ESTIMATING A HÖLDER CONTINUOUS FUNCTION FROM A NOISY SAMPLE VIA SHRINKAGE AND TRUNCATION OF WAVELET COEFFICIENTS

by

Zhongcheng Wang
Purdue University

Technical Report #93-9

Department of Statistics
Purdue University

January 1993

Estimating a Hölder Continuous Function from a Noisy Sample via Shrinkage and Truncation of Wavelet Coefficients

Zhongcheng Wang

Department of Statistics
Purdue University
West Lafayette, IN47907

December 12, 1992

Abstract

We use the wavelet decomposition and reconstruction methods of multiresolution analysis to estimate a Hölder continuous function f from noisy, sampled data in the white noise model $y_i = f(x_i) + \varepsilon_i$ where the random variables ε_i have mean zero and are uncorrelated. The result will be a class of consistent estimators of the regression function f. In wavelet multiresolution analysis, any L_2 function is completely described by its wavelet coefficients. Many important properties of the function can be determined from this sequence of coefficients. We begin with a local optimal-order interpolatory scheme to get the empirical scaling function coefficients at the highest resolution. We then shrink and truncate the wavelet coefficients produced by the multiresolution decomposition so that the noise is reduced. The estimator is the function derived from the multiresolution reconstruction process based on these modified wavelet coefficients. It is local adaptive to respond spetial differences of data at different lacation. The wavelets used in this study are B-spline wavelets which are bi-orthogonal wavelets. The results can also be applied to orthogonal wavelets with essentially no modification.

Keywords: wavelet shrinkage, wavelet threshold, nonparametric regression, B-spline, multiresolution analysis.

1 Introduction

The objective of this article is to use B-spline wavelets (wavelets based on B-splines) to estimate the function f from a noisy sample $y_i = f(x_i) + \varepsilon_i$, $i = 0, \cdots, n$. We assume that the unknown function f is Hõlder continuous and that the design points $\{x_i\}_{i=0,\cdots,n}$ are equally spaced. The random variables (noise) $\{\varepsilon_i\}_{i=0,\cdots,n}$ are uncorrelated with mean zero and variance σ^2. This estimation problem has been studied using many other methods such as kernel estimators, moving linear smoother, penalized smoothing spline etc. Many of these are, in one way or another, linear estimators. The application of wavelet theory in the above problem is an interesting new nonlinear approach.

Since the wavelet transformation of $\{y_i\}$ is the sum of the transformation of $\{f(x_i)\}$ and the transformation of $\{\varepsilon_i\}$, it is important to understand the behavior of the wavelet transformation of white noise. We study the behavior and use our understanding of it to form a class of consistent estimators of the regression function f.

2 B-Wavelet and Interpolation Scheme

Let φ and ψ be a pair of functions (scaling function and wavelet function) which provides a wavelet analysis of $L^2(R)$. Define $\varphi_{j,k}(\cdot) = 2^{j/2}\varphi(2^j \cdot -k)$ and $\psi_{j,k}(\cdot) = 2^{j/2}\psi(2^j \cdot -k)$, and let $V_j = \text{clos}_{L^2(R)}\{\varphi_{j,k} : k \in Z\}$, and $W_j = \text{clos}_{L^2(R)}\{\psi_{j,k} : k \in Z\}$. Then these subspaces of $L^2(R)$ have the following properties:

(1⁰) $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots;$

(2⁰) $\text{clos}_{L^2}\{\bigcup_{j \in Z} V_j\} = L^2(R);$

(3⁰) $\bigcap_{j \in Z} V_j = \{0\};$

(4⁰) $V_{j+1} = V_j \bigoplus W_j, \quad j \in Z;$

(5⁰) $f(x) \in V_j \Longleftrightarrow f(2x) \in V_{j+1}, \quad j \in Z.$

In other words, any function f in $L^2(R)$ has a wavelet series representation:

$$\begin{aligned} f(x) &= \sum_{j,k} d_{j,k}\psi_{j,k}(x) \\ &= \cdots + \mathcal{W}_{-1}f + \mathcal{W}_0 f + \mathcal{W}_1 f + \cdots = \mathcal{P}_M f + \mathcal{W}_M f + \mathcal{W}_{M+1} f + \cdots \qquad (2.\ 1) \end{aligned}$$

where $(\mathcal{W}_j f)(x) = \sum_k d_{j,k}\psi_{j,k}(x)$ is the projection of f onto the space W_j, $(\mathcal{P}_M f)(x) = \sum_k c_{M,k}\varphi_{M,k}(x)$ is the projection of f onto the space V_M, and $\mathcal{P}_{j+1}f(x) = \mathcal{P}_j f(x) + \mathcal{W}_j f(x)$ for all integer j and M.

Wavelets also provide unconditional bases and characterizations for many functional spaces other than $L^2(R)$. For instance, the Hölder spaces $\mathcal{C}^\alpha(R)$ which are defined as:

$$\mathcal{C}^\alpha(R) = \{f \in L^\infty(R);\ \sup_{x,h} \frac{|f(x+h)-f(x)|}{|h|^\alpha} < \infty\} \quad 0 < \alpha < 1$$

$$\mathcal{C}^\alpha(R) = \{f \in C^n(R);\ f^{(n)} \in \mathcal{C}^{\alpha'}\} \qquad \alpha = n + \alpha' \quad 0 < \alpha' < 1$$

A function f is in $\mathcal{C}^\alpha(R)$ if and only if there exists a constant $C < \infty$ such that the coefficients $d_{j,k}$ in the wavelet representation (2.1) of f satisfy

$$(2.\ 2) \qquad |d_{j,k}| \le C2^{-(\frac{1}{2}+\alpha)j} \qquad \forall j,k \in Z$$

Therefore a Hölder continuous function's wavelet coefficients have exponential decay rate. On the other hand, if we use a set of wavelet coefficients which have some exponential decay rate to construct a function, then this function would have a certain degree of continuity, or smoothness. For the wavelet characterization of functional spaces and Hölder continuity, see [1, 2, 3, 6].

