Statistica Sinica **16**(2006), 1255-1271

LONG-MEMORY WAVELET MODELS

Nan-Jung Hsu

National Tsing-Hua University

Abstract: This article presents a novel long-memory wavelet model for approximating a stationary long-memory process. The proposed model is constructed in the wavelet domain in which the dependence structure is characterized by the variances of wavelet coefficients at different scales. This model can be easily incorporated into more complex model structures such as a generalized linear model. For inference, maximum likelihood estimation is derived. In a simulation study, we show that the modeling via wavelets has a good performance both in estimating the long-memory parameter and in predicting future observations under various long-memory processes. For illustration, the methodology is applied to modeling the Nile River data.

Key words and phrases: Discrete wavelet transform, long-range dependence, spectral density.

1. Introduction

Long-memory phenomena have been observed frequently in finance, economics, hydrology, geophysics and many other fields. Generally speaking, a process is called long memory if its spectral density is unbounded at the origin. Consequently, the autocorrelation function decays at a hyperbolic rate such that the absolute values of the autocorrelations are not summable. Conventionally, the class of fractionally integrated autoregressive moving average (ARFIMA) processes (Hosking (1981) and Granger and Joyeux (1980)) is used most widely in modeling time series with both long-memory and short-memory behaviors.

The corresponding maximum likelihood (ML) estimation (Sowell (1992) is computationally intensive due to high-dimensional matrix inversion when dealing with large data sets. In contrast to the ARFIMA models constructed in the time domain, one alternative is to specify the spectral density semi-parametrically for small frequencies in the frequency domain. Accordingly, the corresponding inference for the long-memory parameter is derived in the frequency domain, for example the GPH estimator (Geweke and Porter-Hudak (1983)), the Whittle estimator of Fox and Taqqu (1986) and the semi-parametric estimator of Robinson (1995). Since the semiparametric model only characterizes the long-memory property of the underlying process, not the complete dependence structure of the underlying process, the implications on forecasting and predictions are limited.

It is well known that wavelet methods provide excellent tools for scale analysis of time series. Wavelets are particularly powerful for analyzing long-memory properties due to self-similarity (Allan (1966)), Flandrin (1992), Masry (1993) and Wornell (1993)). In the literature about long memory, wavelet analysis has been used in simulating long-memory realizations and in estimating the long-memory parameter under a fractionally integrated (FI) process (McCoy and Walden (1996) and Jensen (1999a)) and an ARFIMA process (Jensen (1999b, 2000)). These studies proposed alternative likelihood-based estimation procedures for the conventional ARFIMA processes which essentially approximate the likelihood function in terms of the discrete wavelet transform (DWT) of data in which the dependence between wavelet coefficients is ignored. The advantage of using the wavelet approach is to reduce the computational order of calculating the likelihood, which is particularly useful for large data sets. Moreover, the wavelet MLE of the long-memory parameter has been found to be fairly robust to model specification in previous empirical studies. Therefore, when the main interest is only in estimating the long-memory parameter, the wavelet MLE can be viewed as a semiparametric estimator for the long-memory parameter.

Motivated by the variance structure of DWT under a FI process and the success of wavelet MLE for estimating the long-memory parameter, we consider a new class of time series models in the wavelet domain in which the dependence structure is fully determined by the variances and covariances of wavelet coefficients at different scales. By imposing certain constraints on the variances (which are considered as parameters in the new model) of wavelet coefficients at large scales, the new model can exhibit long memory. The constraints only rely on the behavior of the spectral density of a long-memory process near zero frequency. In other words, the new model is fairly semiparametric in specifying the short-memory behavior which corresponds to the behavior of the spectral density away from zero frequency.

In contrast to previous studies focusing on an alternative likelihood-based estimation for ARFIMA processes using wavelet analysis, the primary interest here is to provide an alternative class of models which exhibit long memory. From the modeling point of view, the long-memory structure of the proposed model is similar but not identical to that of a FI model, because the way of constructing long memory in the proposed model is to match the variance structure of wavelet coefficients to that obtained from a FI process only in large scales. Moreover, the short-memory structure of the proposed model, which is fully determined by the variances of wavelet coefficients in small scales and the covariance of scaling function coefficients (parameters in the model), is different from those of ARFIMA models. From the estimation point of view, since the short-memory structure in the proposed model is fairly semiparametric, the estimator of long-memory parameter is expected to be robust. The wavelet MLE suggested by previous studies also has this advantage so that it can be viewed as a semiparametric estimator when data are contaminated. In this case, our model still provides sensible predictions that are not available under the conventional approach. For inference, maximum likelihood estimation is derived for the proposed long-memory wavelet model.

The rest of the paper is organized as follows. Section 2 introduces the proposed long-memory wavelet model. In Section 3, the ML estimation is described for parameter estimation. In Section 4, a simulation study is conducted for investigating the performance of our modeling procedure under various long-memory processes, and for comparison with those using conventional methods. In Section 5, for illustration, the proposed method is applied to the Nile River data. A brief discussion is finally drawn in Section 6.

2. Long-Memory Models in Wavelet Domain

In Section 2.1, we introduce the wavelet transform, and look at its properties in connection with long-memory processes. The proposed long-memory wavelet model is described in Section 2.2.

2.1. Wavelet transform of a long memory process

Suppose $\{x_t\}$ is a long-memory process with long-memory parameter d, so its spectral density satisfies

$$f(\omega) \sim c_f\, \omega^{-2d}, \quad \text{for } \omega \text{ near zero}, \tag{1}$$

where $0 < d < 0.5$ and $c_f > 0$. In the literature, the most popular class of long-memory models consists of ARFIMA processes that satisfy the following difference equation:

$$\phi(B)(1-B)^d x_t = \theta(B)\epsilon_t,$$

where $0 < d < 0.5$, the polynomials $\phi(z)$ and $\theta(z)$ have no common roots, and all roots are outside the unit circle, $\{\epsilon_t\}$ are independent, identically distributed (iid) $N(0,\sigma^2)$.

