Modified periodogram method for estimating the Hurst exponent of fractional Gaussian noise

Yingjun Liu,[1,2] Yong Liu,[2] Kun Wang,[2] Tianzi Jiang,[2,*] and Lihua Yang[1]
[1]*School of Mathematics and Computing Science, Sun Yat-Sen University, Guangzhou, China*
[2]*LIAMA Center for Computational Medicine, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
(Received 10 March 2009; revised manuscript received 13 August 2009; published 17 December 2009)

Fractional Gaussian noise (fGn) is an important and widely used self-similar process, which is mainly parametrized by its Hurst exponent (H). Many researchers have proposed methods for estimating the Hurst exponent of fGn. In this paper we put forward a modified periodogram method for estimating the Hurst exponent based on a refined approximation of the spectral density function. Generalizing the spectral exponent from a linear function to a piecewise polynomial, we obtained a closer approximation of the fGn's spectral density function. This procedure is significant because it reduced the bias in the estimation of H. Furthermore, the averaging technique that we used markedly reduced the variance of estimates. We also considered the asymptotical unbiasedness of the method and derived the upper bound of its variance and confidence interval. Monte Carlo simulations showed that the proposed estimator was superior to a wavelet maximum likelihood estimator in terms of mean-squared error and was comparable to Whittle's estimator. In addition, a real data set of Nile river minima was employed to evaluate the efficiency of our proposed method. These tests confirmed that our proposed method was computationally simpler and faster than Whittle's estimator.

DOI: 10.1103/PhysRevE.80.066207 PACS number(s): 05.45.Df

I. INTRODUCTION

Methods that utilize the characteristics of self-similar processes can effectively model many natural observations. Self-similarity implies that an object looks similar to its zoomed part and has been exploited in many scientific fields, such as geophysics [1,2], traffic flow [3], textures [4,5], biology [6,7], and functional brain signals [8,9]. The relevance of the self-similar property derives from its ability to capture the inner nature of data without introducing artificial and cumbersome models. Among the emerging self-similar models (e.g., [10,11]) are two frequently studied processes, i.e., fractional Gaussian noise (fGn) process and the fractional autoregressive integrated moving average (FARIMA) process. We will confine our attention to fGn, yet almost all the following parameter estimators could apply to either fGn or FARIMA.

Only two parameters, the Hurst exponent (H) and variance, completely specify fGn. Of these, H is the primary parameter. (We will return to a further and more detailed consideration of fGn in Sec. II.) Many articles have proposed methods for estimating the Hurst exponent (H) of fGn. As suggested by Jeong *et al.* [12], the classical methods for estimating H of fGn fall logically into three domains. The first is in the time domain, which includes the aggregated variance method [13], Higuchi's method [14], the rescaled range statistics method [15], and the detrended fluctuation analysis method [6]. The second is in the frequency domain and includes the periodogram method [16], the modified periodogram method [13], and Whittle's estimator [11,17]. The third domain is in the wavelet domain, which includes Wornell's estimator [18], an estimator based on discrete variations of fractional Brownian motion (fBm) [19–21], the wavelet maximum likelihood (WML) estimator [22], and the Abry-Veitch Daubechies wavelet-based estimator [23].

In the past few years, various researchers introduced several new estimators, including the wavelet transform modulus maxima estimator [24], an estimator based on the sample autocorrelation function [25], and a bias-corrected version of the rescaled range statistics estimator [26]. In the present study, we put forward a modified periodogram method for estimating the Hurst exponent. As expected, the proposed method resulted in less bias than did the original periodogram method. Monte Carlo simulation results showed that the proposed estimator is reasonably good at estimating the Hurst exponent of fGns. Moreover, its numerical complexity is low because it is computationally simple.

The paper is organized as follows. Section II contains a brief description of fGn. In Sec. III, we introduce the traditional periodogram method and consider a few modified versions. Then we propose our modified periodogram method in Sec. IV. We provide Monte Carlo simulation results in Sec. V. A real data set of Nile river minima is used as an illustrative example to validate the estimators in Sec. VI. Finally, we end with a discussion and conclusions.

II. FRACTIONAL GAUSSIAN NOISE

fGn $G=(G_t:t=1,2,\ldots,N)$ is a zero mean stationary process, which is defined as the stationary increment of fractional Brownian motion [27]. The Hurst exponent $H \in (0,1)$ and variance $\sigma^2 := \mathrm{Var}(G_t)$ fully characterize fGn. The distribution of fGn can be completely specified by its autocovariances at lags $\tau \in Z$ [11],

$$c(\tau) := \frac{\sigma^2}{2}(|\tau+1|^{2H} - 2|\tau|^{2H} + |\tau-1|^{2H}). \quad (1)$$

The Hurst exponent is a measure of the strength of dependence between the discrete time points G_t, whereas the vari-

*Corresponding author; jiangtz@nlpr.ia.ac.cn

ance is only a scale parameter. Equation (1) reveals that fGn is white Gaussian noise when $H=0.5$. A Hurst exponent $H \in (0,0.5)$ means that fGn is negatively correlated or antipersistent, whereas $H \in (0.5,1)$ means that it is positively correlated or has a long memory [11]. The spectral density function (SDF) of fGn is precisely defined as the Fourier transform of its autocovariance sequence, that is,

$$S(f) := \sum_{\tau=-\infty}^{\infty} c(\tau) e^{-i2\pi f\tau} \tag{2}$$

$$= 4\sigma^2 C_H [\sin^2(\pi f)] \sum_{j=-\infty}^{\infty} \frac{1}{|f+j|^{2H+1}}, \quad -\frac{1}{2} \le f \le \frac{1}{2}, \tag{3}$$

with $C_H := \Gamma(2H+1)\sin(\pi H)/(2\pi)^{2H+1}$ [11,28].

