big(\frac{t+k}{p}\Big)\Big|\,dt + \int_{t_0}^{\frac{p(\ell+1)}{n}} \Big|g_2\Big(\frac{t_\ell+k}{p}\Big) - g_2\Big(\frac{t_0+k}{p}\Big)\Big|\,dt$$
$$+ \int_{t_0}^{\frac{p(\ell+1)}{n}} \Big|g_2\Big(\frac{t_0+k}{p}\Big) - g_2\Big(\frac{t+k}{p}\Big)\Big|\,dt \le C\,n^{-1}\Big(\int_{\frac{p\ell}{n}}^{t_0} dt + 2\int_{t_0}^{\frac{p(\ell+1)}{n}} dt\Big). \tag{5.1.8}$$

Summing up (5.1.7) and (5.1.8) over $\ell = 0, 1, \ldots, n/p - 1$ and recall that there are τ discontinuities, we obtain

$$\big|J_{1,2}\big| \le \frac{1}{p^{1/2}} \cdot C\,(1+\tau)\,n^{-1} \le C\,(1+\tau)\,n^{-1}. \tag{5.1.9}$$

As to the second term J_2, we use the boundedness of g and the uniform 1/2-Hölder condition (3.1.1) for ϕ to derive

$$\big|J_2\big| \le \frac{1}{p^{1/2}} \cdot C\left(\frac{p}{n}\right)^{1/2} = C\,n^{-1/2}. \tag{5.1.10}$$

It is clear that (3.3.11) follows from (5.1.5), (5.1.6), (5.1.9) and (5.1.10).

Remark 5.1. If we write

$$\begin{aligned}\alpha_{jk} &= \int_0^1 g(x)\phi_{jk}(x)dx = \int_0^1 g_1(x)\phi_{jk}(x)dx + \int_0^1 g_2(x)\phi_{jk}(x)dx \\ &= \alpha_{jk,1} + \alpha_{jk,2},\end{aligned} \tag{5.1.11}$$

similarly for $a_{jk,1}$ and $a_{jk,2}$, then Lemma 3.6 shows that $\sup_k |a_{jk,1} - \alpha_{jk,1}| = O(n^{-1/2})$ and $\sup_k |a_{jk,2} - \alpha_{jk,2}| = O(n^{-1/2} + \tau n^{-1})$. Furthermore, if the number of the jump discontinuities $\tau \le \tau_n = O(n^{1/2})$, then $\sup_k |a_{jk,1} - \alpha_{jk,1}| = O(n^{-1/2})$ and $\sup_k |a_{jk,2} - \alpha_{jk,2}| = O(n^{-1/2})$. Similar results hold for β_{jk} and b_{jk}.

Let's restate Lemma 3.7 and provide the proof here.

Lemma 5.2. *Under the assumptions of Theorem 3.1, there exists a positive constant C such that*

$$P\Big(\Big|\hat{\beta}_{jk} - b_{jk}\Big| > \delta_j\Big) \le C\, n^{-1}, \quad \forall j \in [j_0,\, j_1] \ \text{ and } \ k = 0, 1, \cdots, 2^j - 1. \tag{5.1.12}$$

Proof: First let's calculate $E(\hat{\beta}_{jk} - b_{jk})^2$. ¿From (3.1.5) and (5.1), we have

$$\begin{aligned} E(\hat{\beta}_{jk} - b_{jk})^2 &= \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} E\big(\varepsilon_{i_1}\varepsilon_{i_2}\big)\psi_{jk}(x_{i_1})\psi_{jk}(x_{i_2}) \\ &= \frac{2^j}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} r(i_1 - i_2)\psi(2^j x_{i_1} - k)\psi(2^j x_{i_2} - k). \end{aligned}$$

For each fixed $k = 0,\ 1,\ \cdots,\ 2^j - 1$, similar to (5.1.3), we have

$$\begin{aligned} E(\hat{\beta}_{jk} - b_{jk})^2 &= \frac{2^j}{n^2} \sum_{i_1=1}^{n2^{-j}-1} \sum_{i_2=1}^{n2^{-j}-1} r(i_1 - i_2)\psi\Big(\frac{i_1 2^j}{n}\Big)\psi\Big(\frac{i_2 2^j}{n}\Big) \\ &= 2^{-j} C_0 \big(2^j n^{-1}\big)^{\alpha} \left[\int_0^1 \int_0^1 |x - y|^{-\alpha} \psi(x)\psi(y)\, dxdy + o(1) \right], \end{aligned}$$

where the last equality follows from (1.1.5) and a standard limiting argument.