For notational simplicity, let $T = 2^p$ for some $p \in I\!N$ in the following context. The discrete wavelet transform (DWT) of $\boldsymbol{x} = (x_1, \ldots, x_T)'$ satisfies

$$\boldsymbol{w} = \boldsymbol{W}\boldsymbol{x}, \tag{2}$$

where $\boldsymbol{W}$ is the matrix of the discrete wavelet transform associated with the selected wavelet filter (e.g., Haar wavelet) and the selected resolution (indicated by J in the later text). More details about the DWT for time series can be found in Percival and Walden (2000), Vidakovic (1999), and the references therein. The transformed vector $\boldsymbol{w}$ can be decomposed as

$$\boldsymbol{w} \equiv \left(\boldsymbol{d}'_{-1}, \ldots, \boldsymbol{d}'_{-J}, \boldsymbol{c}_{-J}\right)', \tag{3}$$

in which $J \leq p$, $\boldsymbol{d}_j \equiv \{d_{j,k} : k = 1, \ldots, T2^j\}$ comprises the wavelet coefficients at the jth scale for $j = -1, \ldots, -J$, and $\boldsymbol{c}_{-J} \equiv \{c_{-J,k} : k = 1, \ldots, T2^{-J}\}$ comprises the scaling-function coefficients. According to Daubechies (1992), the jth scale ($j = -1, \ldots, -J$) wavelet filter acts as an approximate band-pass filter with octave pass-band $[-2^{j+1}\pi, -2^j\pi) \cup (2^j\pi, 2^{j+1}\pi]$, and the corresponding wavelet coefficients are approximately a bandpass representation of the original data. Moreover, the wavelet filter corresponding to the scaling-function coefficients $\boldsymbol{c}_{-J}$ acts as an approximate band-pass filter with octave pass-band $[-2^{-J}\pi, 2^{-J}\pi]$. In particular, when $J = p$, $\boldsymbol{c}_{-J}$ only contains the single scaling-function coefficient that corresponds to the sample mean of the observed series.

Given a zero-mean long-memory process with the power spectrum in (1), the wavelet coefficients and the scaling-function coefficients also have zero mean, i.e., $E\boldsymbol{d}_j = \boldsymbol{0}$ for $j = -1, \ldots, -J$ and $E\boldsymbol{c}_{-J} = \boldsymbol{0}$. Moreover, the variances of the wavelet coefficients satisfy

$$\sigma_c^2 \equiv \mathrm{Var}\,(c_{-J,k}) \approx \frac{1}{2^{-J}\pi}\int_0^{2^{-J}\pi} f(\omega)d\omega \approx \frac{2^J}{\pi}\int_0^{2^{-J}\pi} c_f\omega^{-2d}d\omega = \frac{c_f}{1-2d}\pi^{-2d}2^{2dJ},$$

$$\sigma_j^2 \equiv \mathrm{Var}\,(d_{j,k}) \approx \frac{1}{2^{j+1}\pi - 2^j\pi}\int_{2^j\pi}^{2^{j+1}\pi} f(\omega)\,d\omega \approx \frac{c_f}{1-2d}\pi^{-2d}2^{-2dj}\left(2^{1-2d}-1\right),$$

for relatively large scale (i.e., $j = -p, -p+1, ..$). Similar results are derived by McCoy and Walden (1996) and Jensen (1999a) for a FI(d) process, and by Jensen (1999b) for an ARFIMA process.

It is well known that the DWT is an excellent decorrelator and therefore, the wavelet coefficients are nearly uncorrelated both within and between scales. In fact, this is the key property for having computational efficiency using wavelet MLE in which the wavelet coefficients are considered independent. For justifying this independent assumption, Dijkerman and Mazumdar (1994) showed that the correlations of wavelet coefficients decay exponentially fast between scales and hyperbolically fast within scales for a fractionally Brownian motion. Similarly, for a fractionally integrated process, Fan (2003) showed that the within-scale correlations decay hyperbolically and the decay rate increases as the length of wavelet filter increases. Moreover, the between-scale correlations decrease to zero uniformly across time lags as the length of wavelet filter increases.

Based on these theoretical properties about the DWT of a long-memory process, we propose a new model in the wavelet domain in which the dependence structure is characterized by the variances of wavelet coefficients at different scales, denoted by $\{\sigma_j^2 : j = -1, \ldots, -J\}$, and the covariance matrix of scaling function coefficients, denoted by $\sigma_c^2 \Gamma_c$ where Γ_c is a correlation matrix. By imposing constraints on σ_j^2 at large scales satisfying $\sigma_j^2 \propto 2^{-2dj}$, which is equivalent to (1), the proposed model exhibits long memory behavior with parameter d. On the other hand, no structure is assumed for σ_j^2 at the small scales which characterize the spectral density away from frequency zero and correspond to short-range dependence. Similar to previous related studies, the correlations in wavelet coefficients within and between scales are assumed to be zero for ease of inference. The scaling function coefficients in $\boldsymbol{c}_{-j}$ preserve information of the spectral density on $[-2^{-J}\pi, 2^{-J}\pi]$ which represents low-frequency variations. Compared to wavelet coefficients, the scaling function coefficients are relatively smooth and the correlations between them cannot be ignored. The correlations in scaling function coefficients is taken into account in our model by assuming an AR(1) dependence structure. This choice is somehow arbitrary and more complex correlation structure can be assumed such as a higher-order AR or even a FI(d). But, in practice, we found that AR(1) works very well to approximate the dependence in scaling function coefficients for the various situations considered in our simulations. The precise model specification is given in the next section.