III. PERIDODOGRAM METHOD

Using a Taylor expansion to Eq. (3), we obtain [11]

$$S(f) \approx \sigma^2 C_H (2\pi)^2 |f|^{1-2H}, \quad -\frac{1}{2} \le f \le \frac{1}{2} \tag{4}$$

in the neighborhood of the origin, and hence SDF is related to frequency by a power law with spectral exponent

$$\gamma = 1 - 2H, \quad -1 < \gamma < 1, \tag{5}$$

i.e.,

$$S(f) \propto |f|^{\gamma} \tag{6}$$

or, equivalently,

$$\ln S(f) \propto \gamma \ln |f|. \tag{7}$$

The approximate linear relationship between $\ln S(f)$ and $\ln|f|$ can be fitted using a least-squares estimate (LSE). Then the slope of the regression line is an estimate of the spectral exponent which can be converted to the estimated H by Eq. (5). As a result, the procedure, based on a logarithmic periodogram, is called the periodogram method [16]. In practice, the linear regression is restricted around the origin in the frequency domain and only the low-frequency part of $O(N^{4/5})$ (N is the length of the time series) of the frequency range should be employed [29].

The periodogram method is a simple and quick estimator. It has received extensive attention, including, for example, the proof of consistency and asymptotic normality for a modified periodogram method provided by Robinson [30], the mean-squared error (MSE) of the periodogram method studied by Hurvich *et al.* [29], and the bias testing by Davidson and Sibbertsen [31] However, some previous studies indicate that the periodogram method leads to a poor estimate of H, especially when the time series length is short [13,20,32]. Flandrin showed that continuous fGn yields a well-defined power spectrum which exactly obeys a power law over all frequencies [33]. However, this does not hold for discrete fGn (for convenience we continue to use the term "fGn" to represent a discrete fGn in this study) [34]. Numerous variants of the periodogram method have been proposed. Shimotsu and Phillips advocated a pooled periodogram regression estimator to allow the use of a larger number of periodogram ordinates to reduce the variance without significantly changing the bias [35], Andrews and Guggenberger included polynomial terms in the frequencies in the narrow-band regression [36], and Moulines and Soulier developed a broadband regression with dummy variables alone controlling for the short-run effects [37]. Each of these three methods is based more or less on the approximation of SDF given in Eq. (4), which will be confined in the linear form of the spectral exponent. As a result, these methods may not perform satisfactorily for short time series. This is why the narrow-band regression of the periodogram method is constrained in the neighborhood of zero. When the logarithmic periodogram is performed over the entire range of the frequency domain, it is referred to as a broadband regression [37]. We will propose a quite different modified periodogram method of broadband regression in the following section.

IV. MODIFIED PERIODOGRAM METHOD

Given a time series $X=(X_0,X_1,\ldots,X_{N-1})$ of length N, a basic estimate of SDF is provided by the periodogram,

$$\hat{S}_N^{(p)}(f) = \frac{1}{N} \left| \sum_{t=0}^{N-1} X_t e^{-i2\pi ft} \right|^2. \tag{8}$$

The expected value of the periodogram, $E[\hat{S}_N^{(p)}]$, is given by Percival and Walden as [38]

$$E[\hat{S}_N^{(p)}(f)] = \frac{\sigma^2}{2N} \sum_{j=-N+1}^{N-1} (|j-1|^{2H} - 2|j|^{2H} + |j+1|^{2H}) \times \cos(2\pi fj)(N-|j|). \tag{9}$$

The estimated spectrum above is an asymptotically unbiased estimator of $S(f)$ [38]. Thus, for any particular frequency, f, let $S_N^{(p)}(f) = E[\hat{S}_N^{(p)}(f)]$, with

$$\lim_{N\to\infty} S_N^{(p)}(f) = S(f). \tag{10}$$

So in practice we have to substitute $S_N^{(p)}(f)$ for $S(f)$. In order to refine the poor approximation that results from Eq. (4), we assumed the exponent to be some function of H, that is,

$$S_N^{(p)}(f) \propto |f|^{\alpha(H)}. \tag{11}$$

Similar to Eq. (7), we have,

$$\ln S_N^{(p)}(f) \propto \alpha(H) \ln |f|. \tag{12}$$

For practical applications, we need to smooth the estimated SDF to reduce the variance. Let $\hat{S}_N^{(p)}(f)$ be the estimated SDF and its smoothed version, $\hat{S}_{N,A}^{(p)}(f)$. To be consistent with $\hat{S}_{N,A}^{(p)}(f)$, we performed the same smoothing procedure on $S_N^{(p)}(f)$ and thus got $S_{N,A}^{(p)}(f)$ (see Fig. 1). After that, we let H vary in the interval (0,1); then we used the LSE of $\ln S_{N,A}^{(p)}(f)$ versus $\ln|f|$ in Eq. (12) to obtain α. Since α is independently related to both H and the data length (N),