Recall that $\delta_j^2 = 2^{3+\gamma} C_2 n^{-\alpha} 2^{-j(1-\alpha)} \ln n$ in (3.1.4). Let $\sigma_j^2 = C_2 n^{-\alpha} 2^{-j(1-\alpha)}$ and $\lambda = 2\sqrt{2^{1+\gamma} \ln n}$, then we have $\delta_j^2 = \lambda^2 \sigma_j^2$. From the above calculation, we see that $E(\hat{\beta}_{jk} - b_{jk})^2 \sim \sigma_j^2$. In view of (3.1.5), (5.1.1) and (1.1.6), we may write $\hat{\beta}_{jk} - b_{jk}$ as an infinite weighted sum of independent random variables $\{\zeta_j,\, j \in \mathbb{Z}\}$:

$$\hat{\beta}_{jk} - b_{jk} = n^{-1} \sum_{i=1}^{n} \varepsilon_i \psi_{jk}(x_i) =: \sum_{s \in \mathbb{Z}} d_{n,s} \zeta_s, \tag{5.1.13}$$

where

$$d_{n,s} = \begin{cases} n^{-1} \sum_{i=1}^{n} b_{i-s} \psi_{jk}(x_i), & \text{if } s \le 0; \\ n^{-1} \sum_{i=s}^{n} b_{i-s} \psi_{jk}(x_i), & \text{if } 0 < s \le n; \\ 0, & \text{otherwise.} \end{cases}$$

Hence, we have $\sum_{s\in\mathbb{Z}} d_{n,s}^2 = E(\hat{\beta}_{jk} - b_{jk})^2 \sim \sigma_j^2$.

For any positive integers n and K, we define

$$S_n = \sum_{s\in\mathbb{Z}} \sigma_j^{-1}\, d_{n,s}\, \zeta_s \quad \text{and} \quad S_{n,K} = \sum_{|s|\le K} \sigma_j^{-1}\, d_{n,s}\, \zeta_s. \tag{5.1.14}$$

Then, as $K \to \infty$, $S_{n,K} \longrightarrow S_n$ almost surely for all integers n. Note that $E(S_{n,K}) = 0$ and, by (2.3.2) and (1.2.1), we have that for all integers $m \ge 3$,

$$\left|\Gamma_m(S_{n,K})\right| = \Big| \sum_{|s|<K} \Big(\frac{d_{n,s}}{\sigma_j}\Big)^m \Gamma_m(\zeta_s)\Big| \le \sum_{|s|<K} \Big|\frac{d_{n,s}}{\sigma_j}\Big|^m \frac{(m!)^{1+\gamma}}{\Delta^{m-2}}. \tag{5.1.15}$$

By using (1.1.7), the Cauchy-Schwarz inequality and the fact that $n^{-1}\sum_{i=1}^n \psi_{jk}^2(x_i)$ $\to 1$, we have

$$\sup_{s\in\mathbb{Z}} d_{n,s}^2 \le C\, n^{-1} \sum_{i=1}^{n} i^{-(1+\alpha)} \le C\, n^{-1}$$

for some finite constant $C > 0$. This implies

$$\sup_{s\in\mathbb{Z}} \frac{d_{n,s}^2}{\sigma_j^2} \le C\,(n^{-1}2^j)^{1-\alpha}. \tag{5.1.16}$$

It follows from (5.1.16) that

$$\begin{aligned} \sum_{|s|<K} \Big|\frac{d_{n,s}}{\sigma_j}\Big|^m &\le \sup_{|s|<K} \Big(\frac{d_{n,s}^2}{\sigma_j^2}\Big)^{(m-2)/2} \cdot \sum_{|s|<K} d_{n,s}^2 \sigma_j^{-2} \\ &\le \Big(C\,(n^{-1}2^j)^{(1-\alpha)/2}\Big)^{m-2}. \end{aligned} \tag{5.1.17}$$

Combining (5.1.15) and (5.1.17) yields

$$\left|\Gamma_m(S_{n,K})\right| \le \Big(\frac{m!}{2}\Big)^{1+\gamma} \frac{2^{1+\gamma}}{\left[C^{-1}\,\Delta\,(n\,2^{-j})^{(1-\alpha)/2}\right]^{m-2}}, \qquad \forall m = 3, 4, \ldots. \tag{5.1.18}$$

That is, $S_{n,K}$ satisfies the condition (2.3.3) with $H = 2^{1+\gamma}$ and $\widetilde{\Delta} = C^{-1}\,\Delta\,(n\,2^{-j})^{(1-\alpha)/2}$.

Since $2^{j_1} \simeq n^{1-\pi}$, we have $\widetilde{\Delta} \ge C^{-1}\,\Delta\, n^{\pi(1-\alpha)/2}$ for all integers $j \in [j_0, j_1]$. Hence $\lambda = 2\sqrt{2^{1+\gamma}\,\ln n} < (H^{1+\gamma}\widetilde{\Delta})^{1/(1+\gamma)}$ for all integers $j \in [j_0, j_1]$, for sufficiently large n.

It follows from Lemma 2.1 that

$$P\Big(|S_{n,K}| > \lambda\Big) \le \exp\Big(-\frac{\lambda^2}{4H}\Big) = n^{-1}. \tag{5.1.19}$$

Let $K \to \infty$ and use Fatou's lemma, we have

$$P\Big(\Big|\hat{\beta}_{jk} - b_{jk}\Big| > \delta_j\Big) = P\big(|S_n| > \lambda\big) \le \liminf_{K\to\infty} P\Big(|S_{n,K}| > \lambda\Big) \le C\, n^{-1}.$$

This finishes the proof of Lemma 3.7.

Remark 5.3. From the proof of Lemma 3.7, we see that by choosing λ appropriately, the tail probability estimate (3.3.13) can be significantly improved.