A function f in $L^2(R)$ can be approximated as closely as desired by its projections on V_j as seen by (2^0). The most important intrinsic property of these spaces is that more and more "variations" of the projections are removed as $j \to -\infty$. In fact, these variations are peeled off, level by level in decreasing order of the "rate of variations" and stored in the complementary subspaces W_j as in (4^0). One of the features of the wavelet theory is that this "peeling off" process can be made very efficient by an application of the property (5^0). That is, once we know the coefficients $\{c_{M,k}\}_{k\in Z}$ of a function $\mathcal{P}_M f(x) = \sum_k c_{M,k}\varphi_{M,k}(x)$ in V_M, then a recursive decomposition algorithm will quickly give us the decomposition

$$(2.\ 3) \qquad \mathcal{P}_M f = \mathcal{W}_{M-1}f + \mathcal{W}_{M-2}f + \cdots + \mathcal{W}_{M-M_0}f + \mathcal{P}_{M-M_0}f$$

where M_0 is an arbitrary integer depending on the application problem. On the other hand, if we know the coefficients $\{d_{M-\ell,k}\}_{k\in Z}$ of $\mathcal{W}_{M-\ell}f$ for $\ell = 1, 2, \cdots, M_0$, and the coefficients $\{c_{M-M_0,k}\}_{k\in Z}$ of $\mathcal{P}_{M-M_0}f$, then the reconstruction algorithm will quickly give us $\mathcal{P}_M f$.

There are many choices of the functions φ and ψ. In [1, 4, 5], it is proved that $\varphi(x) = N_m(x)$ and $\psi(x) = L_{2m}^{(m)}(2x-1)$ provide a wavelet analysis of $L^2(R)$, where N_m denotes the mth order B-spline

$$N_m(x) = (N_{m-1} * N_1)(x) = \int N_{m-1}(x-t)dt$$

with $N_1(x) = I_{[0,1)}(x)$, $L_{2m}(x)$ denotes the $(2m)$th order fundamental cardinal interpolatory spline and its mth derivative is

$$L_{2m}^{(m)}(2x-1) = \sum_{j=0}^{2m-2} \frac{(-1)^j}{2^{m-1}} N_{2m}(j+1) N_{2m}^{(m)}(2x-j)$$

where $N_{2m}^{(m)}$ is the mth derivative of N_{2m}. The wavelets introduced here are called B-wavelets because they are constructed from B-splines. We need to point out that B-wavelets are semi-orthogonal wavelets (see [1,4,5]). In most of our analysis we use cubic ($m = 4$) B-spline wavelets, $\varphi(x) = N_4(x)$ and $\psi(x) = L_8^{(4)}(2x-1)$.

In applications, one often truncates the wavelet series representation of f in (2.1) to $\mathcal{P}_M f$ whose coefficients might be approximated by the observed data. Then one works on the decomposition (2.3) of $\mathcal{P}_M f$.

In the problem of recovering a function f from a noisy sample, we can not get $\mathcal{P}_M f$ because we do not know f, we only have finite number of f observations which are confounded with noise. But there is a candidate to replace $\mathcal{P}_M f$ to give us a starting point.

Applied mathematicians developed a completely local interpolation scheme [1, 4, 5] using mth order B-spline. It utilizes only finite blocks of data information but gives the optimal order of approximation. That is, with a finite sequence $\{w_{-k_m}^{(m)}, \cdots, w_{k_m}^{(m)}\}$

$$(2.\ 4) \qquad (P_M f)(x) = \sum_j \{\sum_i w_{j+m/2-il_m}^{(m)} f(il_m/2^M)\} N_m(2^M x - j)$$

satisfies:

$$(2.\ 5) \qquad \begin{cases} (P_M f - f)(jl_m/2^M) = 0, & j \in Z \\ \|P_M f - f\| = O(1/2^{mM}), & M \to \infty \quad f \in C^m \end{cases}$$

where l_m is a positive integer depending on the order m of the spline. M is an integer and $l_m/2^M$ denotes the sampling rate. Note that $P_M f$ is an mth order spline with knots at $2^{-M}Z$. For $m = 4$, $\{w_{\pm 4}, w_{\pm 3}, w_{\pm 2}, w_{\pm 1}, w_{\pm 0}\} = \{1/48, -1/12, -1/8, 7/12, 29/24\}$ and $l_4 = 2$.

We will use $P_M f$ to approximate $\mathcal{P}_M f$. $P_M f$ depends only on f's values on integers or dyadic numbers, thus it is determined completely by the observations of f, whereas the determination of $\mathcal{P}_M f$ involves convolutions of f with other functions. Both $P_M f$ and $\mathcal{P}_M f$ live in the same subspace V_M. While $\mathcal{P}_M f$ is the L^2 projection of f onto V_M, $P_M f$ is the optimal approximation to f in V_M in the sense of (2.5). Therefore for smooth functions such as Hölder continuous functions, we expect that (2.4) gives a good approximation to $\mathcal{P}_M f$.

3 Decomposition and Reconstruction Algorithm, Representation Function of Sample

The decomposition (2.3) of $\mathcal{P}_M f$ is defined as follows:

$$(3.\ 1) \qquad \begin{cases} (\mathcal{W}_{M-\ell}f)(x) = \sum_n d_n^\ell \psi(2^{M-\ell}x - n) \\ \\ (\mathcal{P}_{M-\ell}f)(x) = \sum_n c_n^\ell \varphi(2^{M-\ell}x - n) \qquad \ell = 1, 2, \cdots, M_0 \end{cases}$$

Notice that we write $c_n^\ell = c_{M-\ell,n}2^{(M-\ell)/2}$, and $d_n^\ell = d_{M-\ell,n}2^{(M-\ell)/2}$. Thus (2.2) becomes

$$(3.\ 2) \qquad |d_n^\ell(f)| \le C2^{-\alpha(M-\ell)} \qquad f \in \mathcal{C}^\alpha$$

The wavelet coefficients at different levels satisfy the relationship

$$(3.\ 3) \qquad \begin{cases} c_k^\ell = \sum_n a_{n-2k} c_n^{\ell-1} \\ \\ d_k^\ell = \sum_n b_{n-2k} c_n^{\ell-1} \qquad \ell = 1, 2, \cdots, M_0 \end{cases}$$

Two constant sequences $\{a_n\}$ and $\{b_n\}$ have at least exponential decay rate and satisfy $\sum a_n = 2$, $\sum b_n = 0$. The decomposed function components $\mathcal{W}_{M-1}f, \cdots, \mathcal{W}_{M-M_0}f$ and $\mathcal{P}_{M-M_0}f$ can be processed by modifying the wavelet coefficients $\{c_n^{M_0}\}$ and $\{d_n^\ell\}$ into $\{\tilde{c}_n^{M_0}\}$ and $\{\tilde{d}_n^\ell\}$. To reconstruct the function $\tilde{f}_M(x) = \widetilde{\mathcal{P}_M f}(x)$, use the reconstruction algorithm:

$$(3.\ 4) \qquad \begin{cases} \tilde{f}_M(x) = \sum_n \tilde{c}_n^0 \varphi(2^M x - n) \\ \\ \tilde{c}_k^{\ell-1} = \sum_n (p_{k-2n}\tilde{c}_n^\ell + q_{k-2n}\tilde{d}_n^\ell) \qquad \ell = 1, 2, \cdots, M_0 \end{cases}$$

for sequences $\{a_n\}$, $\{b_n\}$, $\{p_n\}$, and $\{q_n\}$ as well as their properties, see [1, 4, 5]. The simple recursive form of this wavelet transformation is the property for the implementation of the algorithm.