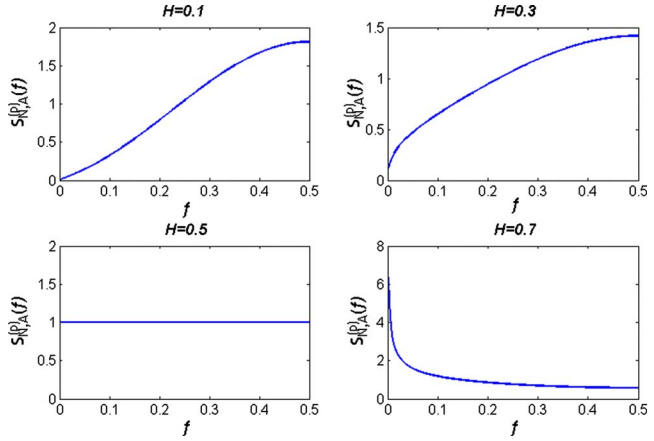

FIG. 1. (Color online) Plots of $S_{N,A}^{(p)}(f)$ versus f. $S_{N,A}^{(p)}(f)$ is smoothed version of the expectation shown in Eq. (9), i.e., $S_N^{(p)}(f)$. The shape of $S_{N,A}^{(p)}(f)$ varies with H. We set $\sigma^2=1$, $N=600$ (only a few H values are shown for simplicity).

we denoted it as $\alpha_{N,H}$. The relationship is shown in Fig. 2. Then, we performed a polynomial fitting of $\alpha_{N,H}$ and H. It is natural to deal with $H<0.5$ and $H>0.5$ separately because they have quite different properties. In the following simulations, we considered different order polynomials and found little difference in efficiency when the order was greater than or equal to 3. In particular, a seventh-order polynomial can provide a good fit empirically. Let $P_L(H)=\Sigma_{i=0}^{7}c_iH^i$ and $P_R(H)=\Sigma_{i=0}^{7}d_iH^i$ be the piecewise fitted polynomial. Then we have

$$\alpha_{N,H} \approx \begin{cases} P_L(H), & H < 0.5 \\ P_R(H), & H \geq 0.5. \end{cases} \tag{13}$$

Note that in practical applications the frequencies (f) are discrete. Without a loss of generality, we assumed a sampling rate equal to 1. Using the direct periodogram method, we then obtained the estimated SDF as $\hat{S}_N^{(p)}(f_i)$, $f_i=i/N$, where $-N/2\leq i\leq N/2-1$. We restricted our study to real-valued time series, so that $\hat{S}_N^{(p)}(f_i)$ is symmetrical. Thus, we only have to consider the last half, i.e., $\hat{S}_N^{(p)}(f_i)$ (with $i=1,\ldots,N/2-1$).

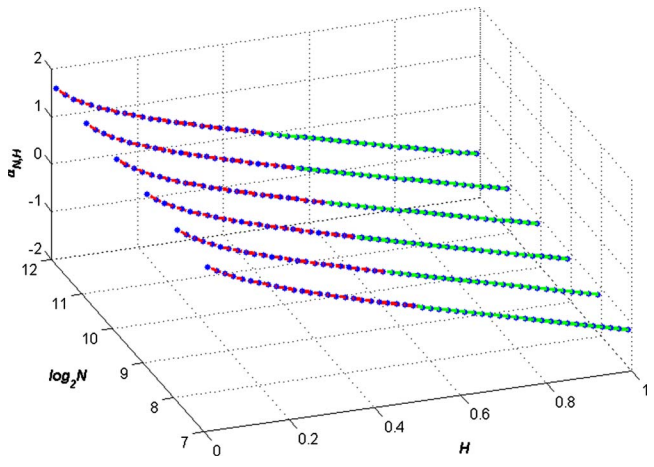

FIG. 2. (Color online) The relationship between $\alpha_{N,H}$, N, and H. Data length N is from 2^7 to 2^{12}. The Hurst exponent H varies from 0 to 1 with increments of 0.01. Blue dots indicate values of $\alpha_{N,H}$ with different N and H (only half of dots are drawn for viewing). Dashed red curves are fitted polynomial $[P_L(H)]$ when $H<0.5$. Solid green curves are the fitted polynomial $[P_R(H)]$ when $H>0.5$.

Therefore, the modified H estimation procedures can be fully specified as follows:

Step 1. Let H vary by small increments over the interval (0,1) and then use LSE to calculate α, i.e., $\alpha_{N,H}$, with the given data length N, that is,

$$(\alpha_{N,H}, c_{N,H}) = \arg\min_{(\alpha,c)} \sum_{i=1}^{N/2-1} |\ln S_{N,A}^{(p)}(f_i) - \alpha \ln|f|_i - c|^2, \tag{14}$$

where $f_i=i/N$ and c is a constant.