Lemma 5.4. *Under the assumptions of Theorem 3.1,*

$$I_1 := \sum_k E\big(\hat{\alpha}_{j_0k} - \alpha_{j_0k}\big)^2 = o\Big(n^{-2\sigma\alpha/(2\sigma+\alpha)}\, \log_2 n\Big).$$

Proof: Note that

$$I_1 \le 2\Big[\sum_k E(\hat{\alpha}_{j_0k} - a_{j_0k})^2 + \sum_k (a_{j_0k} - \alpha_{j_0k})^2\Big] =: 2(I_{11} + I_{12}).$$

As to the first term, we may apply the similar calculation as that in Lemma 3.7 to derive

$$\begin{aligned} I_{11} &= \sum_k 2^{-j_0} C_0 \big(2^{j_0} n^{-1}\big)^\alpha \Big[\int_0^1 \int_0^1 |x-y|^{-\alpha} \phi(x)\phi(y)\, dxdy + o(1)\Big] \\ &= \sum_{k=0}^{2^{j_0}-1} 2^{-j_0} C_0 \big(2^{j_0} n^{-1}\big)^\alpha \iint |x-y|^{-\alpha} \phi(x)\phi(y)\, dxdy + o\big((2^{j_0} n^{-1})^\alpha\big) \\ &\le C\big(2^{j_0} n^{-1}\big)^\alpha = o\Big(n^{-2\sigma\alpha/(2\sigma+\alpha)}\, \log_2 n\Big), \end{aligned}$$

where the last equality follows from our choice of j_0 with $2^{j_0} \simeq \log_2 n$.

As to the second term, since $\tau \le \tau_n = O(n^{\theta+0.25\alpha(2r+1)^{-1}}) = O(n^{1/2})$, from Lemma 5.1 and Remark 5.1, we have

$$I_{12} = O\Big(2^{j_0} n^{-1}\Big) = o\Big(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n\Big).$$

Together with term I_{11}, this proves Lemma 5.4.

Lemma 5.5. *Under the assumptions of Theorem 3.1,*

$$I_2 := \sum_{j=j_0}^{j_\sigma} \sum_k E\big(\hat{\theta}_{jk} - \beta_{jk}\big)^2 \le C n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n,$$

where $\hat{\theta}_{jk} = \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > \delta_j)$ *and* $j_\sigma = j_\sigma(n)$ *such that* $2^{j_\sigma} \simeq \big(n^{-1} \log_2 n\big)^{-\alpha/(2\sigma+\alpha)}$.

Proof: Notice $\hat{\theta}_{jk} = \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > \delta_j)$, we have

$$\begin{aligned} I_2 &\le 2 \sum_{j=j_0}^{j_\sigma} \sum_k E\Big[\beta_{jk}^2 I\big(|\hat{\beta}_{jk}| \le \delta_j\big)\Big] + 2 \sum_{j=j_0}^{j_\sigma} \sum_k E\Big[\big(\hat{\beta}_{jk} - \beta_{jk}\big)^2 I\big(|\hat{\beta}_{jk}| > \delta_j\big)\Big] \\ &=: 2(I_{21} + I_{22}). \end{aligned} \tag{5.1.20}$$

Also,

$$\begin{aligned} I_{21} &\le \sum_{j=j_0}^{j_\sigma} \sum_k \beta_{jk}^2 I\big(|\beta_{jk}| \le 2\delta_j\big) + \sum_{j=j_0}^{j_\sigma} \sum_k \beta_{jk}^2 P\big(|\hat{\beta}_{jk} - \beta_{jk}| > \delta_j\big) \\ &=: I_{211} + I_{212}. \end{aligned} \tag{5.1.21}$$

Since there are at most 2^j non-zero terms of β_{jk}'s and $\delta_j^2 = 2^{3+\gamma} C_2 n^{-\alpha} 2^{-j(1-\alpha)} \ln n$, we have

$$I_{211} \le \sum_{j=j_0}^{j_\sigma} \sum_k 4\delta_j^2 \ \le C \log_2 n \cdot n^{-\alpha} \sum_{j=j_0}^{j_\sigma} 2^{j\alpha} \le C n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n. \tag{5.1.22}$$

As to the term I_{212}, from (3.3.12) in Lemma 3.6 and our choice of τ, it is easy to see that $\sup_k |b_{jk} - \beta_{jk}| < \delta_j$ for all $j \in [j_0, j_\sigma]$. Thus, $I_{212} = O\big(\sum_{j=j_0}^{j_\sigma} \sum_k \beta_{jk}^2 P\big(|\hat{\beta}_{jk} - b_{jk}| >$

$\delta_j)$). Write

$$\beta_{jk} = \int g\psi_{jk} = \int g_1\psi_{jk} + \int g_2\psi_{jk} =: \beta_{jk,1} + \beta_{jk,2}$$

as in Remark 5.1. Since $g_1 \in \mathcal{G}^{\sigma}_{\infty,\infty}$, we have $\beta^2_{jk,1} = O(2^{-j(1+2\sigma)})$. As to $\beta_{jk,2}$, since $g_2 \in P_{d\tau A}$ and our wavelet ψ has $r\,(r > d)$ vanish moments, there are at most τ non-zero $\beta_{jk,2}$ terms with $\beta^2_{jk,2} = O(2^{-j})$. Thus, apply Lemma 3.7, we have

$$I_{212} \le C\sum_{j=j_0}^{j_\sigma} 2^j 2^{-j(1+2\sigma)} n^{-1} + C\sum_{j=j_0}^{j_\sigma} \tau 2^{-j} n^{-1} = o(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n). \tag{5.1.23}$$

