Ordinary Least Squares Estimate of the Fractional Differencing Parameter Using Wavelets as Derived from Smoothing Kernels

Mark J. Jensen
Department of Economics, Southern Illinois University
Carbondale, IL 62901

May 25, 1995

Abstract

This paper develops a consistent OLS estimate of a fractionally integrated processes' differencing parameter, using continuous wavelet theory as constructed from smoothing kernels. We show that a log-log linear relationship exists between the variance of the wavelet coefficient and the level at which the fractionally integrated processes is smoothed. This linear relationship occurs because the self-simularity property of the fractionally integrated process and the self-similarity of the wavelet causes the smoothing level to continually appear in the wavelet transformation. Since the wavelet coefficient can be interpreted as the kth order details of the series at some level of smoothing, we also show that the above log-log relationship can be derived from the variance of the 1st order derivative of the time series smoothed by a kernel that is well localized in both time and frequency space. Lastly, we derive the asymptotic biasness and variance of the OLS estimate and test our consistent estimate with a number of Monte Carlo experiments.

Keywords: Fractionally Integrated Processes, Long-Memory, Smoothing Kernels, Wavelets

1 Introduction

Wavelet analysis is a relatively new development in the area of applied mathematics that is just now receiving the attention of statisticians[1]. Wavelets were first introduced in seismology [Morlet (1983)] to provide a time dimension to seismic analysis that Fourier analysis lacked. The wavelets generality and its strong results quickly revealed that they were advantageous in other areas ranging from signal [Kronland-Martinet, Morlet and Grossman (1987)] to numerical analysis [Beylkin, Coifman, and Rokhlin (1991)].

By design the wavelets usefulness rests in their ability to localize a process in time-frequency space. At high frequency levels, the wavelet is tight in shape (small time interval) and is able to focus in on short lived phenomena like singularity points, while at low frequencies the wavelet is stretched out in shape, making it well suited in identifying long periodic processes. By moving from high to low levels of frequency the wavelet is able to zoom in on a processes behavior at a point in time and identify either singularities or alternatively zoom out and reveal the long and smooth features of a signal.

Wavelets can be thought of as the derivative at any order k of a smoothing kernel, under the assumption that the smoothing kernel has at least k-ordered derivatives. Like any smoothing kernel, the kernel from which wavelets are formed is well localized in time space. But unlike normal unimodular smoothing kernels, the smoothing kernel used in deriving a wavelet can take on negative value. This feature of the smoothing kernel enables the wavelet to be well localized in frequency space, improving the decorrelation between the wavelet coefficients, and enabling the wavelet's bandwidths to be increased (decreased) to capture the long and smooth (short and discontinuous) characteristics of a time series.

Suppose the bandwidth of a smoothing kernel is smaller than the interval over which a time series features a periodic behavior. In order for the smoothing kernel to correctly characterize this periodicity, convolutions of the kernel and time series must be spread out over time and calculated often over the periodic interval. If the characteristic of the time series involves a time interval that is smaller than the kernel's bandwidth, then a large number of convolutions must be spread out in frequency. In both situations the smoothing

[1] See Donoho and Johnstone (1994), (1995a), (1995b), and Donoho, et.al. (1995)

kernel is dependent on the chosen bandwidth. Since wavelets use every possible bandwidths available in analyzing a time series, wavelet analysis does not suffer from these problems.

In this paper we show that there exists a log-log linear relationship between the variance of the wavelet coefficient and the bandwidth, equal to the fractional differencing parameter. This linear relationship lends itself nicely to the estimation of the fractional differencing parameter by ordinary least squares. We show that the OLS estimate is a consistent estimate of the fractional differencing parameter and derive its small sample biasness and variance.

In Section 2 we provide the background and motivation for using wavelets. Some of the material of Section 2 can be found in other treatments on wavelets [Daubechies (1988) and Mallat (1989)], but a large percentage of Section 2 is presented in a manner that is parculiar to a wavelet being defined as the kth order derivative of a smoothing kernel. Section 3 defines the class of fractionally integrated processes, their self-similarity and long-memory properties. In Section 4 we draw heavily on the self-similarity property of the fractionally integrated processes and of the wavelet to establish the log-log linear relationship between the variance of the wavelet coefficient and the fractional differencing parameter. We then provide some of the asymptotic properties of the OLS estimate. Lastly, in Section 5 we conduct a Monte Carlo simulation study to determine the robustness of the OLS estimate of the fractional differencing parameter to different values of the fractional differencing parameter and signal length.

2 Wavelet Theory

A continuous wavelet is defined as

$$\psi_{a,b}(t) \quad = \quad |a|^{-1/2}\psi\left(\frac{t-b}{a}\right) \tag{1}$$

where $a \neq 0$ and b is any real number. $\psi_{a,b}$ is simply the dilation (by a) and translation (by b) of the function ψ. If $a > 1$, ψ flattens out horizontally, while $0 < a < 1$ tightens ψ.[2] For this reason a is referred to as the scaling parameter. The $|a|^{-1/2}$ term is a normalizing

[2]In theory a can take on any real value other than zero. However, we make the simplifying assumption of restricting a to the unit interval. The results we find are also obtainable with scaling parameters outside the unit interval with only slight modifications to the notation.

constant that insures that $\psi_{a,b}$ has an inner product equal to one. When $a > 1$ the $|a|^{-1/2}$ term causes the vertical height of ψ to be scaled down, while when $0 < a < 1$ the vertical height is increased. If ψ is well localized around zero, changing the value of b shifts ψ over the time arguments, allowing $\psi_{a,b}$ to be well localized around the translation point b. b is referred to as the translation parameter.

The function $\psi(t)$ is commonly referred to as the 'mother' wavelet. In order for a function to qualify as a 'mother' wavelet it must satisfy the admissibility condition, $\int \psi(t)dt = 0$. This is a necessary condition insuring smoothness and localization in frequency and time space. The admissibility condition can also be interpreted as requiring $\psi(t)$ to be nonunimodular, hence, the name wavelets.

The regularity properties of the 'mother' wavelet can be extended to require $\psi(t)$ to have more than one vanishing moments or to possess higher ordered continuous derivatives, i.e. $\int t^r \psi(t)dt = 0$ where $r = 0, 1, 2, \ldots, M-1$, and/or $\psi(t) \in C^k$. Increasing the number of vanishing moments is important in the results of this paper and will become more apparent once the theory is fully developed.