The cubic B-spline interpolation scheme represents the sample

$$(3.\ 5) \qquad \{y_i\}_{i=0,\cdots,n} = \{f(x_i) + \varepsilon_i\}_{i=0,\cdots,n} \qquad x_i = i2^{1-M}$$

as a spline series

$$\begin{aligned} P_M y(x) &= \sum_n c_n^0 \varphi(2^M x - n) \\ &= \sum_n c_n^0(f)\varphi(2^M x - n) + \sum_n c_n^0(\varepsilon)\varphi(2^M x - n) \\ (3.\ 6) \qquad &= P_M f(x) + P_M \varepsilon(x) \end{aligned}$$

where $\{c_n^0\} = \{c_n^0(f) + c_n^0(\varepsilon)\}$ is the approximation to $\{c_{M,n}\}$ which are the coefficients in $\mathcal{P}_M f + \mathcal{P}_M \varepsilon$, and

$$\left\{ \begin{array}{l} c_n^0(f) = \sum_i w_{n+2-2i} f(x_i) \\ \\ c_n^0(\varepsilon) = \sum_i w_{n+2-2i} \varepsilon_i \end{array} \right. \tag{3. 7}$$

We will use (3.6) to approximate

$$\mathcal{P}_M y = \mathcal{P}_M f + \mathcal{P}_M \varepsilon \tag{3. 8}$$

where $\mathcal{P}_M \varepsilon := \mathcal{P}_M [P_M \varepsilon] = P_M \varepsilon$. And (3.8) will serve as our theoretical model.

The decomposition of $\mathcal{P}_M y$ is

$$\left\{ \begin{array}{l} \mathcal{P}_M y(x) = \mathcal{W}_{M-1} y(x) + \cdots + \mathcal{W}_{M-M_0} y(x) + \mathcal{P}_{M-M_0} y(x) \\ \\ \mathcal{W}_{M-\ell} y(x) = \mathcal{W}_{M-\ell} f(x) + \mathcal{W}_{M-\ell} \varepsilon(x) = \sum_n (d_n^\ell(f) + d_n^\ell(\varepsilon)) \psi(2^{M-\ell} x - n) \\ \\ \mathcal{P}_{M-M_0} y(x) = \mathcal{P}_{M-M_0} f(x) + \mathcal{P}_{M-M_0} \varepsilon(x) = \sum_n (c_n^{M_0}(f) + c_n^{M_0}(\varepsilon)) \varphi(2^{M-M_0} x - n) \end{array} \right. \tag{3. 9}$$

We call function $P_M y$ the representation function of the sample (3.5) because the noisy sample of the underlying function f is just the exact sample of function $P_M y$ at some dyadic numbers. Since the sample is the composition of the underlying function and noise, we do not simply do reconstruction to get the original noisy sample back, we modify the wavelet coefficients to get a smoothed estimation $\tilde{f}^M = \widetilde{\mathcal{P}_M f}$ of true underlying function f. In order to do so, we have to be able to see how the random noise affects the wavelet coefficients.

4 Analysis of Variance of Wavelet Decomposition

We use $P_M y$ defined in (3.6) to represent the sample (3.5). For a noise free sample from a cubic polynomial segment, $P_M y$ reproduces the polynomial segment. In other words, $P_M f = f$ if f is a cubic polynomial. This can be shown directly from (3.6). Such low degree polynomial reproducing ability is especially important for smoothing.

For noisy data, $P_M y(x) = P_M f(x) + P_M \varepsilon(x)$ has expectation $P_M f(x)$. By (2.5) we know that $P_M f(x)$ approaches $f(x)$ as $M \to \infty$ for f having a certain degree of continuity, therefore the expectation of the function $P_M y$ would approach the truth f. However we can not use $P_M y$ to estimate f simply because it is too noisy.

Assuming the variance of $\{\varepsilon_i\}_{i=0,\cdots,n}$ is σ^2, the variance of $\mathcal{P}_M\varepsilon(x)$ is between $0.6592\sigma^2$ and σ^2. In fact, the variance of $\mathcal{P}_M\varepsilon(x)$ is a periodical function with period 2^{1-M}.

$$\sigma^{-2}Var(\mathcal{P}_M\varepsilon(2^{1-M}x)) = \sigma^{-2}Var(\sum_n\sum_i w_{n+2-2i}\varepsilon_i\varphi(2x-n)) = \sum_i(\sum_n w_{n-2i}\varphi(2x+2-n))^2$$

This is also the variance of $P_M\varepsilon(x)$ or $\mathcal{P}_My(x)$ or $P_My(x)$. Figure 1 is the plot of one period of the variance of $\mathcal{P}_M\varepsilon(2^{-M}x)/\sigma$ as a function of x.

For analyzing the variance of $\mathcal{W}_{M-1}y, \cdots, \mathcal{W}_{M-M_0}y$, and $\mathcal{P}_{M-M_0}y$, it is enough only to trace the stochastic part in the decomposition, i.e. let $f(x) = 0$. We think of $\{\varepsilon_i\}_{i\in Z}$ as a stationary process so that we do not have to worry about the boundary problem. The idea of viewing this sequence as a stationary process will help us to see how the variation of $\{\varepsilon_i\}$ carry on from one layer to another by means of the spectral density.