Now let's consider the second term I_{22}. Apply Lemma 3.6 and $E(\hat{\beta}_{jk} - b_{jk})^2 \sim \sigma_j^2$ as that in Lemma 5.4, we have

$$\begin{aligned} I_{22} &\le 2\Big[\sum_{j=j_0}^{j_\sigma}\sum_k E\big(\hat{\beta}_{jk} - b_{jk}\big)^2 + \sum_{j=j_0}^{j_\sigma}\sum_k \big(\beta_{jk} - b_{jk}\big)^2\Big] \\ &\le C\sum_{j=j_0}^{j_\sigma}\sum_k n^{-\alpha} 2^{-j(1-\alpha)} + C\sum_{j=j_0}^{j_\sigma} 2^j\big(n^{-1} + \tau^2 n^{-2}\big) \\ &\le Cn^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n, \end{aligned} \tag{5.1.24}$$

where the last inequality follows from our choice $\tau \le \tau_n$, $\sigma < r$ and $1 \le r$. Combining with (5.1.20), (5.1.21), (5.1.22) and (5.1.23), this completes the proof of the lemma.

Lemma 5.6. *Under the assumptions of Theorem 3.1,*

$$I_3 := \sum_{j=j_\sigma+1}^{j_1}\sum_k E\big(\hat{\theta}_{jk} - \beta_{jk}\big)^2 \le Cn^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n,$$

where $\hat{\theta}_{jk} = \hat{\beta}_{jk} I(|\hat{\beta}_{jk}| > \delta_j)$ *and* $j_\sigma = j_\sigma(n)$, *such that* $2^{j_\sigma} \simeq \big(n^{-1}\log_2 n\big)^{-\alpha/(2\sigma+\alpha)}$.

Proof: As in Lemma 5.5, we have

$$\begin{aligned} I_3 &\le 2\sum_{j=j_\sigma+1}^{j_1}\sum_k E\Big[\beta^2_{jk} I\big(|\hat{\beta}_{jk}| \le \delta_j\big)\Big] + 2\sum_{j=j_\sigma+1}^{j_1}\sum_k E\Big[\big(\hat{\beta}_{jk} - \beta_{jk}\big)^2 I\big(|\hat{\beta}_{jk}| > \delta_j\big)\Big] \\ &=: 2(I_{31} + I_{32}). \end{aligned} \tag{5.1.25}$$

Also,

$$I_{31} \le \sum_{j=j_\sigma+1}^{j_1} \sum_k \beta_{jk}^2 I\big(|\beta_{jk}| \le 2\delta_j\big) + \sum_{j=j_\sigma+1}^{j_1} \sum_k \beta_{jk}^2 P\big(|\hat{\beta}_{jk} - \beta_{jk}| > \delta_j\big) \tag{5.1.26}$$

$$=: I_{311} + I_{312}.$$

Let's consider term I_{311} first. From Remark 5.1, we only need to prove

$$I_{311,l} = \sum_{j=j_\sigma+1}^{j_1} \sum_k \beta_{jk,l}^2 I\big(|\beta_{jk,l}| \le 2\delta_j\big) \le Cn^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n, \quad l = 1,\ 2. \tag{5.1.27}$$

Since $\beta_{jk,1}^2 = O(2^{-j(1+2\sigma)})$, we have

$$I_{311,1} \le C \sum_{j=j_\sigma+1}^{j_1} 2^j \cdot 2^{-j(1+2\sigma)} \le C2^{-2\sigma j_\sigma} = Cn^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n.$$

For the second term $I_{311,2}$, since $g_2 \in P_{d\tau A}$ and our wavelet ψ has r vanish moments with $r > d$, there are at most τ non-zero coefficients $\beta_{jk,2}$. Because $|\beta_{jk,2}| \le 2\delta_j$ for these τ terms, we have

$$I_{311,2} \le C \sum_{j=j_\sigma+1}^{j_1} \tau\delta_j^2 \le C\tau n^{-\alpha} 2^{-(1-\alpha)j_\sigma} \le Cn^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n,$$

the last inequality follows from $\tau \le \tau_n = O(n^{\theta+0.25\alpha(2r+1)^{-1}})$. Thus we prove (5.1.27).

As to the term I_{312}, we have, for any positive number α_1 and α_2 such that $\alpha_1+\alpha_2 = 1$,

$$I_{312} = \sum_{j=j_\sigma+1}^{j_1} \sum_k \beta_{jk}^2 P\big(|\hat{\beta}_{jk} - b_{jk}| > \alpha_1\delta_j\big) + \sum_{j=j_\sigma+1}^{j_1} \sum_k \beta_{jk}^2 I\big(|b_{jk} - \beta_{jk}| > \alpha_2\delta_j\big).$$

Since we can choose α_1 large enough, close to 1, from Lemma 3.7, the first term in I_{312} is bounded by $C\sum_{j=j_\sigma+1}^{j_1} 2^j 2^{-j} n^{-1} = o(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n)$.