Let $x(t)$ be a real valued time series. The wavelet coefficients for $x(t)$ are equal to the inner product

$$< x, \psi_{a,b} >= |a|^{-1/2} \int x(t) \psi \left(\frac{t-b}{a} \right) dt. \tag{2}$$

Later on we will show that the wavelet coefficients $\{< x, \psi_{a,b} >\}_b$ represent the details of the signal $x(t)$ at the scale a.[3]

Grossman and Morlet (1984), desiring to find a transformation who's dilations and translation would represent a function in Hardy space and could be used to reconstruct the function, found that the wavelet coefficients, $< x, \psi_{a,b} >$, completely characterizes x in the L^2 sense of

$$\int | < x, \psi_{a,b} > |^2 db \quad = \quad \int |x(t)|^2 dt$$

and could be used to reconstruct x by

$$x \quad = \quad C_\psi^{-1} \int \int a^{-2} < x, \psi_{a,b} > \psi_{a,b} \, da \, db$$

[3]Mallat (1989) used multiresolution analysis to show that discrete wavelets coefficients represent the details of a signal at a given scale and that the wavelet forms a basis of $L^2(\Re)$.

where $C_\psi^{-1} = 2\pi \int |\hat{\psi}(\xi)|^2 |\xi|^{-1} < \infty$. Note that the admissibility condition, $\int \psi(x)dx = 0$, is implied by $C_\psi^{-1} < \infty$ if $\psi(t)$ has sufficient decay.

Thus, $\psi_{a,b}$ constitutes a basis for the Hardy space of functions, and more regular wavelets have been shown by others to form a basis for L^2, Sobolev and Hölder spaces [Meyer (1990)]. If ψ is carefully constructed, $\psi_{a,b}$ also provides an orthonormal basis for these function spaces. The wavelet we have chosen to use in this paper are of this construction.

Let $\phi(t)$ be a function where

$$\int \phi(t)dt = 1 \tag{3}$$

$$\int t^r \phi(t)dt = 0 \qquad r = 1, 2, \ldots, M-1 \tag{4}$$

and $\partial^{k-1}\phi(t)/\partial t$ has either a compact support or

$$\lim_{t \to \pm\infty} \frac{\partial^{k-1}\phi(t)}{\partial t} = 0. \tag{5}$$

We assume that ψ has k ordered derivatives. In the wavelet literature $\phi(t)$ is known as the scaling function. Since in general a function only needs to satisfy Eq. (3) in order to be a scaling function, Eq (4) and (5) defines a certain type of scaling function. In particular, $\phi(t)$ has some vanishing moments.

For our purposes $\phi(t)$ can be thought of as a smoothing kernel who's expected value is zero, but unlike a typical smoothing kernel $\phi(t)$ can take on negative values. By viewing $\phi(t)$ as a smoothing kernel, the scaling coefficient, $< x, \phi >$, can be interpreted as the average value of $x(t)$, and the dilated and translated scaling coefficient, $< x, \phi_{a,b} >$, represent the average of $x(t)$, at resolution a, around b. Note that $\phi_{a,b}$ is the dilation and translation of $\phi(t)$.

For expository purposes we define the 'mother' wavelet in terms of the scaling function ϕ as

$$\psi(t) = (-1)^k \frac{\partial^k}{\partial t} \phi(t). \tag{6}$$

The wavelet admissibility condition follows as a result of Eq. (5). From this definition the wavelet coefficient can be interpreted as a measure of the rate of change (details) in the

average of $x(t)$ at scale a. This is found by using the definition of the wavelet coefficient and the 'mother' wavelet to obtain

$$\begin{aligned} < x, \psi_{a,b} > &= \int x(t)\psi_{a,b}\, dt \\ &= (-1)^k |a|^k \frac{\partial^k}{\partial t} \int x(t)\phi_{a,b}(t)dt \\ &= |a|^k \frac{\partial^k}{\partial b} \int x(t)\phi_{a,b}(t)dt \\ &= |a|^k \frac{\partial^k}{\partial b} < x, \phi_{a,b} > . \end{aligned} \tag{7}$$

Thus, the wavelet coefficient is proportional to the kth order rate of change in the average of $x(t)$ around b at the scale a. It follows that $\{< x, \psi_{a,b} >\}_b$ is a measure of the kth order details of $x(t)$ at scale a.

Other interpretations that aid in understanding the scaling and wavelet coefficients are found by expanding $x(t)$ in a Taylor series around b. Assuming that $x(t) \in C^{M-1}$

$$x(t) = x(b) + x'(b)(t-b) + \cdots + \frac{1}{M-1} x^{(M-1)}(b)(t-b)^{M-1} + C(t)(t-b)^M$$

where $C(t)$ is a finite constant dependent on t. Multiplying by $\psi_{a,b}$ and integrating over t we find

$$< x(t), \phi_{a,b} > \leq |a|^{1/2} x(b) + K|a|^{M-1/2} \int t^m \phi(t)dt. \tag{8}$$

for some finite constant K. It also follows from the above Taylor series that the absolute value of the wavelet coefficients,

$$| < x, \psi_{a,b} > | \leq K|a|^{M-1/2} \int |\psi(y)||y|^M dy. \tag{9}$$

Since both the level of approximation achieved by the scaling coefficient and the rate of decay of the wavelet coefficients depend on the number of vanishing moments, M, Eq. (8) and (9) show the benefits of choosing a wavelet with a large number of vanishing moments. We use the benefit of decaying wavelet coefficients to provide statistical inference for the fractional differencing parameter.

2.1 Example

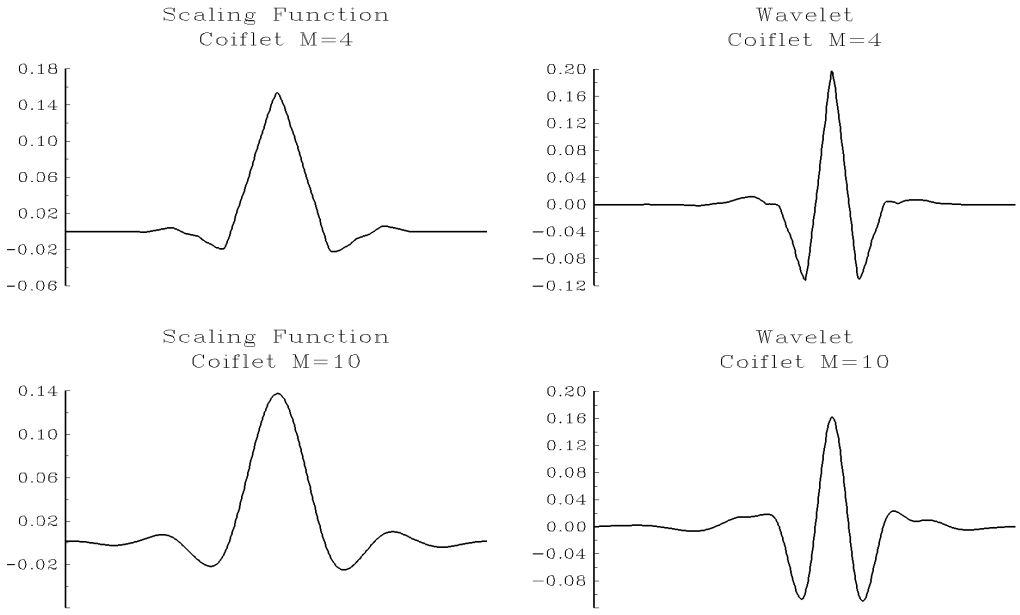

Figure 1: Coiflet Scaling Functions and Wavelets

An example, and the chosen wavelet for the simulations performed in this paper, is the Coiflet wavelet [Daubechies (1992) p. 258]. The Coiflet family of wavelets are derived from a scaling functions with a nonzero zero-order moment and at least one vanishing moment at higher orders. The corresponding wavelets have just as many vanishing moment as their scaling function. In fig. 1 the Coiflet's scaling and wavelet functions are plotted for $M = 4, 10$. Notice that in both cases the scaling functions are well localized in time around zero. Also notice that by increasing M, $\psi(t)$ and $\phi(t)$ become smoother, which in turn improves their regularity properties.