2.2. A long-memory wavelet model

In this section, a linear model constructed in the wavelet domain is introduced. First, select a $T \times T$ wavelet transform matrix $\boldsymbol{W}$ corresponding to some wavelet filter and a resolution J with $1 \leq J \leq [\log_2 T]$. Let $\boldsymbol{w} = (\boldsymbol{d}'_{-1}, \ldots, \boldsymbol{d}'_{-J}, \boldsymbol{c}'_{-J})$ be a $T \times 1$ random vector with $\boldsymbol{d}_j = (d_{j,1}, \ldots, d_{j,n_j})'$ for $j = -1, \ldots, -J$, where $n_j = T2^j$. The elements $\{d_{j,k}\}$ are assumed to be iid random variables from $N(0, \sigma_j^2)$ for each scale j, and independent between scales; $\boldsymbol{c}_{-J}$ is assumed to be $N(\boldsymbol{0}, \sigma_c^2 \boldsymbol{\Gamma}_c)$ and independent of $\{\boldsymbol{d}_j\}$, where $\{\sigma_j^2\}$ and σ_c^2 satisfy

$$\sigma_j^2 = \begin{cases} \tau_j^2, & j = -1, \ldots, -K, \\ \tau_0^2\,(2^{1-2d} - 1)2^{-2dj}, & j = -(K+1), \ldots, -J, \\ \tau_0^2\, 2^{2dJ}, & j = c, \end{cases} \tag{4}$$

for a given K $(0 \leq K \leq J-1)$, and $\mathbf{\Gamma}_c$ is an $(n_c \times n_c)$ correlation matrix $(n_c = T2^{-J})$ satisfying

$$\mathbf{\Gamma}_c = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \\ \phi & 1 & \phi & \ddots & \\ \phi^2 & \ddots & \ddots & \ddots & \phi^2 \\ \vdots & \ddots & \ddots & \ddots & \phi \\ & & \phi^2 & \phi & 1 \end{pmatrix}, \tag{5}$$

thus an AR(1) dependence structure.

Suppose $\boldsymbol{y} = (y_1, \ldots, y_T)'$ are the observed data. A long-memory wavelet model for $\boldsymbol{y}$ is defined as

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{W}'\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)'$ is the design matrix in which the ith row corresponds to the covariate variables for the ith observation, $\boldsymbol{\beta}$ is the corresponding regression coefficient vector, $\boldsymbol{W}'$ is the inverse of wavelet transform matrix, $\boldsymbol{w}$ satisfies the variance structures in (4) and (5), and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_T)'$ are measurement errors with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, independent of $\boldsymbol{w}$. For $\boldsymbol{y}$, the mean structure is characterized by $\boldsymbol{Z\beta}$ and the dependence structure is characterized by the variance of $\boldsymbol{w}$. In particular, the parameters $\tau_0^2 > 0$ and $d \in [0, 0.5)$ model the long-memory behavior and $\{\tau_j^2 > 0 : j = -1, \ldots, -K\}$ and $|\phi| < 1$ model the short-memory behavior. The measurement error ϵ_t is considered in our model to account for the nugget effect which is often observed in financial and environmental time series with long memory.

In the proposed model, $\mathrm{Var}\,(\boldsymbol{y}) = \boldsymbol{W}'\Lambda_w\boldsymbol{W} + \sigma_\epsilon^2\boldsymbol{I}$, where

$$\boldsymbol{\Lambda}_w \equiv \mathrm{Var}\,(\boldsymbol{w}) = \begin{pmatrix} \sigma_{-1}^2 I & & & & \\ & \sigma_{-1}^2 I & & & \\ & & \ddots & & \\ & & & \sigma_{-J}^2 I & \\ & & & & \sigma_c^2 \mathbf{\Gamma}_c \end{pmatrix}. \tag{7}$$

Note that $\mathrm{Var}\,(\boldsymbol{y})$ is not a Toeplitz matrix which means the covariance function is not stationary across time. The non-stationarity is caused by assuming independence of wavelet coefficients between and within scales in the construction. Although the covariance functions are varying from row to row in $\mathrm{Var}\,(\boldsymbol{y})$, their patterns are quite similar and their differences are small. From the modeling point of view, the proposed model provides a good approximation to a stationary long memory process in the sense that the covariance structure of the proposed model in the wavelet domain provides a good approximation to that of a stationary long-memory process.

In determining the wavelet transform matrix $\boldsymbol{W}$, the resolution J should be selected in advance. Generally speaking, for modeling long-memory, J should be selected as large as possible to capture information about the spectrum near zero frequency. One also needs to select K in the proposed model, it plays the same role as choosing Fourier frequencies in the regression for calculating the GPH estimator (Geweke and Porter-Hudak (1983)). Consider two extreme cases: when $K = 0$, d is involved in the specification of all frequencies; when $K = J-1$, d is only involved in the specification of the largest scale variations that correspond to information about the spectrum near zero frequency. In practice, J and K determine the number of parameters in our model and both can be selected via a model selection criterion such as AIC.

3. Estimation Method

One advantage of our model is that likelihood-based inference is much simpler and computationally efficient than under conventional ARFIMA models. In this section, maximum likelihood estimation is derived for the long-memory wavelet model defined in Section 2.2.