As to the second term in I_{312}, based on Lemma 3.6, we have for all $j \in [j_0, j_1]$, $|b_{jk} - \beta_{jk}| < \alpha_2\delta_j$ for sufficient large n. Therefore this term is negligible. Together with (5.1.26) and (5.1.27), we prove the bound for term I_{31}.

As to the term I_{32}, for any $\eta_1 \in (0,\ 1)$, we have

$$\begin{aligned} I_{32} \le & \sum_{j=j_\sigma+1}^{j_1} \sum_k E\Big[\big(\hat{\beta}_{jk}-\beta_{jk}\big)^2 I\big(|\beta_{jk}| > \eta_1\delta_j\big)\Big] \\ & + \sum_{j=j_\sigma+1}^{j_1} \sum_k E\Big[\big(\hat{\beta}_{jk}-\beta_{jk}\big)^2 I\big(|\hat{\beta}_{jk}-\beta_{jk}| > (1-\eta_1)\delta_j\big)\Big] \qquad (5.1.28) \\ =: & \ I_{321} + I_{322}. \end{aligned}$$

Let's consider I_{321} first. Applying the same argument as in I_{22}, using Lemma 3.6 and noticing there are at most τ terms that $|\beta_{jk}| > \eta_1\delta_j$, we have

$$\begin{aligned} I_{321} \le & C \sum_{j=j_\sigma+1}^{j_1} \sum_k n^{-\alpha} 2^{-j(1-\alpha)} I\big(|\beta_{jk}| > \eta_1\delta_j\big) + C \sum_{j=j_\sigma+1}^{j_1} \tau\big(n^{-1}+\tau^2 n^{-2}\big) \qquad (5.1.29) \\ =: & \ I_{3211} + I_{3212}. \end{aligned}$$

For the second term I_{3212}, based on the boundness of $\tau \le \tau_n$ in Theorem 3.1, we have $I_{3212} \le Cj_1\tau n^{-1} + Cj_1\tau^3 n^{-2} = o(n^{-2\sigma\alpha/(2\sigma+\alpha)}\ \log_2 n)$.

As to the first term I_{3211}, we can consider $I_{3211,1}$ and $I_{3211,2}$, respectively. For the term $I_{3211,2}$, we have $I_{3211,2} \le C\sum_{j=j_\sigma+1}^{j_1} \tau n^{-\alpha} 2^{-j(1-\alpha)}$, which is the same as $I_{311,2}$. As to the term $I_{3211,1}$, since $\beta_{jk,1}^2 > \eta_1^2\delta_j^2$ in $I_{3211,1}$, we have, for any $t > 0$,

$$\begin{aligned} I_{3211,1} &\le Cn^{-\alpha} \sum_{j=j_\sigma+1}^{j_1} \sum_k 2^{-j(1-\alpha)} \Big(\beta_{jk,1}^2 \eta_1^{-2} \delta_j^{-2}\Big)^t \\ &= \frac{Cn^{\alpha(t-1)}}{(\log_2 n)^t} \sum_{j=j_\sigma+1}^{j_1} \sum_k \beta_{jk,1}^{2t} 2^{-j(1-\alpha)(1-t)} \\ &\le \frac{Cn^{\alpha(t-1)}}{(\log_2 n)^t} \sum_{j=j_\sigma+1}^{j_1} 2^{-j(1+2\sigma)t} 2^{-j(1-\alpha)(1-t)} \\ &= o\Big(n^{-2\sigma\alpha/(2\sigma+\alpha)}\ \log_2 n\Big). \end{aligned}$$

Together with I_{3212}, we prove the bound for I_{321}. In order to prove the Lemma, in view of (5.1.28), it remains to bound the last term I_{322}.

As before, we may write

$$I_{322} \leq 2 \sum_{j=j_\sigma+1}^{j_1} \sum_k E\Big[\big(\hat{\beta}_{jk} - b_{jk}\big)^2 I\big(|\hat{\beta}_{jk} - \beta_{jk}| > (1-\eta_1)\delta_j\big)\Big]$$
$$+ 2 \sum_{j=j_\sigma+1}^{j_1} \sum_k E\Big[\big(b_{jk} - \beta_{jk}\big)^2 I\big(|\hat{\beta}_{jk} - \beta_{jk}| > (1-\eta_1)\delta_j\big)\Big] \qquad (5.1.30)$$
$$=: 2(I_{3221} + I_{3222}).$$

Apply Lemma 3.6 to see that, when n is sufficiently large, $|b_{jk} - \beta_{jk}| < (1-\eta_1)\delta_j$ for all j and k. Thus the term I_{3221} is equivalent to

$$\sum_{j=j_\sigma+1}^{j_1} \sum_k E\Big[\big(\hat{\beta}_{jk} - b_{jk}\big)^2 I\big(|\hat{\beta}_{jk} - b_{jk}| > (1-\eta_1)\delta_j\big)\Big].$$

Now we apply Hölder's inequality, for any positive numbers a and b such that $1/a + 1/b = 1$, we have it's bound

$$\sum_{j=j_\sigma+1}^{j_1} \sum_k \Big[E\big(\hat{\beta}_{jk} - b_{jk}\big)^{2a}\Big]^{1/a} \Big[P\big(|\hat{\beta}_{jk} - b_{jk}| > (1-\eta_1)\delta_j\big)\Big]^{1/b}. \qquad (5.1.31)$$