We earlier showed that the approximating ability of $< x, \phi_{a,b} >$ and the rate of decay in $< x, \psi_{a,b} >$ improves as M increases. However, as the number of vanishing moments increases the size of the wavelet's support also increases [Daubechies (1992)]. If $\psi(t)$ is a Coiflet wavelet with M vanishing moments, the width of $supp(\psi(t)) = 3M - 1$.[4] This property is easily seen in the scaling function of fig. 1 where the support for $M = 10$ extends well beyond the support for $M = 4$.

[4]The Daubechies (1988) family of wavelets possess the smallest possible support for M vanishing moments. If the wavelet has M vanishing moments the smallest support width a wavelet can have is $2M$. Hence, the Daubechies wavelet is superior to the Coiflet wavelet because of its support size. However, the Daubechies wavelet is asymmetric in time dimension and for this reason we choose to use the Coiflet wavelet.

To show how simple a function can be, and still qualify as a wavelets, we present one of the oldest and well known wavelets. The Haar wavelet is a member of the Daubechies class, the symmetric biorthogonal class and many other classes of wavelets, and equals

$$\psi(t) = \begin{cases} 1, & 0 \leq t < 1/2 \\ -1, & 1/2 \leq t < 1 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Hence, it follows that the corresponding scaling function is the indicator function $\chi(t)_{[0,1)}$.

3 Fractionally Integrated Series

Let $x(t)$ represent a time series with the form

$$(1-L)^d x(t) = \epsilon(t) \tag{11}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Without losing any generality we assume that $\sigma_\epsilon^2 = 1$. The lag operator, L, is defined as $Lx(t) = x(t-1)$. $x(t)$ is an integrated series with differencing parameter d, denoted as $I(d)$, where d is normally an integer. Granger and Joyeux (1980) and Hosking (1981) assume that, in addition to taking on integer values, d can also equal non-integer values, in particular $|d| < 0.5$. Their reasoning is that many series have spectrums that increase without bound as $\omega \to 0$, but when the series is differenced the spectrum equals zero. Furthermore, by allowing d to be a non-integer, a wide variety of processes with spectrums of different shapes are made available with the power spectrum,

$$S(\omega) \sim 1/|\omega|^{2d} \qquad \text{as } \omega \to 0. \tag{12}$$

In Eq. (12), Mandelbrot and Van Ness' (1968) statistical self-similarity property is evident. For any a, $S_x(\omega) = |a|^{2d} S_x(a\omega)$, i.e. the statistical properties of $x(t)$ remain the same regardless of the resolution of $x(t)$.

$x(t)$'s autocovariance function is

$$\begin{aligned} R_x(k) &= E[x(t)\, x(t+k)] \\ &= K|k|^{2d-1} \end{aligned} \tag{13}$$

as $|k| \to \infty$, for some finite constant K. Since

$$\begin{aligned} R_x(ak) &\approx |a|^{2d-1} R_x(k) \\ &= |a|^{2H} R_x(k) \end{aligned} \tag{14}$$

the statistical self-similarity parameter is $H = d - 1/2$.

In Eq. (13), the long-term memory property of the fractional integrated model is also evident. When $d > 0$ the autocovariance function decays hyperbolically, as opposed to a short-memory processes' exponential decay. This same slow decay also occurs when $d < 0$. However, instead of having positive long-term dependence the process has negative dependence.

It is possible to write the variance of the scaling coefficients as a function of x's autocovariance function, $R_x(k)$. By taking advantage of the self-similarity of $x(t)$, each time the scaling factor a decreases it is like measuring the $I(d)$ process at points found in between those already measured. As a result of this increasing resolution the a^{2d-1} term continues to pop up. If the wavelet is defined as the 1st order derivative of the scaling function, the self-similarity of $R_x(k)$ and the scaling parameter has a relationship equal to

$$var(< x, \psi_{a,b} >) = |a|^{2d} R_x(k).$$

It will become more apparent how this statistical self-similarity property of an $I(d)$ process allows us to estimate d, but for the moment we only need to stress how important this behavior is in reaching our results. It enables us to measure $x(t)$ are different levels of resolution and to know that the autocovariance functions from different levels resolutions will be related.

4 OLS Estimate of d Using Wavelets

Let the series $x(t)$ be a mean zero fractionally integrated process with $|d| < 1/2$, and define

$$R(a) = E[< x, \psi_{a,b} >^2]. \tag{15}$$

By the wavelet definition given in Eq. (6), $R(a)$ can be written in terms of the scaling coefficients as

$$R(a) \quad = \quad |a|^{2k} E\left[\left\{\frac{\partial}{\partial b} < x, \phi_{a,b} >\right\}^2\right]. \tag{16}$$

Hence, $R(a)$ is a measure of the variance in the kth order difference of $x(t)$ when the series is smoothed to the scale a.

We now draw on the statistical properties of the wavelet coefficients as spelled out in the following theorem to obtain a consistent estimate of the $I(d)$'s fractional differencing parameter. From the statistical properties of a fractionally integrated process, we arrive at the follow theorem.

Theorem 1 *The wavelet coefficients,* $< x(t), \psi_{a,b} >$, *associated with a mean zero* $I(d)$ *process with* $|d| < 1/2$ *are distributed* $\mathcal{N}(0, \sigma^2|a|^{2d})$, *where* σ^2 *is a finite constant.*

Proof: See Appendix A.

By Theorem 1, $R(a) = \sigma^2|a|^{2d}$. Taking the logarithmic transformation of $R(a)$, we obtain the log-log relationship

$$\ln R(a) \quad = \quad \ln \sigma^2 + d \ln |a|^2 \tag{17}$$

where $\ln R(a)$ is linearly related to $\ln |a|^2$ by the fractional differencing parameter, d. Hence, the unknown d of a fractionally integrated series can be calculated by the ordinary least squares estimate, $\hat{d}$.

To perform this OLS regression we require a measure for the dependent variable $R(a)$. Fortunately, if a large number of wavelet coefficients are available for the scale a, the sample variance of $< x, \psi_{a,b} >$ provides us with a consistent estimate of $R(a)$. At scale a and $x(t)$ measured over $t = 0, 1, \ldots, T$, where $T \geq a^{-1} - 1$, the sample variance of the wavelet coefficients equals

$$\bar{R}(a) = \frac{1}{N(a)} \sum_{n=0}^{a^{-1}-1} < x, \psi_{a,b_n} >^2 \tag{18}$$

where $b_n = a * n$ and $N(a) = a^{-1}$.