Let $\boldsymbol{\theta} \equiv (\sigma_\epsilon^2, d, \tau_0^2, \phi, \{\tau_j^2 : j = -1, \ldots, -K\})'$ be the parameter vector determining Var $(\boldsymbol{y})$. Under the long-memory wavelet model in (6), $\boldsymbol{y} \sim N(\boldsymbol{Z\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{W}'\boldsymbol{\Lambda}_w\boldsymbol{W} + \sigma_\epsilon^2\boldsymbol{I}$. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, respectively. Then

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) = \left(\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{y}, \tag{8}$$

where $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\left(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right). \tag{9}$$

Since $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{W}'(\boldsymbol{\Lambda}_w + \sigma_\epsilon^2\boldsymbol{I})\boldsymbol{W} \equiv \boldsymbol{W}'\boldsymbol{\Sigma}_*(\boldsymbol{\theta})\boldsymbol{W}$, (8) and (9) can be written as

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}) = \left(\boldsymbol{Z}_*'\boldsymbol{\Sigma}_*^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{Z}_*\right)^{-1} \boldsymbol{Z}_*'\boldsymbol{\Sigma}_*^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{y}_*, \tag{10}$$

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_*(\boldsymbol{\theta})| - \frac{1}{2}\left(\boldsymbol{y}_* - \boldsymbol{Z}_*\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right)'\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\theta})\left(\boldsymbol{y}_* - \boldsymbol{Z}_*\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\right), \tag{11}$$

where $\boldsymbol{y}_* = \boldsymbol{W}\boldsymbol{y}$ is the DWT of $\boldsymbol{y}$ and $\boldsymbol{Z}_* = \boldsymbol{W}\boldsymbol{Z}$ consists of the DWT for each column of $\boldsymbol{Z}$. Since $\boldsymbol{\Sigma}_*$ is a block-diagonal matrix consisting of a $(T-n_c)\times(T-n_c)$ diagonal matrix and an $n_c \times n_c$ non-diagonal matrix, this estimation procedure only involves n_c-dimensional matrix operations and therefore is computationally feasible for very large data sets.

Moreover, by Searle (1970), the asymptotic variance of the MLE for $(\boldsymbol{\theta}, \boldsymbol{\beta})$ under (6) is the inverse of the information matrix satisfying

$$I = \begin{pmatrix} I_{\theta\theta} & \mathbf{0} \\ \mathbf{0} & I_{\beta\beta} \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\boldsymbol{B}_* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_*'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{Z}_* \end{pmatrix},$$

where $\boldsymbol{B}_*$ is the matrix with ijth element tr $\left[\boldsymbol{\Sigma}_*^{-1}(\partial\boldsymbol{\Sigma}_*/\partial\theta_i)\boldsymbol{\Sigma}_*^{-1}(\partial\boldsymbol{\Sigma}_*/\partial\theta_j)\right]$.

4. Numerical Simulation

The performance of the proposed model for approximating long-memory phenomenon is investigated for several long-memory processes in this section. First, we examine the performance for estimating the long-memory parameter in Section 4.1. Then, the performance for forecasting future observations is examined in Section 4.2.

4.1. Estimating the long-memory parameter

The performance for estimating the long-memory parameter d under various long-memory processes is studied. Our intention is to assess the stability of our modeling scheme across a range of data generating mechanisms. Several long-memory processes are considered for data generation, including FI(d), ARFIMA(1,d,1) with $d = 0.4$ and both processes with additive noise. For each process, 100 realizations with sample size $T = 1,024$ (i.e., $p = 10$) are generated. The data are generated in Splus using the function "arima.fracdiff.sim" in which each series with length 6,024 is generated first, and only the last 1,024 observations are used to form a realization. For each realization, four parametric models, FI(d), ARFIMA(1,d,1) and our long-memory wavelet models with $K = 2$ and $K = 4$ and $\boldsymbol{Z} = \mathbf{1}$, are fitted and the corresponding maximum likelihood estimates of d is calculated. In our simulation, the symmlet wavelet (s8) is used in the DWT and the resolution J is set to be seven. For comparison, the GPH estimates that consider the first $[T^\alpha]/2$ periodogram ordinates in the regression with $\alpha = 0.8$ (GPH1) and $\alpha = 0.6$ (GPH2) are also computed. Note that, for $T = 1,024$, the GPH estimator with $\alpha = 0.8$ only uses one fourth of all periodogram ordinates in the regression. Similarly, the ML estimator of d under our long-memory wavelet model with $K = 2$ (LW1) is solved only using the information about the spectrum over $[0, \pi/4]$. Essentially, these two estimators use the same amount of information in estimating d except the former is a semiparametric estimator and the latter is a parametric estimator. It is similar for the GPH estimator with $\alpha = 0.6$ and the ML estimator in a long-memory wavelet model with $K = 4$ (LW2), in which only the information about the spectrum over $[0, \pi/16]$ is used for estimating d.

The first process considered is FI(0.4). The estimation results are presented in the first part of Table 1 and Figure 1. As expected, the MLE under the true model (i.e., FI(d)) has the best performance in terms of smallest bias and root mean squared error. The second best estimator is the MLE based on our long-memory wavelet model with $K = 2$, which is superior to GPH1. Even though the model is misspecified, the long-memory wavelet model does a good job in recovering d. The performance of GPH2 and the MLE under long-memory wavelet model with $K = 4$ is obviously worse than the other estimators since less information is used in estimating d.

Table 1. Biases, standard errors (s.d) and root mean squared errors (rmse) of GPH estimates with $\alpha = 0.8$ (GPH1) and $\alpha = 0.6$ (GPH2), ML estimates under FI(d), ARFIMA(1,d,1) and the long-memory wavelet models with $K = 2$ (LW1) and $K = 4$ (LW2) under various true processes. ($T = 1,024$, 100 replications).