Step 2. Find the piecewise polynomial fitting of $\alpha_{N,H}$ and H in Eq. (14), i.e., $\alpha(H)$.

Step 3. Estimate the SDF of the given time series X_t $(0\leq t\leq N-1)$ via the direct periodogram method, i.e., $\hat{S}_N^{(p)}(f_i)$ $(1\leq i\leq N/2-1)$.

Step 4. Get the smoothed version $\hat{S}_{N,A}^{(p)}(f_i)$ by averaging the estimated SDF $\hat{S}_N^{(p)}(f_i)$.

Step 5. Perform a linear regression fitting of $\ln \hat{S}_{N,A}^{(p)}(f_i)$ versus $\ln|f_i|$. The slope K and intercept C are given by

$$(K, C) = \arg\min_{(k,c)} \sum_{i=1}^{N/2-1} |\ln \hat{S}_{N,A}^{(p)}(f_i) - k \ln|f_i| - c|^2. \tag{15}$$

Step 6. In the end, a proper $\hat{H}$ in the interval (0,1), which is the estimate of the Hurst exponent of X_t, can be found as the solution to the equation $P_L(H)=K$ [or $P_R(H)=K$] by a dichotomous search method.

According to Eq. (14), $\alpha_{N,H}$ is the optimal α from the perspective of LSE, which certainly refines the approximation of the SDF compared with the original approximated version in Eq. (6). The piecewise polynomial $\alpha(H)$ is a generalization of the linear function $(\gamma=1-2H)$. However, in steps 1 and 2, the accuracy of $\alpha(H)$ is related to the size of the increments of H in the interval (0,1). The smaller the increment, the greater the accuracy of $\alpha(H)$ that can be obtained. On the other hand, the smaller the increment, the longer the computing time needed. Hence, a compromise must be made between accuracy and computing time. We chose an increment of 0.01 in the simulation.

Note that $\alpha(H)$ is a modified version of the linear function $\gamma=1-2H$, and it is a monotonously decreasing function of H (Fig. 2). This will ensure only one proper $\hat{H}$ will be found in step 6.

It is not difficult to see that, for the same data length N, $\alpha(H)$ only has to be calculated once (steps 1 and 2 above) for either simulated or actual data. Hence, the proposed method is appropriate for the analysis of large samples of fGns that have the same data length. It is commonly known that a fast Fourier transform requires only $O(N \ln N)$ operations and

that the dichotomous search method demands $O(\ln N)$ operations. Thereby our algorithm has a numerical complexity on the order of $O(N \ln N)$.

Theoretically, let $x_i = \ln|f_i|$ and $\bar{x} = [1/(N/2-1)]\Sigma_{i=1}^{N/2-1} \ln|f_i|$, from Eq. (15), we have

$$\hat{\alpha}(H) = K = \frac{\sum_{i=1}^{N/2-1} \ln \hat{S}_{N,A}^{(p)}(f_i)(x_i - \bar{x})}{\sum_{i=1}^{N/2-1} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{N/2-1} \{\ln S_{N,A}^{(p)}(f_i) + \ln[\hat{S}_{N,A}^{(p)}(f_i)/S_{N,A}^{(p)}(f_i)]\}(x_i - \bar{x})}{\sum_{i=1}^{N/2-1} (x_i - \bar{x})^2}. \tag{16}$$

For simplicity, we will consider the estimate without smoothing. In this situation, we let $\hat{\alpha}_1(H)$ and $\hat{H}_1$ instead of $\hat{\alpha}(H)$ and $\hat{H}$, respectively. The expression of $\hat{\alpha}_1(H)$ is given as follows:

$$\hat{\alpha}_1(H) = \frac{\sum_{i=1}^{N/2-1} \{\ln S_N^{(p)}(f_i) + \ln[\hat{S}_N^{(p)}(f_i)/S_N^{(p)}(f_i)]\}(x_i - \bar{x})}{\sum_{i=1}^{N/2-1} (x_i - \bar{x})^2}. \tag{17}$$