Let $\{c_n^0\}_{n\in Z} = \{c_n^0(\varepsilon)\}_{n\in Z}$ be defined as in (3.7), and $\{\varepsilon_i\}_{i\in Z}$ be a white noise $WN(0,\sigma^2)$. A simple calculation shows that $Var(c_{2k}^0) = \frac{1699}{1172}\sigma^2$, $Var(c_{2k+1}^0) = \frac{25}{36}\sigma^2$, and $Cov[c_{2k}^0, c_{2k+\ell}^0] = 0$ if $\ell \geq 9$. Although this calculation indicates that $\{c_n^0\}$ is not a stationary process, the sequence $\{c_n^1\}$ is a stationary process with autocovariance function

$$\gamma_1(h) = \sigma^2\sum_i(\sum_j a_jw_{j+2+2i})(\sum_j a_jw_{j+2+2i-2h}) \tag{4. 1}$$

This is because $\{c_n^1\}$ can be written as a filtered process from $\{\varepsilon_i\}$.

$$\begin{aligned} c_n^1 &= \sum_j a_{j-2n}c_j^0 = \sum_j a_{j-2n}\sum_i w_{j+2-2i}\varepsilon(i) \\ &= \sum_i\sum_j a_jw_{j+2n+2-2i}\varepsilon(i) = \sum_i\sum_j a_jw_{j+2+2i}\varepsilon(n-i) \end{aligned}$$

where $\sum_i|\sum_j a_jw_{j+2+2i}| < \infty$.

The sequences $\{c_n^\ell\}$ defined by (3.3) recursively are also stationary processes for all $\ell \geq 2$ as long as $\{c_n^{\ell-1}\}$ is a stationary process and $\{a_j\}$ is in ℓ^1 space. We state our observations as the following propositions.

Let us consider the operator

$$\Theta(\mathbf{a},t) = \sum_{j=-\infty}^{\infty} a_jF^{t+j} \tag{4. 2}$$

the downsampling linear filter with weights $\{a_j\}$. where F is the forward shift operator

$$F^jX_t = X_{t+j}$$

We can rewrite (3.3) as

$$(4.\ 3) \qquad \begin{cases} c_k^\ell = \Theta(\mathbf{a}, k)c_k^{\ell-1} \\ \\ d_k^\ell = \Theta(\mathbf{b}, k)c_k^{\ell-1} \qquad \ell = 1, 2, \cdots, M_0 \end{cases}$$

Notice that this linear filter is not a time-invariant linear filter, but it will take any stationary process to another stationary process.

Proposition 4.1 *If $\{c_k^{\ell-1}\}$ is a stationary process with autocovariance function $\gamma_{\ell-1}(\cdot)$, then $\{c_k^\ell$ as defined in (4.3) is a stationary process with autocovariance function*

$$(4.\ 4) \qquad \gamma_\ell(h) = \sum_{j,k=-\infty}^{\infty} a_j a_k \gamma_{\ell-1}(2h + j - k)$$

The proofs of this proposition and the next two are given in a seperate study in order to focus on discussing the wavelet function estimator.

This proposition says that $\{c_k^\ell\}$ is well defined from $\{c_k^{\ell-1}\}$ via (4.3) and every sequence $\{c_k^\ell\}$, $\ell \geq 1$, is a stationary process. The autocovariance functions satisfy the relationship (4.4). In fact, two autocovariance functions $\gamma_\ell(\cdot)$ and $\gamma_{\ell-1}(\cdot)$ of the stationary processes $\{c_k^{\ell-1}\}$ and $\{c_k^\ell\}$ are not only connected by (4.4), but also by the corresponding spectral distribution functions or spectral density functions.

Proposition 4.2 *If $\{c_k^{\ell-1}\}$ is a stationary process with spectral density function $f_{\ell-1}(\cdot)$, then the spectral density function $f_\ell(\cdot)$ of the stationary process $\{c_k^\ell\}$ as defined in (4.3) is*

$$(4.\ 5) \qquad f_\ell(\lambda) = \frac{1}{2}g(\frac{\lambda}{2})f_{\ell-1}(\frac{\lambda}{2}) + \frac{1}{2}g(\pi - \frac{\lambda}{2})f_{\ell-1}(\pi - \frac{\lambda}{2}) \qquad 0 < \lambda \leq \pi$$

where $g(\lambda) = |\sum_{j=-\infty}^{\infty} a_j e^{ij\lambda}|^2$.

Proposition 4.3 *There is a θ such that $f_\ell = O(\theta^\ell)$.*

A function sequence like the one defined in (4.5) has some interesting properties. Similar function sequences have been studied in dynamic system and ergodic theory. A Ruelle's Perron-Froebenius theorem type result can be proved.

Proposition 4.3 tell us that the variances of $\{c_n^\ell\}$ and $\{d_n^\ell\}$ decrease exponentially as ℓ increases.

$$E[d_n^\ell(\varepsilon)]^2 \leq C\theta^\ell\sigma^2 \tag{4. 6}$$

for some θ which depends on the sequence $\{a_j\}$.

The Constant C and θ can be estimated numerically. In the case of cubic B-wavelets, θ is close to one half.

5 A Class of Consistent Estimators

For a given sampling rate 2^{1-M}, one can only do a limited number of decomposition before boundary effects seriously show up. That is to say that M_0, the number of decomposition steps, is an integer which can not be too far from M. When ℓ is small, $\mathcal{W}_{M-\ell}y$ in (3.9) contains the high frequency portion of $\mathcal{P}_{M-\ell}y$, and this portion is mainly the information about the random noise. Therefore we need to get rid of it. As ℓ increases, $\mathcal{W}_{M-\ell}y$ carries less and less information from the stochastic part, and the fluctuation of $\mathcal{W}_{M-\ell}y$ comes mainly from the true signal f. The modification of $\mathcal{W}_{M-\ell}y$ for large ℓ contributes little to clean the noise, but tortures the meaningful information of the true signal contained in $\mathcal{W}_{M-\ell}y$ and in turn affects the quality of the reconstruction of f. To balance between cleaning the noise and conserving the true information, we should focus on modifying $\mathcal{W}_{M-\ell}y$ for small ℓ.

Let $\{m_\ell\}_{\ell=1}^{M_0} = \{m_\ell(M_0)\}_{\ell=1}^{M_0}$ be a sequence of M_0 numbers which may depend on M_0 and are between zero and one. Let $\tilde{f}(x) = \widetilde{\mathcal{P}_M y}(x)$ be the function derived from the multiresolution reconstruction process (3.4) based on the modified wavelet coefficients $\{\tilde{d}_n^\ell\} = \{m_\ell d_n^\ell\}$.

$$\begin{aligned}
\tilde{f}(x) &= \sum_{\ell=1}^{M_0} m_\ell \mathcal{W}_{M-\ell}y(x) + \mathcal{P}_{M-M_0}y(x) \\
&= \sum_{\ell=1}^{M_0} m_\ell[\mathcal{W}_{M-\ell}f(x) + \mathcal{W}_{M-\ell}\varepsilon(x)] + \mathcal{P}_{M-M_0}f(x) + \mathcal{P}_{M-M_0}\varepsilon(x)
\end{aligned} \tag{5. 1}$$

If we use $\tilde{f}(x) = \widetilde{\mathcal{P}_M y}(x)$ as an estimator of $f(x)$, then we have the following results.