True Model	Estimation Error	GPH1 $\alpha = 0.8$	GPH2 $\alpha = 0.6$	FI	ARFIMA	LW1 $K = 2$	LW2 $K = 4$
FI(0.4)	bias	-0.011	-0.027	-0.005	-0.076	-0.018	-0.083
	s.d.	0.064	0.159	0.028	0.081	0.042	0.083
	rmse	0.065	0.161	0.028	0.111	0.045	0.117
FI(0.4)+WN	bias	-0.086	-0.054	-0.139	-0.114	-0.070	-0.102
	s.d.	0.077	0.151	0.028	0.097	0.053	0.088
	rmse	0.115	0.160	0.142	0.149	0.088	0.135
ARFIMA(1,0.4,1)	bias	-0.254	-0.002	-0.294	-0.098	-0.096	-0.122
$\phi_1 = 0.5$, $\theta_1 = 0.8$	s.d.	0.103	0.199	0.031	0.094	0.113	0.109
	rmse	0.274	0.198	0.296	0.135	0.148	0.164
ARFIMA(1,0.4,1)+WN	bias	-0.263	-0.051	-0.338	-0.214	-0.179	-0.176
$\phi_1 = 0.5$, $\theta_1 = 0.8$	s.d.	0.113	0.269	0.028	0.101	0.151	0.120
	rmse	0.285	0.273	0.339	0.236	0.234	0.213

The second process considered is FI(0.4) plus additive white noise where the variance of the white noise is equal to the innovation variance of the FI process. The estimation results are presented in Table 1 and Figure 2. In this experiment, the MLE based on the long-memory wavelet model with $K = 2$ beats the GPH estimators and the MLEs among all misspecified candidate models. It provides some evidence that the long-memory wavelet model is very feasible for approximating a long-memory process, and the corresponding MLE of d is fairly robust. Compared to LW1 and LW2, the GPH estimator is also robust but less efficient.

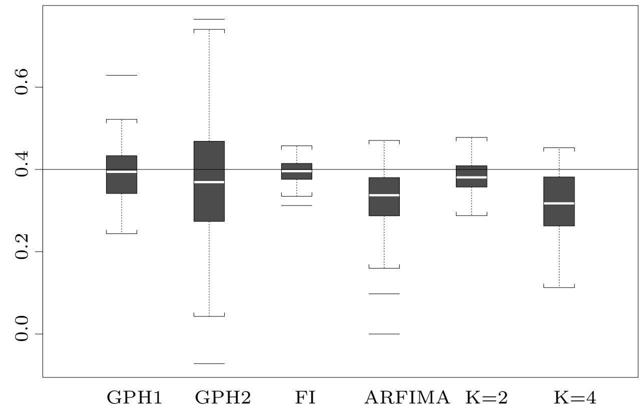

Figure 1. Boxplots of GPH estimates with $\alpha = 0.8, 0.6$ and ML estimates under FI(d), ARFIMA(1,d,1) and long-memory wavelet models with $K = 2, 4$ for FI(0.4) process. ($T = 1,024$, 100 replications).

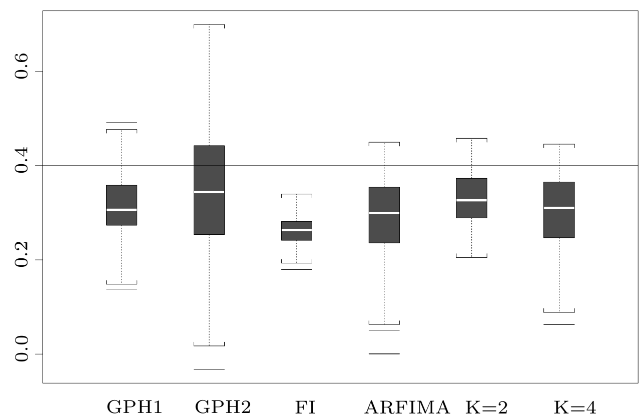

Figure 2. Boxplots of GPH estimates with $\alpha = 0.8, 0.6$ and ML estimates under FI(d), ARFIMA(1,d,1) and long-memory wavelet models with $K = 2, 4$ for FI(0.4) plus white noise process. ($T = 1,024$, 100 replications).

The third process considered is ARFIMA(1,0.4,1) with the difference equation: $(1-0.5B)(1-B)^{0.4}x_t = (1+0.8B)\epsilon_t$. The estimation results are presented in Table 1 and Figure 3. Similar to the first case, the MLE under the true ARFIMA(1,d,1) model has the best performance as expected. However, the ML estimator based on our wavelet model with $K = 2$ is fairly competitive to the best even though the model is misspecified. Moreover, the ML estimator based on our model with $K = 4$ performs better than the GPH estimators and the ML estimator based on a FI(d) model.

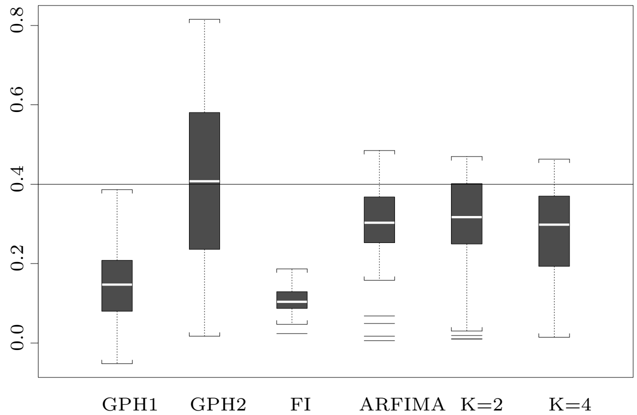

Figure 3. Boxplots of GPH estimates with $\alpha = 0.8, 0.6$ and ML estimates under FI(d), ARFIMA(1,d,1) and long-memory wavelet models with $K = 2, 4$ for ARFIMA(1,0.4,1) process. ($T = 1,024$, 100 replications).