$E[\ln \hat{S}_N^{(p)}(f_i)]$ is not equal to $\ln E[\hat{S}_N^{(p)}(f_i)]$ because of the known the properties of expectation. From the work of Geweke and Porter-Hudak [16], we know that when f approaches zero, the term $\ln\{\hat{S}_N^{(p)}(f)/S(f)\}$ is asymptotically independent and identically distributed (i.i.d.) and its asymptotic mean is $-\gamma$ (γ is Euler's constant 0.577 21…) and its variance is $\pi^2/6$. It would be not too difficult to extend the results to the whole frequency domain for the terms $\ln\{\hat{S}_N^{(p)}(f)/S_N^{(p)}(f)\}$ since the equation $E[\hat{S}_N^{(p)}(f)] = S_N^{(p)}(f)$ holds for arbitrary N and f. Then we could derive $E[\hat{\alpha}_1(H)] \approx \alpha_1(H)$ and $\mathrm{Var}[\hat{\alpha}_1(H)] \approx \pi^2/3N$ from Eq. (16). Therefore, the estimate $\hat{H}_1$ is asymptotically unbiased. Moreover, taking advantage of the simple linear approximate relationship of $\alpha(H)$ and H, i.e., $\alpha(H) \approx 1-2H$, the estimate $\hat{H}_1$'s variance is approximately equal to $\pi^2/12N$. Furthermore, we found experimentally that the distribution of the terms $\ln\{\hat{S}_N^{(p)}(f)/S_N^{(p)}(f)\}$ is approximately normal; so, conveniently, we can assume that the estimate $\hat{H}_1$ obeys the normal distribution, i.e., $\hat{H}_1 \sim N(H, \pi^2/12N)$. We can obtain its confidence intervals $(H - [\pi/\sqrt{12N}]Z_{\beta/2}, H + [\pi/\sqrt{12N}]Z_{\beta/2})$ (where $1-\beta$ is the confidence level and $Z_{\beta/2}$ is the $1-\beta/2$ quantile of a standard normal distribution). Similarly, we can deduce the statistical results of $\hat{\alpha}(H)$ and $\hat{H}$; that is, the estimated $\hat{H}$ is also asymptotically unbiased and yields an asymptotic i.i.d. sample. However, due to the smoothing procedure, we find that, if $\mathrm{Var}[\hat{\alpha}(H)] < \mathrm{Var}[\hat{\alpha}_1(H)]$, then $\mathrm{Var}(\hat{H}) < \mathrm{Var}(\hat{H}_1)$. The confidence intervals of the estimate $\hat{H}$ would also be smaller than the estimated $\hat{H}_1$. These results will be verified in the later simulations.

V. MONTE CARLO SIMULATIONS

In this section, we present the results of Monte Carlo simulations using simulated data to cross validate the modified periodogram estimator (MPE) by comparison with the original periodogram estimator (PE) and three other commonly used estimators: Whittle's estimator [17], the WML, estimator [22,39] and an estimator which is based on discrete variations of the fBm [19–21,40]. Each of these estimators is described in greater detail in the Appendix.

A. Simulated data

Several algorithms [11,41], which have been comparatively evaluated [20], are available to simulate fGn. We adopted the Wood-Chan method, initially proposed by Davies and Harte [42] and improved later by Wood and Chan [43], to generate the fGn simulations because it has been established as being exact and fast [20].

First, we performed an estimation of the simulated data, each of which consisted of 1000 fGns of various lengths ($N=600, 400, 200$) and with different values of H ($H=0.1, 0.2, \ldots, 0.9$). Then we produced 1000 simulations of fGns for each H (only $H=0.3, 0.5, 0.7$ are considered here for the sake of brevity) with lengths varying from $N=2^7$ to $N=2^{12}$ to test how errors of estimates varied with increasing data length. For simplicity, we set $\sigma^2=1$ for all simulations since σ is only a scale parameter.

B. Results

The results are summarized in Figs. 3 and 4. Specifically, Fig. 3 presents the results of the estimators in the form of box plots, which summarize the deviation of the estimated H from the nominal values for the estimators (only the results of $N=600$ are presented for the sake of brevity). Figure 4 gives a detailed description of the abilities of the estimators for estimating fGns of different values of H or various data lengths. Efficiency is quantified in terms of the root-mean-squared error (RMSE) (i.e., $\sqrt{\text{variance}+\text{bias}^2}$) between the nominal and the estimated H.

Comparative evaluation of the deviation of MPE and PE from the nominal values is shown in the first two boxes in each panel of Fig. 3. The corresponding RMSE is shown in Fig. 4 (the top row of the first column). It is clear that PE is a fairly rough estimator and is only acceptable near $H=0.5$ (Fig. 3). PE seems to be greatly underestimated when $H<0.5$ and overestimated when $H>0.5$ (Fig. 3). MPE makes a definite improvement; not only does the median of the estimated H agree with the nominal H, but also the variance is smaller, especially when estimating extreme values of H (Fig. 3). From box plots in Fig. 3 and RMSE in Fig. 4 (where the bottom row is a close-up of the first row without PE), we found that all methods except PE are reasonably good esti-

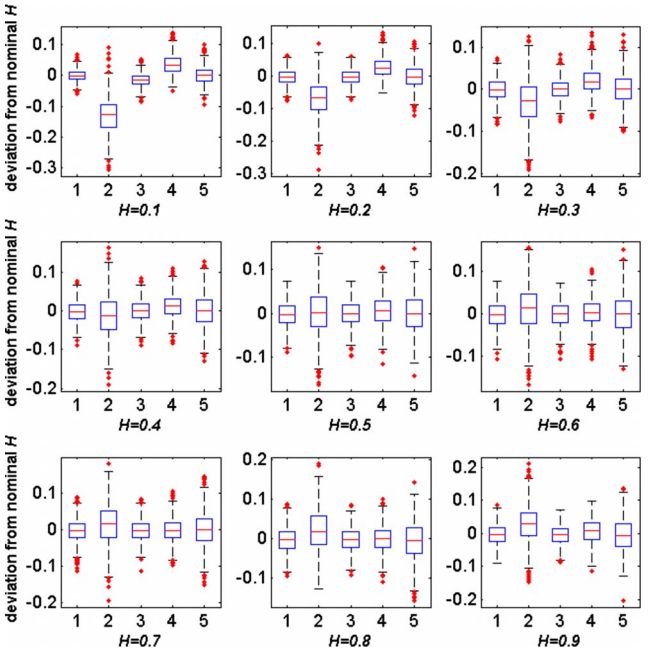

FIG. 3. (Color online) Comparative evaluation of bias and efficiency of five estimators of the Hurst exponent H of fGn with $N=600$. Box plots in each panel summarize the deviation of the estimated H from the nominal values for the estimators. The variance of all simulations was set at $\sigma^2=1$. (Estimators) 1: MPE; 2: PE; 3: Whittle; 4: WML; 5: fBm.