Theorem 5.1 *If the sequence* $\{m_\ell\}_{\ell=1}^{M_0}$ *is such that*

$$\sum_{\ell=1}^{M_0}[|1-m_\ell|2^{-\alpha(M-\ell)} + M_0 m_\ell^2\theta^\ell] \to 0 \qquad as \quad M_0 \to \infty \tag{5. 2}$$

then $\tilde{f}$ is a consistent estimator of $f \in C^\alpha$ in the sense that

$$MSE[\tilde{f}(x)] \to 0 \quad \textit{as} \quad M_0 \to \infty$$

PROOF: The absolute value of the bias of $\tilde{f}(x)$ is

$$\begin{aligned} |bias(\tilde{f}(x))| &= |E(\tilde{f}(x) - f(x))| \\ &= |\sum_{\ell=1}^{M_0} m_\ell \mathcal{W}_{M-\ell} f(x) + \mathcal{P}_{M-M_0} f(x) - f(x)| \\ &= |\sum_{\ell=1}^{M_0} (m_\ell - 1)\mathcal{W}_{M-\ell} f(x) + \mathcal{P}_M f(x) - f(x)| \\ &\leq \sum_{\ell=1}^{M_0} |1 - m_\ell||\mathcal{W}_{M-\ell} f(x)| + |\mathcal{P}_M f(x) - f(x)| \\ &= A_1 + A_2 \qquad (5.\ 3) \end{aligned}$$

If $f \in C^\alpha$, then we have (3.2) $|d_n^\ell(f)| \leq C2^{-\alpha(M-\ell)}$. Notice that the support of ψ is from 0 to 7, we have

$$|\mathcal{W}_{M-\ell} f(x)| = |\sum_n d_n^\ell(f)\psi(2^{M-\ell}x - n)| \leq 7|\psi|C2^{-\alpha(M-\ell)}$$

Thus $A_1 \leq C\sum_{\ell=1}^{M_0} |1 - m_\ell| 2^{-\alpha(M-\ell)} \to 0$ as $M_0 \to \infty$ under the condition (5.2).

The second term A_2 goes to zero as $M \to \infty$ because of the properties (1^o) and (2^o) of the multiresolution analysis of L^2 and the continuity of $f(x)$ and $\mathcal{P}_M f(x)$.

In words, if the sampling rate is high, then the initial resolution level M is high, and the bias square of $\tilde{f}(x)$ is small.

Now the variance of $\tilde{f}(x)$ is

$$\begin{aligned} Var(\tilde{f}(x)) &= E[\sum_{\ell=1}^{M_0} m_\ell \mathcal{W}_{M-\ell}\varepsilon(x) + \mathcal{P}_{M-M_0}\varepsilon(x)]^2 \\ &\leq (M_0 + 1)\sum_{\ell=1}^{M_0} [m_\ell^2 E(\mathcal{W}_{M-\ell}\varepsilon(x))^2 + E(\mathcal{P}_{M-M_0}\varepsilon(x))^2] \end{aligned}$$

Since

$$\begin{aligned} E(\mathcal{W}_{M-\ell}\varepsilon(x))^2 &= E[\sum_n d_n^\ell(\varepsilon)\psi(2^{M-\ell}x - n)]^2 \\ &\leq 7|\psi|^2 \sum_{0<2^{M-\ell}x-n<7} E(d_n^\ell(\varepsilon))^2 \\ &\leq 7^2|\psi|^2 E(d_{\cdot}^\ell(\varepsilon))^2 \\ &\leq C\theta^\ell \end{aligned}$$

and similarly

$$E[\mathcal{P}_{M-M_0}\varepsilon(x)]^2 \leq 4^2|\varphi|^2 E(c_{\cdot}^{M_0}(\varepsilon))^2 \leq C\theta^{M_0}$$

We have

$$Var(\tilde{f}(x)) \leq C(M_0+1)[\sum_{\ell=1}^{M_0} m_\ell^2\theta^\ell + \theta^{M_0}] \tag{5. 4}$$

Where θ is the variance decreasing factor in (4.16).

By combining (5.3) and (5.4), we see that under the condition (5.2) the mean squared error of $\tilde{f}(x)$ goes to zero as M_0 goes to infinity. □

REMARK: There are many sequence $\{m_\ell\}_{\ell=1}^{M_0}$ satisfing the condition (5.2). For example, let $m_\ell = \theta^{\frac{M_0-\ell}{2}}$ for $1 \leq \ell \leq \frac{M_0}{2}$, and $m_\ell = 1$ for $\frac{M_0}{2} < \ell \leq M_0$, then the sequence $\{m_\ell\}_{\ell=1}^{M_0}$ works.

Corollary 5.1 *If the sequence $\{m_\ell\}_{\ell=1}^{M_0}$ is such that*

$$M_0\sum_{\ell=1}^{M_0} m_\ell^2\theta^\ell \to 0 \quad as \quad M_0 \to \infty \tag{5. 5}$$

and if $M - M_0 \to \infty$, then $\tilde{f}$ is a consistent estimator of $f \in C^\alpha$ in the sense that

$$MSE[\tilde{f}(x)] \to 0 \quad as \quad M_0 \to \infty$$

PROOF: $\sum_{\ell=1}^{M_0} |1-m_\ell|2^{-\alpha(M-\ell)} \to 0$ as $M-M_0 \to \infty$. This and (5.5) give the condition (5.2) in Theorem 5.1. □

Corollary 5.2 *$\mathcal{P}_{M-M_0}y(x)$ is a consistent estimator of $f \in C^\alpha$ in the sense that*

$$MSE[\mathcal{P}_{M-M_0}y(x)] \to 0 \quad as \quad M_0 \to \infty \quad and \quad M-M_0 \to \infty$$

PROOF: Set $m_\ell = 0$ in Corollary 5.1. □

Before we introduce a more general theorem, we need the following lemma.