¿From Lemma 3.7, let $\eta_1 > 0$ be small enough, we derive

$$\big[P(|\hat{\beta}_{jk} - b_{jk}| > (1-\eta_1)\delta_j)\big]^{1/b} = O(n^{-1/b}).$$

For the expectation term we write $E(\hat{\beta}_{jk} - b_{jk})^{2a} = \sigma_j^{2a} E(\sum_{s\in\mathbb{Z}} \sigma_j^{-1} d_{n,s}\zeta_s)^{2a}$ and apply Rosenthal's inequality (Härdle, *et al.*, p.244) and the calculation as in Lemma 3.7 to show that this moment exists and is bounded by a constant multiple of σ_j^{2a}. Putting this together we see that (5.1.31) is (up to a constant) at most

$$\sum_{j=j_\sigma+1}^{j_1} \sum_k \sigma_j^2\, n^{-1/b} \leq \sum_{j=j_\sigma+1}^{j_1} 2^j \sigma_j^2\, n^{-1/b} \leq C 2^{\alpha j_1}\, n^{-\alpha-\frac{1}{b}} \leq C\, n^{-\pi\alpha-1/b}. \qquad (5.1.32)$$

Now we choose $a \geq (2\sigma+\alpha)/(2\sigma(1-\alpha)+\alpha)$, so that $1/b \geq 2\sigma\alpha/(2\sigma+\alpha)$. We can show the last term in (5.1.32) is bounded by $Cn^{-\pi\alpha-2\sigma\alpha/(2\sigma+\alpha)}$. Therefore we obtain that $I_{3221} = o(n^{-2\sigma\alpha/(2\sigma+\alpha)}\, \log_2 n)$.

Similar to I_{3221}, we write

$$I_{3222} \le \sum_{j=j_\sigma+1}^{j_1} \sum_k \big(b_{jk}-\beta_{jk}\big)^2 P\big(|\hat\beta_{jk}-b_{jk}| > \alpha_1(1-\eta_1)\delta_j\big) + \sum_{j=j_\sigma+1}^{j_1} \sum_k \big(b_{jk}-\beta_{jk}\big)^2 I\big(|b_{jk}-\beta_{jk}| > \alpha_2(1-\eta_1)\delta_j\big). \tag{5.1.33}$$

The bound for the first term follows from Lemma 3.6 and Lemma 3.7, while the second term is negligible too. Combining (5.1.30), we get bound for I_{322}, which, together with (5.1.28) and (5.1.29), proves the lemma.

Lemma 5.7. *Under the assumptions of Theorem 3.1,*

$$I_4 := \sum_{j=j_1+1}^{\infty} \sum_k \beta_{jk}^2 = o\Big(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n\Big).$$

Proof: Write $\beta_{jk} = \int g\psi_{jk} = \int g_1\psi_{jk} + \int g_2\psi_{jk} =: \beta_{jk,1} + \beta_{jk,2}$ as in Remark 5.1. In order to prove the lemma, it suffices to show

$$I_{4,l} := \sum_{j=j_1+1}^{\infty} \sum_k \beta_{jk,l}^2 = o\Big(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n\Big), \quad l = 1,\ 2.$$

As to $I_{4,1}$, because of the compact support of g and ψ, we have, for any level j, there are at most 2^j non-zero coefficients $\beta_{jk,1}$'s. Also from $g_1 \in \mathcal{G}^\sigma_{\infty,\infty}$, we have $\beta^2_{jk,1} = O(2^{-j(1+2\sigma)})$. Thus

$$I_{4,1} \le C \sum_{j=j_1+1}^{\infty} 2^{-2\sigma j} = C2^{-2\sigma j_1} = o(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n), \tag{5.1.34}$$

where the last equality follows from our choice of j_1 with $2^{j_1} \simeq n^{1-\pi}$ and $\pi < 0.75(2r+1)^{-1}$.

As to the second term $I_{4,2}$, since there are at most τ discontinuities for any level j and $\beta^2_{jk,2} = O(2^{-j})$ for those at most τ coefficients, we have

$$I_{4,2} \le C \sum_{j=j_1+1}^{\infty} 2^{-2\sigma j} + C \sum_{j=j_1+1}^{\infty} \tau 2^{-j}. \tag{5.1.35}$$

From the facts that $\tau \le \tau_n = O(n^{\theta+0.25(2r+1)^{-1}})$ and $2^{j_1} \simeq n^{1-\pi}$ with $\pi < 0.75(2r+1)^{-1}$, one can verify $\sum_{j=j_1+1}^{\infty} \tau 2^{-j} = \tau 2^{-j_1} = o(n^{-2\sigma\alpha/(2\sigma+\alpha)} \log_2 n)$. Combining this with (5.1.34) and (5.1.35) completes the proof of the lemma.

5.2 Proof of Theorem 3.8

For proving Theorem 3.8, we will replace Lemma 3.7 by the following tail probability estimate. It is here that the condition on η in (3.4.1) will be used.