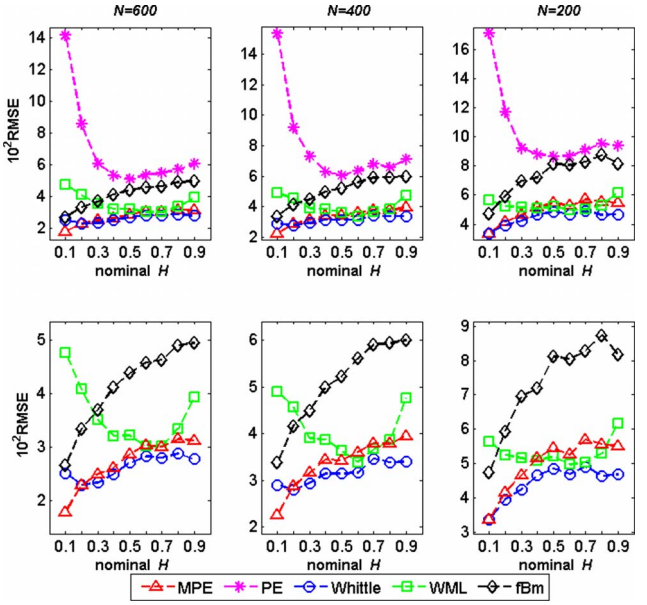

FIG. 4. (Color online) RMSE in estimates of the Hurst exponent according to the estimators involved. Left column: data length $N=600$; middle column: data length $N=400$; right column: data length $N=200$. Top row: RMSE of five estimators when the H value is from 0.1 to 0.9 in increments of 0.1; bottom row: a close-up of the low values of RMSE showed in the top row with PE omitted. Point marker codes for estimator: MPE (△), PE (∗), Whittle (○), WMNL (□), and fBm (◇).

mators of H when $H=0.1-0.9$. However, the WML estimator was biased toward overestimation when $H<0.5$. Additionally, the fBm-based estimator suffers from a large variance. As a result, in terms of RMSE, Whittle's estimator showed the best performance of the four estimators, and the next best was MPE.

Note that an important issue in studying fGn is the length of the time series that is needed to produce an estimate of acceptable precision. To investigate this, we examined the estimators' standard deviation (SD) and RMSE at different lengths of the time series, i.e., from $N=2^7$ to $N=2^{12}$. We produced 1000 simulations of fGn for each H (only $H=0.3,0.5,0.7$ are considered here for the sake of brevity) and evaluated the Hurst exponent for each time series. The results are depicted in Fig. 5, which reflects the estimators' performance for fGns of different lengths.

We found that the SD and RMSE of the estimators all decreased generally as the data length increased. In particular, for $N\geq 4096$, all the estimators have largely equivalent performance in terms of MSE (see RMSE shown in the right column of Fig. 5), which also suggests that the proposed estimator is reasonably good.

In order to verify the theoretical derivation in Sec. IV, we compared the approximate theoretical SD and confidence intervals with the Monte Carlo counterparts from the simulated data (the same data used in Fig. 5) with and without smoothing (the confidence level was chosen as 0.95) (Fig. 6). Both the Monte Carlo SD and confidence intervals of the MPE estimate with smoothing ($\hat{H}$) are obviously smaller than the MPE estimate without smoothing ($\hat{H}_1$); this finding validates the advantage of the smoothing procedure in estimation. Note that the approximate SD and confidence intervals (dashed blue lines) behave less efficiently than the Monte Carlo mainly because of the simple linear relationship between $\alpha(H)$ and H that was adopted in the theoretical derivation in Sec. IV. However, the approximate results without smoothing supply the upper bounds of the MPE estimate's variance and confidence intervals (Fig. 6). The precise SD

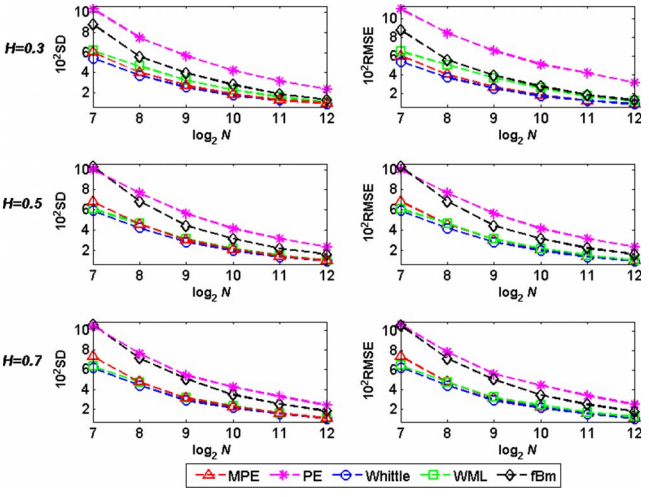

FIG. 5. (Color online) SD and RMSE of all estimators. For different H's, SDs, and RMSEs both decrease consistently as data length increases. We produced 1000 simulations using the Wood and Chan algorithm for each H. Point markers code for the estimators: MPE (△), PE (∗), Whittle (○), WML (□), and fBm (◇).