The statistical properties of the OLS estimate $\hat{d}$ can be calculated by expanding $\ln \bar{R}(a)$ around $\ln R(a)$ in the following Taylor series

$$\ln \bar{R}(a) \quad = \quad \ln R(a) + \frac{\bar{R}(a) - R(a)}{R(a)} - \frac{1}{2}\left[\frac{(\bar{R}(a) - R(a))^2}{R(a)^2}\right]. \tag{19}$$

In order to show that $\hat{d}$ is a consistent estimate of the fractional differencing parameter, we need the following theorem showing that the $< x, \psi_{a,b} >$ are asymptotically independent.

Theorem 2 *If $\psi(t)$ has $M \geq 1$ vanishing moments with support $[-K_1, K_2]$ where $K_1 \geq 0$ and $K_2 \geq 0$ and $x(t)$ is $I(d)$ with $|d| < 1/2$ then $< x, \psi_{a,b} >$ is asymptotically independent in both time and scale space since $< x, \psi_{a,b} >$'s correlation decays as $\mathcal{O}(|b_1 - b_2|^{2(d-M)-1})$ in time space and as $\mathcal{O}(|1/a|^{2(d-M)-1})$ in scale space, for all b_1 and b_2 such that $|b_1 - b_2| > K_1 + K_2$.*

Proof: See Appendix B.

Since $|d| < 1/2$, $2(d - M) - 1$ will always be negative and the correlation will always decay even for the least regular of wavelets, eg. the Haar wavelet. However, as we mentioned early in Section 2, the larger the number of vanishing moments the wider the wavelet's support, i.e. larger values for K_1 and K_2 and less wavelet coefficients satisfying the condition $|b_1 - b_2| > K_1 + K_2$. Thus, by choosing a wavelet with a large M, the rate of decay in $< x, \psi_{a,b} >$'s autocovariances will increase, but over a smaller set of wavelet coefficients.

Fortunately, even though theory says the decay of the wavelet correlation only occur when the difference in the translation parameters are outside the cone $K_1 + K_2$, simulations studies have shown that the effective support of a wavelet is much smaller than their theoretical support[5]. In these simulations the wavelet coefficient's correlation were found to rapidly decay for translation arguments inside the cone. We too have found this to be the case as evidenced in fig. 2.

In fig. 2, we graph the autocovariance function of $< x, \psi_{a,b} >$, where $x \sim I(0.25)$ and $a = 2^{-8}$. Notice that the autocovariances easily falls within two standard errors, $\pm 2/\sqrt{a^{-1}}$, of zero.

By Theorem 1 and 2, $R(a)^{-1/2} < x, \psi_{a,b} > \sim \mathcal{N}(0, 1)$ and is asymptotically independent as a tends to zero. Hence, $R(a)^{-1} \sum_b < x, \psi_{a,b} >^2 \sim \chi^2_{a^{-1}}$, where a^{-1} is the degree of freedom.

[5]See Daubechies (1988), Tewfik and Kim (1992), Kaplan and Kuo (1993), and Flandrin (1991)

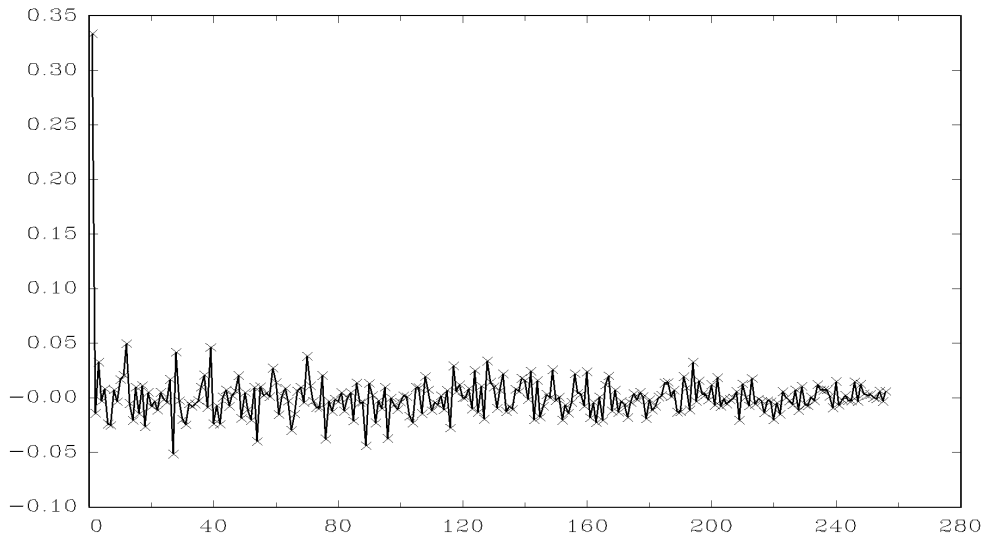

Figure 2: Autocovariance Function of $< x, \psi_{a,b} >$, where $a = 2^{-8}$

It follows that $\bar{R}(a)$ has mean $R(a)$ and variance $2(\sigma^2|a|^{2d+1/2})^2$. Because $\frac{1}{a^{-2}} var(\sum_b < x, \psi_{a,b} >^2)$ converges to zero as a tends to zero, by Markov's law of large numbers $\bar{R}(a)$ will tend in probability to $R(a)$ as $a \to 0$.

The above Taylor series representation of $\ln \bar{R}(a)$ can now be written as

$$\ln \bar{R}(a) \quad = \quad \ln R(a) + o_p(1).$$

Substituting $\sigma^2|a|^{2d}$ for $R(a)$, we find

$$\ln \bar{R}(a) \quad = \quad \ln \sigma^2 + d \ln |a|^2 + o_p(1). \tag{20}$$

In other words, as $a \to 0$ the OLS estimate of the log-log relationship's slope provides a consistent estimate of $x(t)$'s fractional differencing parameter.

4.1 Discrete Wavelet Analysis

Because an observed $I(d)$ processes is discrete, whereas our theory is for continuous time, we must think of the measured data as a approximation of $x(t)$ that has been uniformly sampled at a scale equal to the inverse of the series' length. In other words, the measured fractionally integrated series $x(t)$, where $t = 0, 1, \ldots, T-1$, equals the scaling coefficients

$< x, \phi_{1/T,b} >$ where $b = 0, 1, \ldots, T-1$. From this interpretation it is safe to say that the scaling and wavelet coefficients associated with higher resolution ($a < 1/T$) are meaningless since there is no information contained in between the observed values of $x(t)$. Hence, the finest scale the wavelet coefficients can be calculated is $a = 1/T$.

When analyzing a measured time series the discrete wavelet

$$\psi_{2^{-m},2^{-m}n}(t) = \psi\left(\frac{t - 2^{-m}n}{2^{-m}}\right)$$

where $m = 0, 1, \ldots, \log_2 T - 1$ and $n = 0, 1, \ldots, 2^m - 1$, is commonly used. When a time series has T observations we will have at most $[\log_2 T]$ observations in the regression of $\ln R(a)$ on $\log |a|^2$. As a result, the number of observations used in the OLS regression is dependent on the observations of $x(t)$ doubling in number.