Lemma 5.8. *Under the assumptions of Theorem 3.8, there exists a positive constant C such that*

$$P\Big(|\hat{\beta}_{jk} - b_{jk}| > \delta_j\Big) \le C\, n^{-1}, \quad \forall j \in [j_0, j_1] \ \text{ and } \ k = 0, 1, \cdots, 2^j - 1. \tag{5.2.1}$$

Proof: As in the proof of Lemma 3.7 we write

$$\hat{\beta}_{jk} - b_{jk} = n^{-1} \sum_{i=1}^{n} \varepsilon_i \psi_{jk}(x_i) =: \sum_{s \in \mathbb{Z}} d_{n,s} \zeta_s. \tag{5.2.2}$$

For any positive integers n and K, we define $X_s = X_{n,s} := \sigma_j^{-1} d_{n,s}\, \zeta_s$ for all $|s| \le K$. Then the sequence of random variables $\{X_s, |s| \le K\}$ are independent, $E(X_s) = 0$, $E(|X_s|^{2+\eta}) < \infty$, and the partial sum $S_{n,K}$ in (5.1.14) can be written as $S_{n,K} = \sum_{|s| \le K} X_s$. In the notation of Lemma 3.5 we have

$$B_K = \sum_{|s| \le K} \frac{d_{n,s}^2}{\sigma_j^2} \to 1 \quad \text{as } K \to \infty. \tag{5.2.3}$$

It follows from (5.1.16) that the generalized Lyapunov fraction L_K satisfies

$$\begin{aligned} L_K &= \frac{1}{B_K^{1+\eta/2}} \sum_{|s|\le K} E\big(|X_s|^{2+\eta}\big) \\ &\le \frac{1}{B_K^{1+\eta/2}} \max_{|s|\le K} \left(\frac{d_{n,s}}{\sigma_j}\right)^{\eta} \sum_{|s|\le K} \frac{d_{n,s}^2}{\sigma_j^2} E\big(|\zeta_s|^{2+\eta}\big) \qquad (5.2.4) \\ &\le C\big(n^{-1}2^j\big)^{\eta(1-\alpha)/2}. \end{aligned}$$

Since $2^{j_1} \simeq n^{1-\pi}$, we have $L_K \le Cn^{-\pi\eta(1-\alpha)/2}$ for all K large and all integers $j \in [j_0, j_1]$. Hence condition (3.3.1) is satisfied provided n is large enough.

Let us take $\lambda = \sqrt{\eta\pi(1-\alpha)\ln n}$. Then $\delta_j^2 = \lambda^2\sigma_j^2$ and $\lambda \le \sqrt{2\,\ln(1/L_K)}$. Therefore, by Lemma 3.5, we derive that for n and K large enough

$$P\Big(|S_{n,K}| > \lambda\Big) \le C_5\left(1 - \Phi\big(\lambda\big)\right) \le Cn^{-1}, \qquad (5.2.5)$$

thanks to the assumption that $\eta\pi(1-\alpha) \ge 2$. Let $K \to \infty$ and use Fatou's lemma, we have

$$P\Big(\Big|\hat{\beta}_{jk} - b_{jk}\Big| > \delta_j\Big) = P\big(|S_n| > \lambda\big) \le \liminf_{K\to\infty} P\Big(|S_{n,K}| > \lambda\Big) \le C\,n^{-1}.$$

This finishes the proof of Lemma 5.8.

Proof of Theorem 3.8: The proof of Theorem 3.8 is almost the same as that of Theorem 3.1, replacing Lemma 3.7 by Lemma 5.8 everywhere, where the condition on η in (3.4.1) will be used. The only other place we need to modify is in proving an upper bound for (5.1.31). We will take $2a = 2 + \eta$. During the proof, we need that $\eta \ge \frac{4\sigma\,\alpha}{2\sigma(1-\alpha)+\alpha}$, which is satisfied thanks to (3.4.1). The rest of the proof remains valid. For the sake of simplicity, we omit the details.

Bibliography

[1] Amosova, N. N. (2002). Necessity of the Cramér, Linnik and Statulevičius conditions for the probabilities of large deviations. *J. Math. Sci.* (New York) **109**, 2031–2036.

[2] Amosova, N. N. (1972). Limit theorems for the probabilities of moderate deviations. (Russian) *Vestink Leningrad Univ.* **13**, 5–14.

[3] Amosova, N. N. (1982). Probabilities of moderate deviations. *J. Math. Sci.* **20**, 2123–2130.

[4] Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimator in nonparametric regression: a comparative simulation study. *J. Statist. Software.* **6**, 1C86.

[5] Baillie, R. T. (1996). Long memory processes and fractional integration in econometrics. *J. Econometrics* **73**, 5–59.

[6] Bentkus, R. and Rudzkis, R. (1980). Exponential estimates for the distribution of random variables. (Russian) *Litovsk. Mat. Sb.* **20**, 15–30.

[7] Beran, J. (1994). *Statistics for Long Memory Processes.* Chapman and Hall, New York.

[8] Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

[9] Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika.* **76**, 503-514

[10] Cai, T. T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898C924

[11] Cai, T. T. (2002). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sinica.* **12**, 1241C1273

[12] Cai, T. T. and Silverman, B. W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya B,* **63**, 127C148

[13] Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Assoc.* **92**, 1413C1421

[14] Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc. B,* **62**, 681C698

[15] Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadephia.

[16] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet-shrinkage. *Biometrika.* **81**, 425–455.