Lemma 5.1 *For constants a, λ, and random variable X with mean 0 and variance σ^2,*

$$Var[(a+X)I_{|a+X|\geq\lambda}] \leq \sigma^2 + \min\{a^2, 4\sigma^2+4\lambda^2\}$$

$$|E[(a+X)I_{|a+X|\geq\lambda} - a]| = |-E[(a+X)I_{|a+X|<\lambda}]| \leq \min\{\lambda, a+\sigma\}$$

PROOF: $Var[(a+X)I_{|a+X|\geq\lambda}] \leq E[(a+X)I_{|a+X|\geq\lambda}-a]^2 = E(X^2 I_{|a+X|\geq\lambda})+a^2E(I_{|a+X|<\lambda})$ If $|a| \leq 2\lambda$ than $a^2E[I_{|a+X|<\lambda}] \leq 4\lambda^2$, otherwise $a^2E[I_{|a+X|<\lambda}] \leq 4\sigma^2$ by Chebyshov inequality. □

Donoho and Johnstone [5] showed that "threshold" nonlinearities provide near-minimax behavior. We consider two possible threshold nonlinearities together with shrinkage on wavelet coefficients to construct the estimation. Let $\{\lambda_\ell\}$ be a seqence of nonnegative numbers.

1. "hard" nonlinearity: $d_n^{*\ell} = (d_n^\ell(f) + d_n^\ell(\varepsilon))1_{\{|d_n^\ell(f)+d_n^\ell(\varepsilon)|\geq\lambda_\ell\}}$

2. "soft" nonlinearity:
$d_n^{*\ell} = (d_n^\ell(f) + d_n^\ell(\varepsilon))1_{\{|d_n^\ell(f)+d_n^\ell(\varepsilon)|\geq\lambda_\ell\}} + \lambda_\ell 1_{\{(d_n^\ell(f)+d_n^\ell(\varepsilon))\leq-\lambda_\ell\}} - \lambda_\ell 1_{\{(d_n^\ell(f)+d_n^\ell(\varepsilon))\geq\lambda_\ell\}}$

Consider the estimator $\hat{f}(x)$ which is the function derived from the multiresolution reconstruction process (3.4) based on the modified wavelet coefficients $\{\tilde{d}_n^{*\ell}\} = \{m_\ell d_n^{*\ell}\}$,

$$\begin{aligned} \hat{f}(x) &= \sum_{\ell=1}^{M_0} m_\ell \mathcal{W}_{M-\ell}^* y(x) + \mathcal{P}_{M-M_0} y(x) \\ &= \sum_{\ell=1}^{M_0} m_\ell[\mathcal{W}_{M-\ell}^* f(x) + \mathcal{W}_{M-\ell}^* \varepsilon(x)] + \mathcal{P}_{M-M_0} f(x) + \mathcal{P}_{M-M_0}\varepsilon(x) \end{aligned} \tag{5. 6}$$

where $\mathcal{W}_{M-\ell}^* y(x)$ is based on the modified coefficients $\{\tilde{d}_n^{*\ell}\}$, and $\{\tilde{d}_n^{*\ell}\}$ is the shrinkage of either the "hard" nonlinearity or the "soft" nonlinearity of $\{d_n^\ell\}$.

Theorem 5.2 *If the sequence $\{m_\ell\}_{\ell=1}^{M_0}$ and $\{\lambda_\ell\}_{\ell=1}^{M_0}$ are such that*

$$\sum_{\ell=1}^{M_0}[|1-m_\ell|2^{-\alpha(M-\ell)} + M_0 m_\ell^2(\theta^\ell + \lambda_\ell^2)] \to 0 \quad as \quad M_0 \to \infty \tag{5. 7}$$

then $\hat{f}$ is a consistent estimator of $f \in \mathcal{C}^\alpha$ in the sense that

$$MSE[\hat{f}(x)] \to 0 \quad as \quad M_0 \to \infty$$

PROOF: The absolute value of the bias of $\hat{f}(x)$ is

$$\begin{aligned} |bias(\hat{f}(x))| &= |E(\hat{f}(x) - f(x))| \\ &= |\sum_{\ell=1}^{M_0} m_\ell E\mathcal{W}_{M-\ell}^* y(x) + E\mathcal{P}_{M-M_0} y(x) - f(x)| \end{aligned}$$

$$
\begin{aligned}
&= |\sum_{\ell=1}^{M_0}(m_\ell E\mathcal{W}^*_{M-\ell}y(x) - \mathcal{W}_{M-\ell}f(x)) + \mathcal{P}_M f(x) - f(x)| \\
&\le \sum_{\ell=1}^{M_0} m_\ell |E\mathcal{W}^*_{M-\ell}y(x) - \mathcal{W}_{M-\ell}f(x)| + \\
&\quad \sum_{\ell=1}^{M_0} |1 - m_\ell||\mathcal{W}_{M-\ell}f(x)| + |\mathcal{P}_M f(x) - f(x)| \\
&= A_0 + A_1 + A_2 \qquad (5.\ 8)
\end{aligned}
$$

We have seen in the proof of theorem 5.1 that A_1 and A_2 go to zero as M_0 increase.

In the case of "hard" nonlinearity,

$$|Ed_n^{*\ell} - d_n^\ell(f)| = | - E[d_n^\ell I_{|d_n^\ell| \le \lambda_\ell}]| \le \lambda_\ell$$

In the case of "soft" nonlinearity,

$$
\begin{aligned}
|Ed_n^{*\ell} - d_n^\ell(f)| &\le |E[(d_n^\ell(f) + d_n^\ell(\varepsilon))1_{\{|d_n^\ell(f)+d_n^\ell(\varepsilon)| \ge \lambda_\ell\}}] - d_n^\ell(f)| + \\
&\quad \lambda_\ell |E[1_{\{(d_n^\ell(f)+d_n^\ell(\varepsilon)) \le -\lambda_\ell\}} - 1_{\{(d_n^\ell(f)+d_n^\ell(\varepsilon)) \ge \lambda_\ell\}}]| \\
&\le 2\lambda_\ell
\end{aligned}
$$

thus

$$A_0 \le \sum_{\ell=1}^{M_0} m_\ell \sum_n |Ed_n^{*\ell} - d_n^\ell(f)||\psi(2^{M-\ell}x - n)| \le \sum_{\ell=1}^{M_0} 14|\psi| m_\ell \lambda_\ell \le 14|\psi| M_0 [\sum_{\ell=1}^{M_0} m_\ell^2 \lambda_\ell^2]^{\frac{1}{2}}$$

which approaches to zero as M_0 increase because of the condition (5.7).