The scaling parameter is a multiple of 1/2 so that the wavelet coefficients can be calculated with the Fast Wavelet Transform [Mallat (1989)]. Like the Fast Fourier Transform, the Fast Wavelet Transform uses quadrature mirror filters that zoom out on the original series $x(t)$ by a factor of 2. Both the Fast Wavelet and Fourier transformation require 2^{max} observations where $max \in \boldsymbol{Z}^+$ or zero padding is used.

4.2 Biasness of $\hat{d}$

Let $y_m = \ln |a_m|^2 - \frac{1}{max}\sum_{k=0}^{max-1} \ln |a_k|^2$, where $a_m = 2^{-m}$. It follows that the OLS estimate of the fractional differencing parameter, $\hat{d} = [\sum_{m=0}^{max-1} y_m^2]^{-1}[\sum_{m=0}^{max-1} y_m \ln \bar{R}(a_m)]$. Expanding $\hat{d}$ in a Taylor series around $R(a_m)$ we find

$$\begin{aligned} \hat{d} &= \left[\sum_{m=0}^{max-1} y_m^2\right]^{-1} \left[\sum_{m=0}^{max-1} y_m \ln R(a_m)\right] + \left[\sum_{m=0}^{max-1} y_m^2\right]^{-1} \\ &\quad \times \left[\sum_{m=0}^{max-1} y_m \frac{\bar{R}(a_m) - R(a_m)}{R(a_m)}\right] + \mathcal{O}_p\left(\frac{var\,\bar{R}(a_m)}{R(a_m)^2}\right) \end{aligned} \tag{21}$$

Substituting $\sigma^2 |a_m|^{2d}$ for the first $R(a_m)$ in the RHS of (21), the biasness of $\hat{d}$ is calculated to equal

$$\hat{d} - d = \left[\sum_{m=1}^{max-1} y_m^2\right]^{-1} \left[\sum_{m=0}^{max-1} y_m \frac{\bar{R}(a_m) - R(a_m)}{R(a_m)}\right] + \mathcal{O}_p\left(\frac{var\,\bar{R}(a_m)}{R(a_m)^2}\right). \tag{22}$$

Since $\bar{R}(a_m)$ tends to $R(a_m)$ as $a_m \to 0$ and $\sum_m y_m^2$ is bounded away from zero, Eq. (22) once again shows that $\hat{d}$ is a consistent estimate of the fractional differencing parameter.

To determine the variance of $\hat{d}$, we must calculate the variance of Eq. (22)'s first term. Using the results found in Section 4, we can show

$$\frac{var\bar{R}(a_m)}{R(a_m)^2} = 2|a_m|. \tag{23}$$

Combining (23) with (22) we can deduce that

$$\hat{d} - d = \theta^{1/2}|a_m|^{1/2}Z + o_p(|a_m|^{1/2}) \tag{24}$$

where $\theta = \theta(a_1, a_2, \ldots, a_m)$ is a constant and Z is a random variable with unit variance.

5 Simulations

To determine the robustness of the OLS estimate $\hat{d}$ to different signal lengths and values of d, we conduct a Monte Carlo experiment where artificial $I(d)$ processes are generated. For a number of years generating a series that exhibited long-memory has been a synthesis problem. Many of the known methods require large amounts of memory and are computationally intensive.[6] With this in mind we chose the method provided by Davies and Harte (1987) because of its computation and memory efficiencies.[7]

To insure that the results from our Monte Carlo simulations only report the statistical properties of $\hat{d}$ and not the affect zero padding has on the FWT, the generated fractionally integrated processes will have length equal to $T = 2^{max}$, where $max = 8, 9, 10, 11$. In addition, the simulated time series will be pure long-memory processes, uncontaminated from additive white noise.[8] In each experiment, 1,000 $I(d)$ processes are generated and estimated.

5.1 Results

Table 1 and fig. 2 summarizes our findings on $\hat{d}$ when the Coiflet wavelet with ten vanishing moments is used to calculate the wavelet coefficients and every scale parameter is included

[6] The method most often used to generate $I(d)$ processes are McLeod and Hipel (1978) and Hosking (1984). Both these methods require $\mathcal{O}(T^2)$ memory and take $\mathcal{O}(T^3)$ computations to generate an $I(d)$ process of length T.

[7] See either Davies and Harte (1987) or Jensen (1995) for a complete description on this methodology. Also this method is very similar in nature to the approach described in Feurerverger, Hall, and Wood (1994).

[8] Jensen (1995) contaminated the $I(d)$ processes to determine how well the wavelet filtered out the unwanted noise.

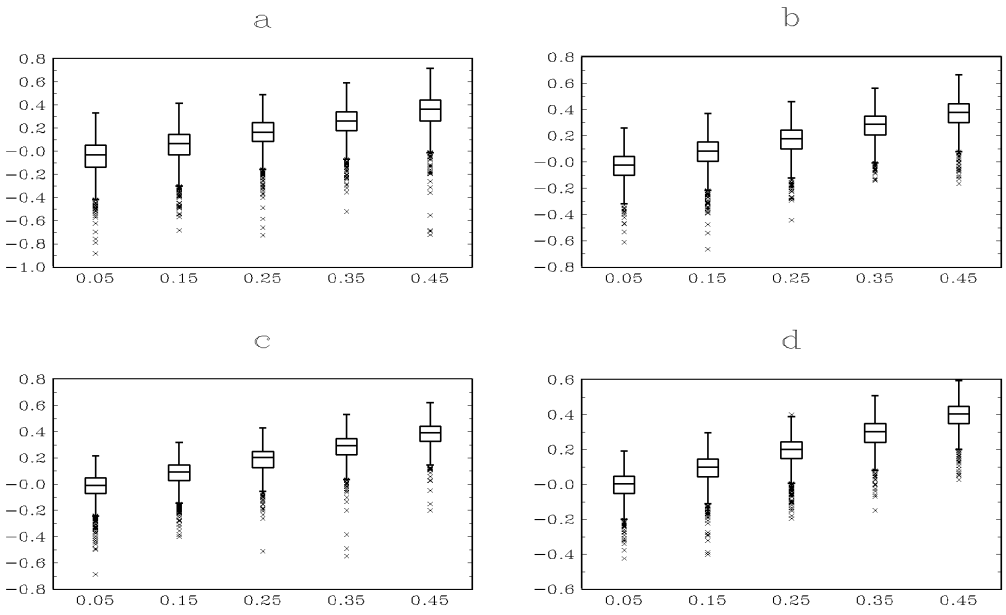

Figure 3: Box Plots: $m \in \{0, 1, \ldots, max - 1\}$, a) $T = 2^8$, b) $T = 2^9$, c) $T = 2^{10}$, d) $T = 2^{11}$

in the regression, i.e. $\{a_m\}_{m=0,1,\ldots,max-1}$. Both the box-plots of fig. 2 and the mean values of $\hat{d}$ over the samples found in table 1 show that $\hat{d}$ under estimates the true fractional differencing parameter. This downward bias in $\hat{d}$ does diminish, however only slightly, as the number of observations increases.