[17] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinking. *J. Amer. Statist. Assoc.* **90**, 1200–1224.

[18] Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

[19] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B.* **57**, 301–369.

[20] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508–539.

[21] Doukhan, P., Oppenheim, G. and Taqqu, M. S. (Eds) (2003). *Theory and Applications of Long-range Dependence.* Birkhäuser, Basel.

[22] Frolov, A. N. (1998). On one-sided strong laws of large increments of sums. *Statist. Probab. Lett.* **37**, 155–165.

[23] Frolov, A. N. (2005). On probabilities of moderate deviations of sums of independent random variables. *J. Math. Sci.* **127**, 1787–1796.

[24] Ghosh, S. and Samorodnitsky, G. (2008). The effect of memory on functional large deviations of infinite moving average processes. *Stoch. Process. Appl.*, to appear.

[25] Giraitis, L., Koul, H. L. and Surgailis, D. (1996). Asymptotic normality of regression estimators with long memory errors. *Statist. Probab. Lett.* **29**, 317–335.

[26] Giraitis, L. and Surgailis, D. (1999). Central limit theorem for the empirical process of a linear sequence with long memory. *J. Statist. Plann. Inference* **80**, 81–93.

[27] Green, P. J. and Silverman, B. W. (1994). it Nonparametric Regression and Generalized Linear Models, A roughness penelty Approach. Chapman and Hall, London.

[28] Grossmann, A. and Morlet, J. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math.* **15**, 723-736

[29] Hall, P. and Hart, J. D. (1990). Nonparametric regression with long-range dependence. *Stoch. Process. Appl.* **36**, 339–351.

[30] Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet method. *Ann. Statist.* **26**, 922–942.

[31] Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9**, 33–50.

[32] Hall, P., and Penev, S. (2001). Cross-validation for choosing resolution level for nonlinear wavelet curve estimators. *Bernoulli* **7**, 317–341.

[33] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). Wavelets, Approximation and Statistical Applications. *Lecture Notes in Statistics* **129**, Springer, New York.

[34] Ho, H. C. and Hsing, T. (1996). On the asymptotic expansion of the empirical process of long memory moving averages. *Ann. Statist.* **24**, 992–1024.

[35] Ho, H. C. and Hsing, T. (1997). Limit theorems for functionals of moving averages. *Ann. Probab.* **25**, 1636–1669.

[36] Johnstone, I. M. (1999). Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statist. Sinica* **9**, 51–83.

[37] Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B.* **59**, 319–351.

[38] Koul, H. L. and Surgailis, D. (1997). Asymptotic expansion of M-estimators with long memory errors. *Ann. Statist.* **25**, 818–850.

[39] Koul, H. L. and Surgailis, D. (2001). Asymptotics of the empirical process of long memory moving averages with infinite variance. *Stoch. Process. Appl.* **91**, 309–336.

[40] Kovac, A. and Silverman, B. W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *J. Amer. Statist. Assoc.* **95**, 172–183.

[41] Li, L. and Xiao, Y. (2007). On the minimax optimality of block thresholded wavelet estimators with long memory data. *J. Statist. Plann. Inference* **137**, 2850-2869

[42] Mallat, S. G. (1989). *A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattn. Anal. Mach. Intell.* **11**, 674-693.

[43] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

[44] Morlet, J., Arens, G., Fourgeau, E. and Giard, D. (1982). Wave propagation and sampling theory. *Geophysics* **47**, 203C236

[45] Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *J.R. Statist. Soc. B* **58**, 463-479.

[46] Petrov, V. V. (1975). *Sums of Independent Random Variables.* Springer-Verlag, New York.

[47] Petrov, V. V. (2002). On probabilities of moderate deviations. *J. Math. Sci.* **109**, 2189–2191.

[48] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28**, 298–335.

[49] Rubin, H. and Sethuraman, J. (1965). Probabilities of moderate deviations. *Sankhya Ser. A* **27**, 325–346.

[50] Rudzkis, R., Saulis, L. and Statulevičius, V. (1978). A general lemma on large deviation probabilities. *Lith. Math. J.* **18**, 226–238.

[51] Saulis, L. and Statulevičius, V. (2000). Limit theorems on large deviations. In: *Limit Theorems of Probability Theory.* (Prokhorov, Yu. V. and Statulevičius, V., editors), pp. 185–266, Springer, New York. 203–236

[52] Shapior, J. M. (1993). Embedded image coding using zerotree of wavelet coefficients. *IEEE trans. on Signal Processing*, **441** 3445–3462.

[53] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman&Hall, London.

[54] T) Triebel, H. (1992). *Theory of Function Spaces II.* Birkhäuser, Basel.

[55] von Sachs, R. and Macgibbon, B. (2000). Non-parametric curve estimation by wavelet thresholding with locally stationary errors. *Scandinavian J. Statist.* **27**, 475–499.

[56] Vidakovic, B. (1999) *Statistical Modeling by Wavelets* A Wiley-Interscience Publication, New York.

[57] Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.* **24**, 466–484.

[58] Wu, Wei Biao and Zhao, Zhibiao (2008). Moderate deviations for stationary processes. *Statist. Sinica* **18**, 769–782.

[59] Zhang, S. and Wong, M. (2003). Wavelet threshold estimation for additive regression models. *Ann. Statist.* **31**, 152–173.