Now the variance of $\hat{f}(x)$ is

$$
\begin{aligned}
Var(\hat{f}(x)) &= E[\sum_{\ell=1}^{M_0} m_\ell \sum_n (d_n^{*\ell} - Ed_n^{*\ell})\psi(2^{M-\ell}x - n) + \mathcal{P}_{M-M_0}\varepsilon(x)]^2 \\
&\le (M_0 + 1)[\sum_{\ell=1}^{M_0} m_\ell^2 E(\sum_n (d_n^{*\ell} - Ed_n^{*\ell})\psi(2^{M-\ell}x - n))^2 + E(\mathcal{P}_{M-M_0}\varepsilon(x))^2]
\end{aligned}
$$

By lemma 5.1, for the "hard" nonlinearity of d_n^ℓ

$$E(d_n^{*\ell} - Ed_n^{*\ell})^2 \le 5E(d_n^\ell(\varepsilon))^2 + 4\lambda_\ell^2 \le C(\theta^\ell + \lambda_\ell^2)$$

for the "soft" nonlinearity of d_n^ℓ

$$
\begin{aligned}
E(d_n^{*\ell} - Ed_n^{*\ell})^2 &\le 2Var[(d_n^\ell(f) + d_n^\ell(\varepsilon))1_{\{|d_n^\ell(f)+d_n^\ell(\varepsilon)| \ge \lambda_\ell\}}] + \\
&\quad 2\lambda_\ell^2 Var[1_{\{(d_n^\ell(f)+d_n^\ell(\varepsilon)) \le -\lambda_\ell\}} - 1_{\{(d_n^\ell(f)+d_n^\ell(\varepsilon)) \ge \lambda_\ell\}}] \\
&\le 10(d_n^\ell(\varepsilon))^2 + 8\lambda_\ell^2 + 2\lambda_\ell^2 \\
&\le C(\theta^\ell + \lambda_\ell^2)
\end{aligned}
$$

therefore we have

$$Var(\hat{f}(x)) \leq C(M_0 + 1)[\sum_{\ell=1}^{M_0} m_\ell^2(\theta^\ell + \lambda_\ell^2) + \theta^{M_0}] \tag{5. 9}$$

Where θ is the variance decreasing factor in (4.16).

By combining (5.8) and (5.9), we see that under the condition (5.7) the mean squared error of $\hat{f}(x)$ goes to zero as M_0 goes to infinity. □

REMARK: Theorem 5.1 is a special case of theorem 5.2 where we do not threshold the wavelet coefficients. It corresponds to setting all λ_ℓ to zero.

Proposition 5.1 *If the sequences $\{m_\ell\}_{\ell=1}^{M_0}$ is such that*

$$M_0 \sum_{\ell=1}^{M_0} m_\ell^2 \theta^\ell \to 0 \quad as \quad M_0 \to \infty \tag{5. 10}$$

and if $M - M_0 \to \infty$, then, for any hard nonlinearity thresholds $\{\lambda_\ell\}_{\ell=1}^{M_0}$, $\hat{f}$ is a consistent estimator of $f \in C^\alpha$ in the sense that

$$MSE[\hat{f}(x)] \to 0 \quad as \quad M_0 \to \infty$$

PROOF: The absolute value of the bias of $\hat{f}(x)$ still has upper bound as in (5.8). The terms A_1 and A_2 go to zero because of $M - M_0 \to \infty$. Since by lemma 5.1,

$$|Ed_n^{*\ell} - d_n^\ell(f)| \leq |d_n^\ell(f)| + \sqrt{Var(d_n^\ell(\varepsilon)} \leq C(2^{-\alpha(M-\ell)} + \theta^{\ell/2})$$

We have that

$$A_0 \leq \sum_{\ell=1}^{M_0} m_\ell |\sum_n |Ed_n^{*\ell} - d_n^\ell(f)||\psi(2^{M-\ell}x - n)| \leq C \sum_{\ell=1}^{M_0} m_\ell (2^{-\alpha(M-\ell)} + \theta^{\ell/2})$$

which approaches to zero as M_0 increase because of the condition (5.10).

By lemma 5.1

$$E(d_n^{*\ell} - Ed_n^{*\ell})^2 \leq E(d_n^\ell(\varepsilon))^2 + |d_n^\ell(f)|^2 \leq C(\theta^\ell + 2^{-2\alpha(M-\ell)})$$

Similar to (5.9), we have

$$Var(\hat{f}(x)) \leq C(M_0 + 1)[\sum_{\ell=1}^{M_0} m_\ell^2(\theta^\ell + 2^{-2\alpha(M-\ell)}) + \theta^{M_0}] \tag{5. 11}$$

which goes to zero as M_0 increases because of (5.10) and $M - M_0 \to 0$.

Therefore the mean squared error of $\hat{f}(x)$ goes to zero as M_0 goes to infinity under the conditions assumed. □

6 A Numerical Example

Here is an example. Let $f(x)$ be a function as in the upper-left box in Figure 2. We observe f with error $\{\varepsilon_i\}_{i=0,\cdots,250}$ at the points $\{i/250\}_{i=0,\cdots,250}$ to get the sample $\{y_i\}_{i=0,\cdots,250} = \{f(i/250) + \varepsilon_i\}_{i=0,\cdots,250}$ as plotted in the upper-right box. The error is distributed normally with mean $\mu = 0$, and variance $\sigma^2 = 1/9$. In the low-left box is the representation function of the sample. And we put the estimator together with the underlying function and the sample in the low-right box.

In Figure 3, the upper-left box contains the reconstruction of f, f is well preserved through the procedure. The upper-right box is the reconstruction from the noise, there is a little trend in the noise.

ACKNOWLEDGEMENTS: I would like to thank Professor Mary Ellen Bock for many constructive discussions. Special thanks to Professors David Donoho and Iain Johnstone for giving me insight guidances as well as references and manuscripts.

Zhongcheng Wang
Department of Statistics
Purdue University
West Lafayette, IN47907
zwang@pop.stat.purdue.edu

References

[1] Chui, C. K. (1992). *An Introduction to Wavelets.* Academic Press: Boston.

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

[3] Meyer, Y. (1990). *Ondelettes.* Hermann: Paris.

[4] Donoho, D. L. & Johnstone, I. M. (1992). Adapting to Unknown Smoothness at a Point via Wavelet Shrinkage. (Technical Report, Department of Statistics, Stanford University).

[5] Donoho, D. L. & Johnstone, I. M. (1992). Ideal Spatial Adaptation by Wavelet Shrinkage. (Technical Report, Department of Statistics, Stanford University).