One possible explanation for the downward bias in $\hat{d}$ is that by using every scale available to estimate d the gain in the number of observations is offset by $\bar{R}(a_m)$ being becoming a poor estimate of $R(a_m)$. The number of wavelet coefficients at scale a_m that can be exactly calculated from a finite series gets to be fewer and fewer as a_m approaches one (m gets closer to zero), so that at the scale, $a_m = 1$, there is only one wavelet coefficient from which to calculate $\bar{R}(1)$. This presents us with a tradeoff between using all the available wavelet coefficients to estimate d, or using only those wavelet coefficients which have a scaling parameter below a specific level which will insure that $\bar{R}(a)$ provides a good estimate of $R(a)$.

Table 2 and fig. 3 contain the results from experiments where the wavelet coefficients

T	d	Mean	SD	MSE
2^8	0.05	-0.0058	0.1498	0.0338
	0.15	0.0041	0.1509	0.0346
	0.25	0.1496	0.1412	0.0300
	0.35	0.2439	0.1428	0.0316
	0.45	0.3359	0.1596	0.0384
2^9	0.05	-0.0038	0.1105	0.0200
	0.15	0.0663	0.1201	0.0216
	0.25	0.1583	0.1198	0.0227
	0.35	0.2689	0.1109	0.0189
	0.45	0.3594	0.1217	0.0230
2^{10}	0.05	-0.0024	0.1038	0.0163
	0.15	0.7512	0.0982	0.0152
	0.25	0.1786	0.1030	0.0157
	0.35	0.2730	0.1063	0.0172
	0.45	0.3778	0.0949	0.0142
2^{11}	0.05	-0.0096	0.0801	0.0100
	0.15	0.0853	0.0849	0.0114
	0.25	0.1876	0.0847	0.0111
	0.35	0.2887	0.0826	0.0106
	0.45	0.3920	0.0797	0.0097

Table 1: Summary Statistics: $m \in \{0, 1, \ldots, max - 1\}$

T	d	Mean	SD	MSE
2^8	0.05	-0.0077	0.1104	0.0155
	0.15	0.1046	0.1103	0.0142
	0.25	0.2140	0.1126	0.0140
	0.35	0.3175	0.1122	0.0136
	0.45	0.4332	0.1135	0.0131
2^9	0.05	0.0069	0.0842	0.0089
	0.15	0.1136	0.0866	0.0088
	0.25	0.2157	0.0942	0.0101
	0.35	0.3238	0.0910	0.0090
	0.45	0.4323	0.0853	0.0076
2^{10}	0.05	0.0152	0.0696	0.0061
	0.15	0.1172	0.0692	0.0059
	0.25	0.2268	0.0698	0.0054
	0.35	0.3280	0.0729	0.0058
	0.45	0.4366	0.0711	0.0052
2^{11}	0.05	0.0172	0.0584	0.0045
	0.15	0.1204	0.0588	0.0043
	0.25	0.2256	0.0591	0.0043
	0.35	0.3340	0.0603	0.0039
	0.45	0.4381	0.0617	0.0039

Table 2: Summary Statistics: $m \in \{1, 2, \ldots, max - 1\}$

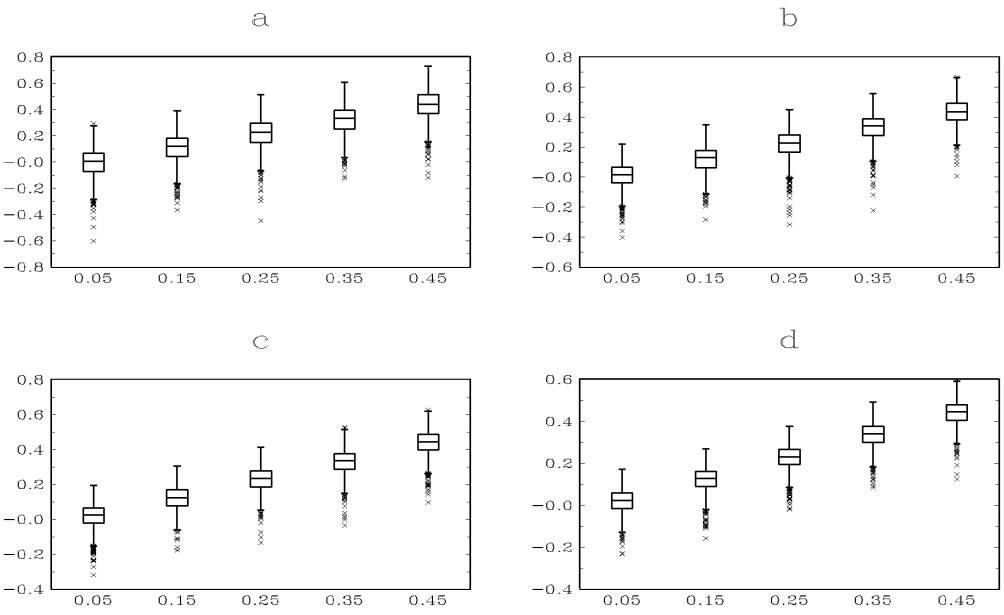

Figure 4: Box Plots: $m \in \{1, 2, \ldots, max - 1\}$, a) $T = 2^8$, b) $T = 2^9$, c) $T = 2^{10}$, d) $T = 2^{11}$

with scaling parameters a_m, where $m = 1, 2, \ldots, max - 1$, were used to estimate d. Comparing these results to those found in table 1 and fig. 2, we find the mean squared error decreases when the coarsest wavelet coefficient is left out of the regression, regardless of the value of d or T. In addition, the biasness found in table 1 and fig. 2 becomes insignificant and each of the 25th-75th quantile boxes in fig. 3 contains the true value of d.

Even though the results found in table 2 and fig. 3 are an improvement over those found in table 1 and fig. 2, $\bar{R}(a_1)$ is a sum of only two wavelet coefficients, $\{< x, \psi_{1/2,2n} >\}_{n=0,1}$, Whereas $\bar{R}(a_2)$ has the four wavelet coefficients $\{< x, \psi_{1/4,4n} >\}_{n=0,1,2,3}$. Hence, to improve the regression at the expense of decreasing the total number of observations, we choose to use the wavelet coefficients with scaling parameters, a_m, where $m = 2, 3, \ldots, max - 1$ to estimate d. The simulation results using these scaling parameters are respectively listed and plotted in table 3 and fig. 4.