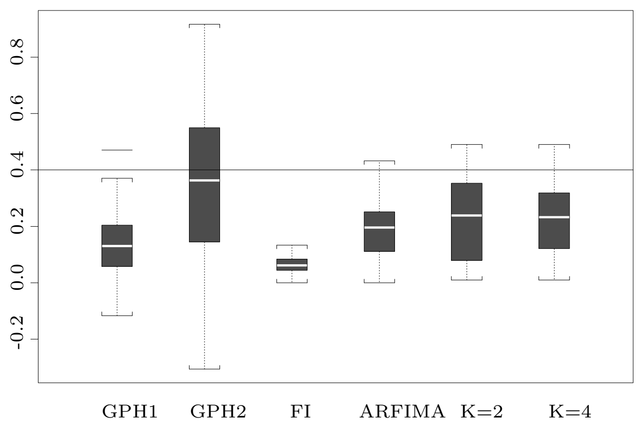

Figure 4. Boxplots of GPH estimates with $\alpha = 0.8, 0.6$ and ML estimates under FI(d), ARFIMA(1,d,1) and long-memory wavelet models with $K = 2, 4$ for ARFIMA(1,0.4,1) plus white noise process. ($T = 1,024$, 100 replications).

The fourth process considered is ARFIMA(1,0.4,1) plus additive white noise where the variance of the white noise is equal to the innovation variance of the ARFIMA process. The estimation results are presented in Table 1 and Figure 4. Again, the MLE based on our model with $K = 4$ has the best performance in terms of smallest root mean squared error.

As expected, the MLE for d based on our model performs well under various long-memory processes; it is robust subject to model misspecification and is more efficient than the commonly used semiparametric GPH estimator. These findings are consistent with those obtained by Jensen (2000) in which the likelihood function under ARFIMA process is approximated in the wavelet domain and the corresponding wavelet MLE of d is found to be robust even the data are contaminated.

4.2. Forecasting future observations

In this section, the forecasting of future observations based on our long-memory wavelet model is studied and compared to those under a FI model and under the true model. The four long-memory processes considered in the previous section are studied. Similar to Section 4.1, 100 realizations of length $1,024$ are generated for each process but only the first $T = 1,000$ observations are used for model fitting and the rest of data are reserved for examining the forecasting performance. In this study, the averaged prediction mean squared error

$$PMSE(h) = \frac{1}{h}\sum_{k=1}^{h}(\hat{y}_{1,000+k} - y_{1,000+k})^2$$

is used to evaluate the forecasting performance for each candidate model. In the following simulations, $PMSE(6)$, $PMSE(12)$ and $PMSE(24)$ are calculated for comparison.

For each data generating mechanism, the best linear predictions of $\{y_{1,000+h} : h = 1,\ldots,24\}$ are calculated based on the fitted long-memory wavelet model with $K = 2$, the fitted FI model with estimated parameters, and the true model with known parameters. In particular, under the fitted long-memory wavelet model, two types of predictors are calculated: one is based on the non-stationary covariance of the fitted model (denoted by LW1) and the other is based on a stationary covariance obtained by averaging all autocovariance functions across time (denoted by LW1-AVG). Under the fitted Gaussian FI model, the best prediction of $\boldsymbol{y}_0 = (y_{1,001},\ldots,y_{1,024})'$ given $\boldsymbol{y} = (y_1,\ldots,y_{1,000})'$ is

$$\hat{\boldsymbol{y}}_0 = \hat{\mu}\mathbf{1} + \hat{\Sigma}_2\hat{\Sigma}_1^{-1}(\boldsymbol{y} - \hat{\mu}\mathbf{1}), \tag{12}$$

where $\hat{\mu}$ is the MLE of the mean parameter, $\hat{\Sigma}_1$ is a $1,000\times 1,000$ variance matrix with ijth element $\hat{\gamma}(i-j)$ and $\hat{\Sigma}_2$ is a $24 \times 1,000$ covariance matrix with ijth element $\hat{\gamma}(1,000+i-j)$ in which

$$\hat{\gamma}(h) = \begin{cases} \hat{\sigma}^2\frac{\Gamma(1-2\hat{d})}{\Gamma^2(1-\hat{d})}; & \text{if } h = 0, \\ \hat{\gamma}(0)\frac{\Gamma(h+\hat{d})\Gamma(1-\hat{d})}{\Gamma(h-\hat{d}+1)\Gamma(\hat{d})}; & \text{if } h > 0, \\ \hat{\gamma}(-h); & \text{if } h < 0, \end{cases}$$

where $\hat{\sigma}^2$ is the MLE of the innovation variance in the fitted FI process. The best prediction of $\boldsymbol{y}_0$ under the true FI model has the same form as (12) with $\hat{\mu} = 0$,

$\hat{d} = 0.4$ and $\hat{\sigma}^2 = 1$. To avoid the circular effect on prediction under our model, the variance matrix with larger dimension m ($m > n$) is constructed as $\Sigma^* = (\boldsymbol{W}^*)'\hat{\Lambda}^*_w \boldsymbol{W}^* + \hat{\sigma}^2_\epsilon \boldsymbol{I}$, where $\boldsymbol{W}^*$ is the wavelet transform matrix for m-dimensional data with the same wavelet filter (symmlet wavelet s8) and the same resolution ($J = 7$), $\hat{\Lambda}^*_w$ has the same form as (7) but with estimated parameters, in which the dimension of each block-diagonal matrix is determined by the number of filters associated with the corresponding scale defined in $\boldsymbol{W}^*$. In this simulation, we use $m = 2,000$. Under the long-memory wavelet model, the prediction of $\boldsymbol{y}_0$ based on the non-stationary covariance is given by

$$\hat{\boldsymbol{y}}_0 = \hat{\beta}\mathbf{1} + \Sigma^*_2(\Sigma^*_1)^{-1}(\boldsymbol{y} - \hat{\beta}\mathbf{1}), \tag{13}$$

where $\hat{\beta}$ is the MLE of β in (6) which represents the mean, Σ^*_1 is a $1,000 \times 1,000$ variance matrix with ijth element γ^*_{ij} and Σ^*_2 is a $24 \times 1,000$ covariance matrix with ijth element $\gamma^*_{1,000+i,j}$ in which γ^*_{ij} is the ijth element in Σ^*. Similarly, the prediction of $\boldsymbol{y}_0$ based on a stationary version of Σ^* has the same form as (13), but replacing Σ^* by Σ^{**} where the ijth element in Σ^{**} satisfies $\gamma^{**}_{ij} = \gamma^{**}(|i-j|)$ and $\gamma^{**}(h)$ is the average over all possible $\gamma^*_{i,i+h}$ in Σ^*.