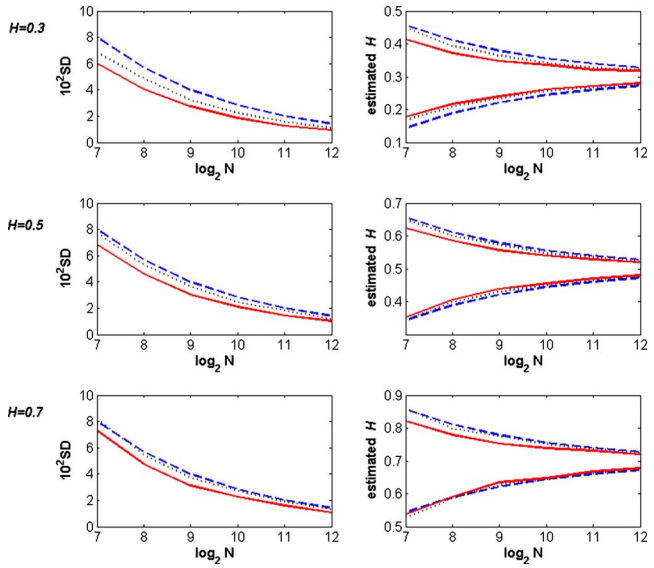

FIG. 6. (Color online) The SD and confidence intervals of theoretical calculation and Monte Carlo simulations on fGn's with different data lengths and H values. $\hat{H}$ and $\hat{H}_1$ denote the MPE estimate with and without smoothing, respectively. Left column: dashed blue lines indicate the approximate theoretical SD of $\hat{H}_1$; dotted black lines indicate the Monte Carlo SD of $\hat{H}_1$; solid red lines indicate the Monte Carlo SD of $\hat{H}$. Right column: dashed blue lines indicate the approximate theoretical confidence intervals of $\hat{H}_1$; dotted black lines indicate the Monte Carlo confidence intervals of $\hat{H}_1$; solid red lines indicate the Monte Carlo confidence intervals of $\hat{H}$. The confidence level was chosen as 0.95.

and confidence intervals should be investigated in future studies.

VI. ILLUSTRATIVE EXAMPLE

We applied the estimators to Nile river minima which were recorded as the yearly minimum water level in the Nile river from 662 to 1284 A.D. [11,44] (Fig. 7). These data are available from [45]. This data set played a key role in the discovery of long-range dependence in hydrological data by Hurst [46]. Statistical modeling of this time series was first

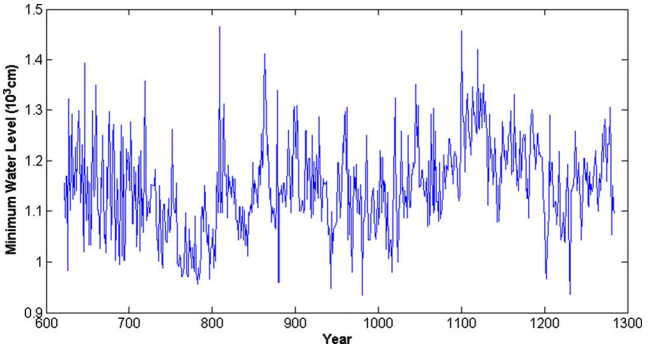

FIG. 7. (Color online) Nile river minimum water levels for 622–1284 A.D.

TABLE I. Estimators of H for Nile river minima.

Estimator	$\hat{H}$
MPE	0.85
PE	0.90
Whittle	0.84
WML	0.82
fBm	0.80

done as an fGn in the doctoral works of Mohr [47] and Graf [48]. Graf reported estimates of H between 0.83 and 0.85 [48]. Using an approximate maximum likelihood estimate, Beran reported estimates of $H=0.84$ for fGn with a 95% confidence interval of 0.79–0.89 [11]. He also established a goodness-of-fit test for the spectral density function of a long memory process and showed that fGn appears to fit the spectral density function of the Nile river series well.

The results are displayed in Table I. Obviously, comparing our results with the above analyses by Graf and Beran indicates that the performance of our proposed method is second only to Whittle's estimator and outperforms the other two estimators (Table I).

VII. DISCUSSION

Wavelet analysis is widely used in estimating the parameters of fGn primarily because of its multiresolution nature and its adaptability to local features in the data. Because the data length is always finite, a boundary effect [49] and an optimal scale range [40] always exist. If only optimal scales are used in the estimation process, as with the fBm-based estimator, the variance will increase. When all the scales are employed, the process may introduce bias into the estimation, as occurs with the WML estimator.