[6] Chui, C. K. & Wang, J. Z. (1991). A Cardinal Spline Approach to Wavelets. *Proceedings of the American Mathematical Society.* **113**, 787–795.

[7] Chui, C. K. & Wang, J. Z. (1991). On Compactly Supported Spline Wavelets and a Duality Principle. To appear, *Trans. Amer. Math. Soc.*

[8] Cohen, A., Daubechies, I., & Feauveau, J. C. (1991). Biorthogonal Bases of Compactly Supported Wavelets. To appear, *Comm. Pure and Appl. Math.*

[9] Brockwell, P. J., & Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer-Verlag: New York.

[10] Wang, Z. (1992). *Some Properties of Spectral Density Functions in Down Sampling a Stationary Process and a Version of Ruelle's General Perron-Froebenius Theorem.* (Technical Report, Department of Statistics, Purdue University).

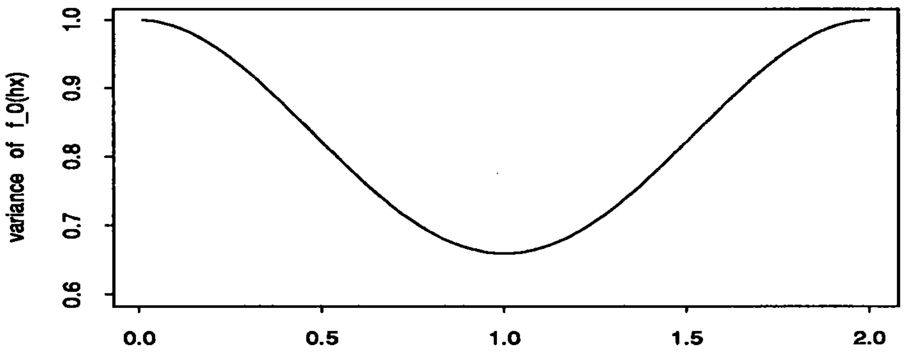

Figure 1: Variance of $\mathcal{P}_M y(hx)/\sigma$

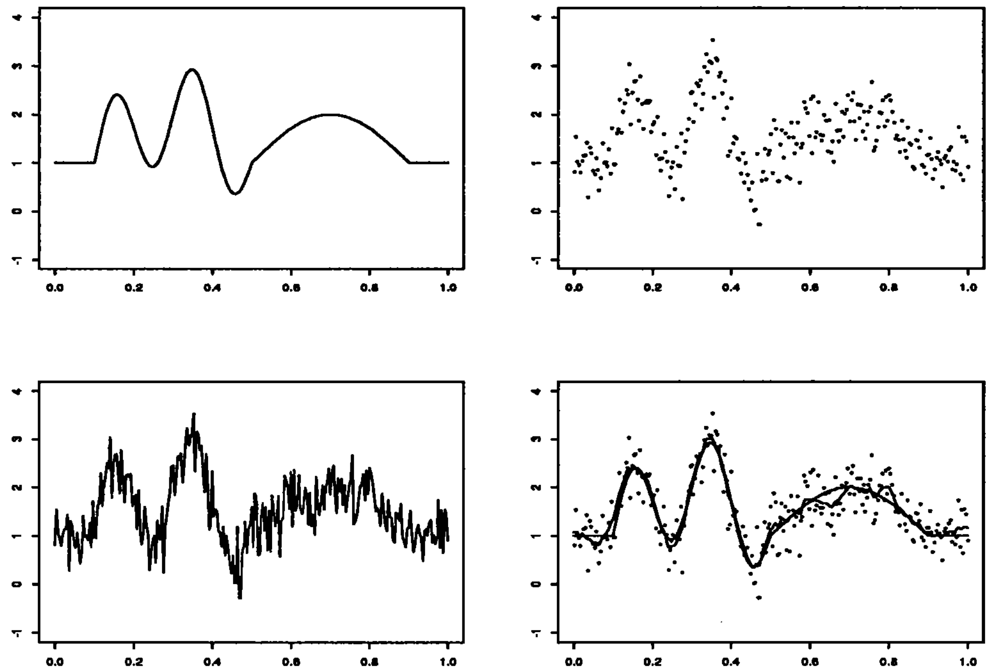

Figure 2: Example of the procedure

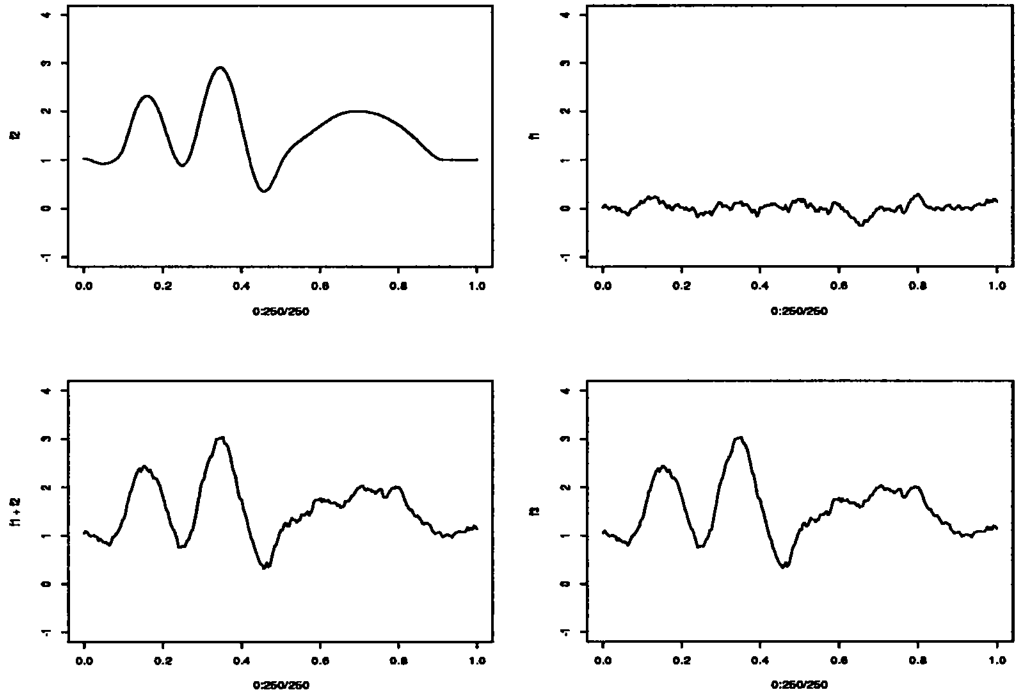

Figure 3: Additive of the Decomposition