The averaged prediction mean squared errors under different models are presented in Table 2. Under the true FI process, the fitted FI model has the best forecasting performance. The fitted long-memory wavelet model is competitive to the true model. Moreover, the prediction based on the averaged autocovariances has better performance than that based on the autocovariances without averaging, but the difference is small. For other data generating mechanisms, the predictions based on all models considered are fairly competitive.

Table 2. Averaged prediction mean squared errors for predicting future observations ($t = 1,001, \ldots, 1,024$) under true model, the fitted FI(d) model and the long-memory wavelet model with $K = 2$ under various long-memory processes. ($T = 1,000$, 100 replications).

True Model	Averaged PMSE	LW1	LW1-AVG	FI	TRUE
FI(0.4)	PMSE(6)	1.5121	1.5056	1.3064	1.4968
	PMSE(12)	1.5770	1.5732	1.4094	1.5632
	PMSE(24)	1.6242	1.6092	1.4627	1.5938
FI(0.4)+WN	PMSE(6)	2.2847	2.2787	2.2897	2.2812
	PMSE(12)	2.4164	2.4151	2.4112	2.4190
	PMSE(24)	2.4580	2.4448	2.4311	2.4433
ARFIMA(1,0.4,1)	PMSE(6)	1.0411	1.0388	1.0399	1.0276
$\phi_1 = 0.5, \theta_1 = 0.8$	PMSE(12)	0.9591	0.9588	0.9644	0.9557
	PMSE(24)	1.0172	1.0179	1.0263	1.0128
ARFIMA(1,0.4,1)+WN	PMSE(6)	2.2345	2.2296	2.2314	2.2177
$\phi_1 = 0.5, \theta_1 = 0.8$	PMSE(12)	2.0927	2.0887	2.0974	2.0818
	PMSE(24)	2.0422	2.0407	2.0494	2.0298

5. Application

We apply the long-memory wavelet model to the Nile River data which are yearly minimum water levels as measured at the Roda Gauge near Cairo (Beran (1994)). The data set contains 663 observations for the years 622AD to 1284AD, where the data are centered about the sample mean and divided by the sample standard deviation for further analysis. The first 650 observations are used in model fitting and the rest of the data are reserved for prediction. In Figure 5, the sample autocorrelations based on the first 650 standardized observations (light vertical lines) show the slow decay suggestive of long memory. We fit the zero-mean, long-memory wavelet models with several different J and K to the first 650 standardized data and the estimation results are summarized in Table 3. From Table 3, the estimates of the long-memory parameter are fairly robust and range from 0.34 to 0.40 under different choices of (J, K). Also, the estimates for τ_j^2 with $j \neq -1$ are robust among different settings indicating the choice of (J, K) is not crucial in our modeling procedures. The estimates $\hat{\tau}_{-1}^2$ and $\hat{\sigma}_\epsilon^2$ are highly correlated, which is reasonable because both of them characterize the high-frequency variations in the data and it is more difficult to identify them individually. When more parameters $\{\tau_j^2\}$ are added to the model to enrich the short-range dependence, σ_ϵ^2 is essentially zero for this particular data set.

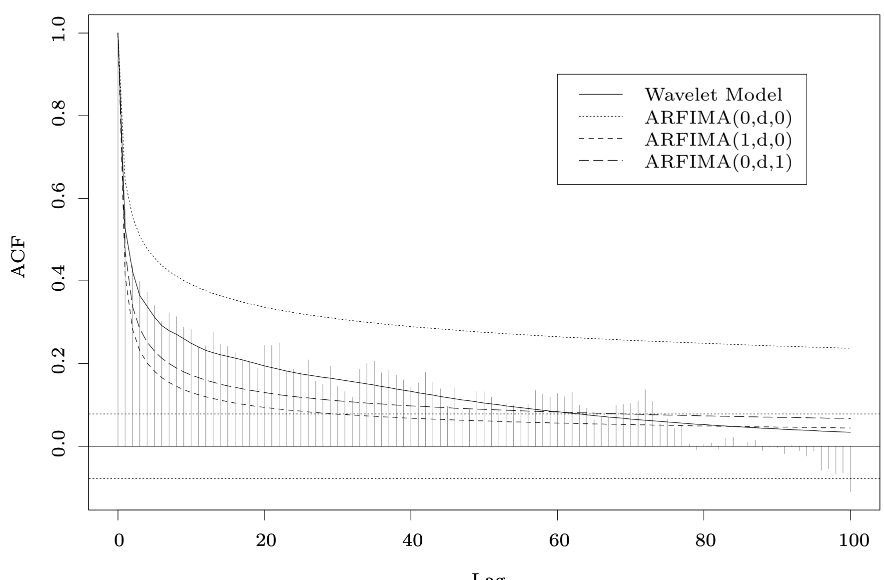

Figure 5. Sample autocorrelations and the fitted autocorrelations under the long-memory wavelet model with $(J, K) = (5, 0)$ and ARFIMA models for standardized Nile River data.

Table 3. Maximum likelihood estimates of parameters in the long-memory wavelet models for the standardized Nile River data.