The improvement from table 2 and fig. 3 to table 3 and fig. 4 are not as dramatic as those found when we left out the coarsest resolution of the wavelet coefficients. The main effect of dropping the next coarsest resolution from the regression was to improve the estimate of d when it is close to zero, especially when $d = 0.05$. In addition, our simulation results

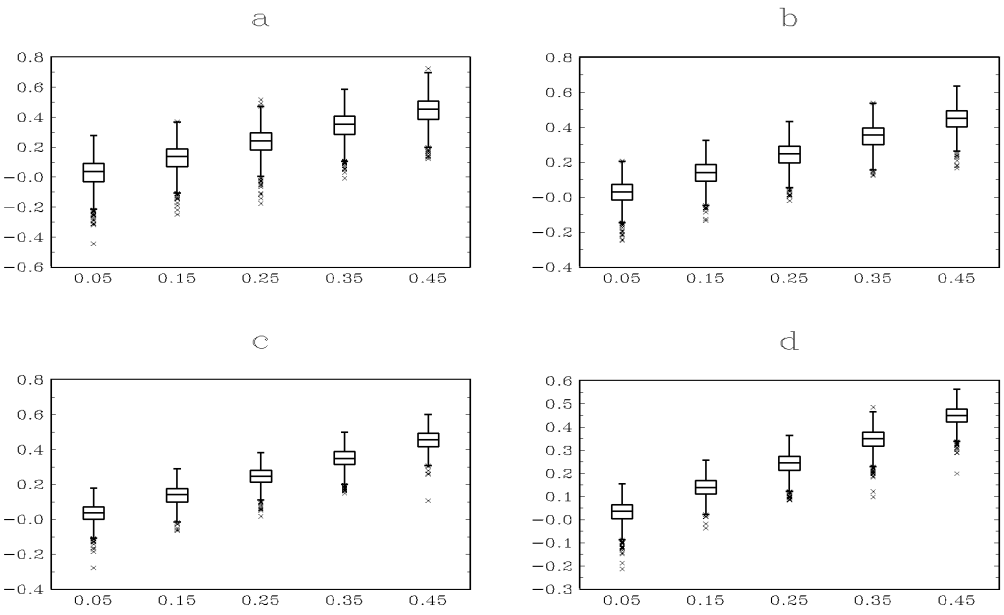

Figure 5: Box Plots: $m \in \{2, 3, \ldots, max - 1\}$, a) $T = 2^8$, b) $T = 2^9$, c) $T = 2^{10}$, d) $T = 2^{11}$

suggest that a sample size of $T \geq 2^8$ is large enough for the asymptotic properties of $\hat{d}$ to take hold. Notice that increasing the sample size above 2^8 only marginally improves the statistical inference of $\hat{d}$ as measured by the MSE and mean of the Monte Carlo experiment.

6 Conclusion

In this paper we have shown that the log-log relationship exists between the variance of the wavelet coefficient and the fractional differencing parameter of a fractionally integrated, and that this relatioship provides a simply method to estimating the differencing parameter. The OLS estimate of the fractional differencing parameter is shown to be consistent when the sample variance of the wavelet coefficient is used in the regression.

The Monte Carlo simulation bore this out, but only after the two sample variances at the coarsest levels of smoothing were left out of the regression. When all the available wavelet coefficient's sample variance were included in the regression, the OLS estimate of the fractional differencing parameter found in our simulations tended to be slightly biased in the downward direction. When the two coarsest sample variances were left out of the

T	d	Mean	SD	MSE
2^8	0.05	0.0238	0.0961	0.0099
	0.15	0.1266	0.0905	0.0087
	0.25	0.2351	0.0931	0.0089
	0.35	0.3413	0.0933	0.0088
	0.45	0.4472	0.0943	0.0089
2^9	0.05	0.0248	0.0686	0.0053
	0.15	0.1353	0.0686	0.0049
	0.25	0.2412	0.0703	0.0050
	0.35	0.3472	0.0681	0.0048
	0.45	0.4479	0.0717	0.0051
2^{10}	0.05	0.0328	0.0055	0.0033
	0.15	0.1363	0.0570	0.0034
	0.25	0.2424	0.0547	0.0031
	0.35	0.3457	0.0558	0.0031
	0.45	0.4521	0.0539	0.0029
2^{11}	0.05	0.0317	0.0463	0.0025
	0.15	0.1383	0.0432	0.0020
	0.25	0.2402	0.0451	0.0021
	0.35	0.3452	0.0473	0.0023
	0.45	0.4473	0.0427	0.0022

Table 3: Summary Statistics: $m \in \{2, 3, \ldots, max - 1\}$

regression, the downward bianess was no longer present at of the sample sizes used in the Monte Carlo experiments. In addition we found that the OLS estimate reacheds its asymptotic properties with a time series of atleast 2^8 observations.

A Proof of Theorem 1

Let $x(t)$ be a mean zero $I(d)$ process with $|d| < 1/2$ and $\sigma_\epsilon^2 = 1$. The expected value of $< x, \psi_{a,b} >$ can easily be shown to equal zero, since

$$E[< x, \psi_{a,b} >] = |a|^{-1/2} \int E[x(t)] \psi \left(\frac{t-b}{a} \right) dt.$$

The variance of the wavelet coefficients, $R(a)$, equals

$$\begin{aligned} R(a) &= E\left[< x, \psi_{a,b} >^2\right] \\ &= |a|^{-1} \int dt \int ds\, E[x(t)x(s)]\, \psi \left(\frac{t-b}{a} \right) \psi \left(\frac{s-b}{a} \right). \end{aligned}$$

Using the fractionally integrated processes' autocovariance function found in Eq. (13) and by a change of variables

$$\begin{aligned} R(a) &= -K|a|^{2d} \int dt \int ds\, |t-s|^{2d-1} \psi(t)\, \psi(s) \\ &= K'|a|^{2d} \int dt\, |t|^{2d-1} \Lambda(1,t) \end{aligned}$$

where $\Lambda(1,t) = \int \psi(s)\psi(s-t)\, ds$ is the wavelet transform of the 'mother' wavelet. Collecting terms we find

$$R(a) = \sigma^2 |a|^{2d}$$

where $\sigma^2 = K' \int dt\, |t|^{2d-1} \Lambda(1,t) < \infty$ since $\Lambda(1,t)$ is finite. Thus, $< x, \psi_{a,b} > \sim \mathcal{N}(0, \sigma^2 |a|^{2d}|)$. Q.E.D.