J	K	$\hat{d}$	$\hat{\tau}_0^2$	$\hat{\phi}$	$\hat{\sigma}_\epsilon^2$	$\hat{\tau}_{-1}^2$	$\hat{\tau}_{-2}^2$	$\hat{\tau}_{-3}^2$	$\hat{\tau}_{-4}^2$	AIC
6	4	0.322	1.096	0.311	0.000	0.390	0.695	0.837	1.543	369.9
6	3	0.340	0.994	0.325	0.000	0.390	0.695	0.837		368.1
6	2	0.361	0.938	0.361	0.000	0.390	0.695			366.6
6	1	0.364	0.852	0.341	0.182	0.209				365.7
6	0	0.337	0.970	0.313	0.003					364.2
5	3	0.344	0.969	0.498	0.000	0.390	0.695	0.837		365.8
5	2	0.361	0.921	0.517	0.000	0.390	0.695			364.1
5	1	0.374	0.798	0.519	0.243	0.147				363.0
5	0	0.340	0.950	0.486	0.019					361.9
4	2	0.362	0.912	0.552	0.000	0.390	0.695			364.3
4	1	0.393	0.724	0.585	0.338	0.052				362.7
4	0	0.347	0.938	0.541	0.040					362.2
3	1	0.407	0.678	0.643	0.383	0.010				363.9
3	0	0.355	0.920	0.581	0.068					364.9

Based on the AIC, the best long-memory wavelet model for the standardized data is the one with $(J, K) = (5, 0)$. The averaged autocorrelation function (averaged across time) for this fitted model is presented in Figure 5 with the solid line. For comparison, three simple ARFIMA models, including a FI(d), an ARFIMA(1,d,0) and an ARFIMA(0,d,1), are fitted to the same data and their corresponding fitted autocorrelation functions are also displayed in Figure 5 with the broken lines. Clearly, the fitted long-memory wavelet model captures the empirical dependence better than any of the ARFIMA models considered. For checking the performance of prediction for the last 13 observations ($t = 651, \ldots, 663$), the averaged prediction mean squared errors (PMSE) $\sum_{t=651}^{663}(y_t - \hat{y}_t)^2/13$ under the fitted long-memory wavelet model with $(J, K) = (5, 0)$ is 0.735 which is slightly smaller than the averaged PMSEs of 0.745, 0.807 and 0.792 under the fitted FI(d), ARFIMA(1,d,0) and ARFIMA(0,d,1), respectively.

6. Conclusions

According to Craigmile, Percival and Guttorp (2001), the between-scale correlations of DWT for a FI process can be made arbitrary small by increasing the length of the wavelet filter, however the within-scale correlations between wavelet coefficients would be fitted better by considering an AR dependence structure. Following this fact, our wavelet model can be extended by allowing an AR dependence between wavelet coefficients within some scales (particularly for larger scales). The extended model becomes even more feasible in approximating both long-memory and short-memory behaviors for a stationary process.

The proposed wavelet model can be easily incorporated into other useful long-memory models such as long-memory stochastic volatility models, multivariate long-memory models, common factor models with long-range dependence and generalized linear long-memory models. As another advantage, for analyzing large data sets, the model selection issue can actually be taken into account in the class of long-memory wavelet models, owing to fast computations. This capability is difficult to achieve by conventional ARFIMA models due to time-consuming computations.

Acknowledgement

This research was supported by National Science Council (Grant No:NSC 92-2118-M-007-012). The author is thankful to Dr. H.-C. Ho, the Editor and two anonymous referees for their helpful comments.

References

Allan, D. W. (1966). Statistics of atomic frequency clocks. *Proc. IEEE* **31**, 221-230.

Beran, J. (1994). *Statistical Methods for Long-memory Processes*. Chapman and Hall, New York.

Craigmile, P. F., Percival, D. B. and Guttorp, P. (2001). The impact of wavelet coefficient correlations on fractionally differenced process estimation. In *European Congress of Mathematics* **II** (Edited by C. Casacuberta, R. M. Miro-Roig, J. Verdera and S. Xambo-Descamps), 591-599. Birkhauser Verlag, Basel.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* CBMS-NSF Series in Applied Mathematics 61. SIAM, Philadelphia.

Dijkerman, R. and Mazumdar, R. (1994). On the correlation structure of the wavelet coefficients of fractionally Brownian motion. *IEEE Trans. Inform. Theory* **40**, 1609-1612.

Fan, Y. Q. (2003). On the approximate decorrelation property of the discrete wavelet transform for fractionally differenced processes. *IEEE Trans. Inform. Theory* **49**, 516-521.

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Trans. Inform. Theory* **38**, 910-917.

Fox, R. and Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Statist.* **14**, 517-532.

Geweke, J. F. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *J. Time Ser. Anal.* **4**, 221-238.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Anal.* **1**, 15-29.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika* **68**, 165-176.

Jensen, M. J. (1999a). Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter. *J. Forecasting* **18**, 17-32.

Jensen, M. J. (1999b). An approximate wavelet MLE of short and long memory parameters. *Stud. Nonlinear Dynamics and Econometrics* **3**, 239-253.

Jensen, M. J. (2000). An alternative maximum likelihood estimator of long-memory processes using compactly supported wavelets. *J. Econom. Dynam. Control* **24**, 361-387.

Masry, E. (1993). The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion. *IEEE Trans. Inform. Theory* **39**, 260-264.

McCoy, E. J. and Walden, A. T. (1996). Wavelet analysis and synthesis of stationary long-memory processes. *J. Comput. Graph. Statist.* **5**, 26-56.

Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Robinson, P. M. (1995). Gaussian semi-parametric estimation of long range dependence. *Ann. Statist.* **23**, 1630-1661.

Searle, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics* **26**, 505-524.

Sowell, F. B. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *J. Econometrics* **53**, 165-188.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. John Wiley, New York.

Wornell, G. W. (1993). Wavelet-based representations for the $1/f$ family of fractal processes. *Proc. IEEE* **81**, 1428-1450.

Institute of Statistics, National Tsing-Hua University, Hsin-Chu, Taiwan 30043.

E-mail: njhsu@stat.nthu.edu.tw

(Received November 2003; accepted June 2005)