B Proof of Theorem 2

Let $\psi(t)$ have $M \geq 1$ vanishing moments and $\alpha = a^{-1}(b_1 - b_2)$. Using the steps found in the proof of Theorem 1, the $corr(< x, \psi_{a,b_1} >, < x, \psi_{a,b_2} >)$ can be written as

$$\begin{aligned} corr(< x, \psi_{a,b_1} >, < x, \psi_{a,b_2} >) &= \frac{\int dt \int ds\, |t-s+\alpha|^{2d-1} \psi(t)\psi(s)}{\int dt \int ds\, |t-s|^{2d-1} \psi(t)\psi(s)} \\ &= K' \int dt \int ds\, |t+\alpha|^{2d-1} \Lambda(1,t) \end{aligned} \tag{25}$$

where $\Lambda(1,t) = \int ds\, \psi(s-t)\psi(s)$ and K' is a finite constant. Let $|\alpha| > K_1 + K_2$, i.e. $\alpha \notin supp(\Lambda(1,t))$, so that $|t+\alpha|^{2d-1}$ is continuously differentiable. By the binomial theorem

$$\begin{aligned} |t+\alpha|^{2d-1} &= |\alpha|^{2d-1}|1+\frac{t}{\alpha}|^{2d-1} \\ &= |\alpha|^{2d-1}\left\{1+\sum_{i=1}^{\infty}\left(\begin{array}{c} 2d-1 \\ i \end{array}\right)\left(\frac{t}{\alpha}\right)^i\right\} \end{aligned} \tag{26}$$

Substituting Eq. (26) into Eq. (25), the correlation becomes

$$\begin{aligned} corr(<x,\psi_{a,b_1}>,<x,\psi_{a,b_2}>) &= K'|\alpha|^{2d-1}\left\{\int dt\,\Lambda(1,t)\right. \\ &\quad \left. + \int dt \sum_{i=1}^{\infty}\left(\begin{array}{c} 2d-1 \\ i \end{array}\right)\left(\frac{t}{\alpha}\right)^i \Lambda(1,t)\right\} \end{aligned} \tag{27}$$

Since $\psi(t)$ has M vanishing moments, $\Lambda(1,t)$ first $2M$ moments are zero (See Jensen (1995) for the proof of this result). Hence,

$$corr(<x,\psi_{a,b_1}>,<x,\psi_{a,b_2}>) = C_1|a|^{-2(d-M)+1}|b_1-b_2|^{2(d-M)-1} + R_{2M+1} \tag{28}$$

where

$$C_1 = K'\frac{(2d-1)!}{2M!(2(d-M)-1)!}\left(\int dt\, t^M \psi(t)\right)^2$$

and

$$R_{2M+1} = K'|\alpha|^{2d-1}\left\{\sum_{i=2M+1}^{\infty}\left(\begin{array}{c} 2d-1 \\ i \end{array}\right)\left(\frac{s-t}{\alpha}\right)^i \psi(t)\,\psi(s)\,dt\,ds\right\} \tag{29}$$

Since $M \geq 1$ and $|d| < 1/2$

$$|R_{2M+1}| \leq C_2|\alpha|^{2d-1}\sum_{i=1}^{\infty}\sup_{(t,s)\in\Omega}\left|\frac{s-t}{\alpha}\right|^{2M+i}$$

where

$$C_2 = K'\left|\left(\begin{array}{c} 2d-1 \\ i \end{array}\right)\right|\left(\int |\psi(t)|dt\right)^2$$

and the set $\Omega = \{(t,s) : -K_1 \leq t,s \leq K_2\}$. Since

$$\sup_{(t,s)\in\Omega}\left|\frac{s-t}{\alpha}\right| < 1.$$

it then follows that

$$|R_{2M+1}| \leq C_3|a|^{-2(d-M)}|b_1-b_2|^{2(d-M)} \tag{30}$$

where C_3 is a finite constant. It follows from Eq. (28) and Eq. (30) that

$$corr(< x, \psi_{a,b_1} >, < x, \psi_{a,b_2} >) \quad = \quad \mathcal{O}\left(|b_1 - b_2|^{2(d-M)-1}\right)$$

and

$$corr(< x, \psi_{a,b_1} >, < x, \psi_{a,b_2} >) \quad = \quad \mathcal{O}\left(\left|\frac{1}{a}\right|^{2(d-M)-1}\right) \tag{31}$$

for all b_1 and b_2 such that $|b_1 - b_2| > K_1 + K_2$. Q.E.D.

References

[1] Beylkin, G., R. Coifman, and V. Rokhlin, (1991) "Fast Wavelet Transfroms and Numerical Algorithms I," *Communications on Pure and Applied Mathematics,* 44, 141-183.

[2] Daubechies, I. (1988) "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, 41, 909-996.

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

[4] Davies, R. and D. Harte (1987) "Tests for Hurst Effect," *Biometrika,* 74, 95-101.

[5] Donoho, D. L., and I.M. Johnstone, (1994) "Ideal Spatial Adaptation Via Wavelet Shrinkage," *Biometrika,* 81, 425-455.

[6] Donoho, D. L., and I.M. Johnstone, (1995a) "Minimax Estimation Via Wavelet Shrinkage," *Annals of Statistics,*, to be published.

[7] Donoho, D. L., and I.M. Johnstone, (1995b) "Adapting to Unknown Smoothness Via Wavelet Shrinkage," *Journal of the American Statistical Association,* to be published.

[8] Donoho, D. L., I.M. Johnstone, G. Kerkyacharian and D. Picard (1995) "Wavelet Shrinkage: Asymptopia?" *Journal of the Royal Statistical Society, B*, 57, 301-337.

[9] Flandrin, P. (1991) "Fractional Brownian Motion and Wavelets," in *Wavelets, Fractals, and Fourier Transformations: New Developments and New Applications,* eds. M. Farge, J.C.R. Hunt, and J.C. Vassilicos, (Oxford Press: Oxford, UK).

[10] Granger, C. and R. Joyeux (1980) "An Introduction to Long-Memory Time Series Models and Fractional Differencing," *Journal of Time Series Analysis*, 1, 15-29.

[11] Grossman, A. and J. Morlet (1984) "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape," *SIAM Journal of Mathematical Analysis,* 15, 723-736.

[12] Hosking, J. R. (1981) "Fractional Differencing," *Biometrika*, 68, 165-176.

[13] Hosking, J. (1984) "Modeling Persistence in Hydrological Time Series Using Fractional Differencing," *Water Resources Research*, 20, 1898-1908.

[14] Jensen M. (1995) "Analysis and Estimation of Fractionally Integrated Processes with Compactly Supported Wavelets", *Journal of Econometrics* to be published.

[15] Kaplan, L. and C. Kuo, (1993) "Fractal Estimation from Noisy Data via Discrete Fractional Gaussian Noise and the Haar Basis," *IEEE Transactions on Signal Processing,* 41, 3554-3562.

[16] Kronland-Martinet, R., J. Morlet, and A. Grossman (1987) "Analysis of Sound Patterns Through Wavelet Transforms," *International Journal of Pattern Recognition and Artificial Intelligence*, 1, 273-301.

[17] Mallat, S. (1989) "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674-693.

[18] Mandelbrot, B. and J. Van Ness (1968) "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Review*, 10, 422-437.

[19] McLeod, B. and K. Hipel (1978) "Preservation of the Rescaled Adjusted Range, I. A Reassessment of the Hurst Phenomenon," *Water Resources Research*, 14, 491-518.

[20] Meyer, Y. (1990) *Ondelettes ét Operérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III: Opérateurs Multilinéaires*, Paris: Hermann.

[21] Morlet, J. (1983) "Sampling Theory and Wave Propagation," in *Acoustic Signal/Image Processing and Recognition*, ed. C. H. Chen (Springer-Verlag: Berlin), 233-261.

[22] Tewfik, A.H., and M. Kim (1992) "Correlation Structure of the Discrete Wavelet Coefficients of Fractional Brownian Motion," *IEEE Transaction on Information Theory,* 38, 904-909.