Whittle's estimator enjoys the same asymptotic properties as the exact MLE. According to the simulation studies in [12,13,50], Whittle's estimator is among the best estimators. Indeed, it also showed the best performance in this study (Figs. 3–5). However, the minimization procedure it employs requires many repetitive calculations, leading to a significantly higher overall cost than the other estimators.

In our proposed modified periodogram estimator, we generalized the approximation of SDF from a simple linear form $(1-2H)$ to a piecewise polynomial form $[\alpha(H)]$, which improved the previously poor approximations from the perspective of LSE. Moreover, the averaging procedure decreased the variance of the estimates. As a result, our modified periodogram performed much better than the original periodogram method.

Our simulations also suggest that it outperforms the popular WML estimator in terms of MSE and is comparable to Whittle's estimator. In addition, it is computationally simple and has a low numerical complexity. Consequently, it is much faster than Whittle's estimator.

The real data of Nile river minima were used to validate the estimators. Whittle's estimator gave the best performance

and the next was our proposed method, which was coincident with the results of simulations in Sec. V.

VIII. CONCLUSION

Based on the periodogram method, we proposed a modified estimator for the Hurst exponent of fGn in this study. The improved approximation of SDF from the perspective of LSE markedly reduced the bias. Moreover, the averaging procedure introduced significantly reduced the variance of the estimates. Thus, it yielded a rapid and efficient estimate of H in fGn with $0<H<1$.

ACKNOWLEDGMENTS

This work was partially supported by the Excellent SKL Project of Natural Science Foundation of China Grant No. 60723005, the Natural Science Foundation of China Grant No. 30900476, and the Natural Science Foundation of China Grant Nos. 60873088, 10631080, 60475042. We appreciate the assistance of Dr. Rhoda E. Perozzi and Dr. Edmund F. Perozzi.

APPENDIX: THREE OTHER ESTIMATORS

1. Whittle's estimator

The likelihood function of the fGn $G=(G_1,\ldots,G_N)^T$, with covariance Σ matrix depending on an unknown parameter vector $\theta:=(H,\sigma^2)\in(0,1)\times\mathbb{R}_+$, is

$$L(G;\theta) := (2\pi)^{-N/2}|\Sigma(\theta)|^{-1/2}e^{-(1/2)G^T\Sigma^{-1}(\theta)G}, \quad \text{(A1)}$$

where $|\Sigma(\theta)|$ denotes the determinant of the matrix $\Sigma(\theta)$. The maximum likelihood estimator (MLE) of θ is obtained by maximizing the logarithmic-likelihood function

$$\mathrm{LL}(G;\theta) := \ln L(G;\theta) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma(\theta)| - \frac{1}{2}G^T\Sigma^{-1}(\theta)G. \quad \text{(A2)}$$

The exact MLE is quite time consuming and unstable for practical estimations. An alternative approximating version to the MLE was proposed by Whittle [17] (see also [11,20,51]). The key idea is that $\ln|\Sigma(\theta)|$ is approximated by $N\int_{-1/2}^{1/2}\ln S(f)df$ and $G^T\Sigma^{-1}(\theta)G$ is approximated by $N\int_{-1/2}^{1/2}[\hat{S}^{(p)}(f)/S(f)]df$ in Eq. (A2).

2. WML estimator

Using wavelet decomposition on the fGn, G, we have the following likelihood function:

$$L(G;\theta) := (2\pi)^{-N/2}|\tilde{\Sigma}(\theta)|^{-1/2}e^{-(1/2)G_w^T\tilde{\Sigma}^{-1}(\theta)G_w}, \quad \text{(A3)}$$

where θ is defined as above, $G_w := (a_{-J,1},d_{-J,1},\ldots,d_{-1,1},\ldots,d_{-1,2^{J-1}})^T$, is the coefficients of the wavelet transform and J is the maximum level decomposition of the signal according to some wavelet. $\tilde{\Sigma}(\theta)$ is the covariance matrix of G_w, which is almost diagonal and can be approximated by the diagonal matrix [22]. A Daubechies wavelet with four vanishing moments (db4) is often used.

3. Estimator based on discrete variations of the fBm

As previously noted, an fGn process can be regarded as an incremental process of an fBm. Equally, if $G=(G_1,\ldots,G_N)$ is a finite sample of a fGn with the Hurst exponent, the variable $B=(B_1,\ldots,B_N)$,

$$B_t = \sum_{l=1}^{t} G_l. \quad \text{(A4)}$$

The convergence of the kth absolute moment of discrete variations is defined by

$$S^N(k,a) = \frac{1}{N-l}\sum_{i=l}^{N-1}|V^a(i/N)|^k \quad (k>0). \quad \text{(A5)}$$

The parameter a denotes a filter of length $l+1$ and with an order, $p\geq 1$, verifying $\Sigma_{q=0}^{l}a_q q^r=0$; for $r=0,\ldots,p-1$, V^a is derived as follows:

$$V^a\left(\frac{i}{N}\right) = \sum_{q=0}^{l} a_q B\left(\frac{i-q}{N}\right), \quad \forall \quad i\in\{l,\ldots,N-1\}. \quad \text{(A6)}$$

Assume M to be the upper bound of the scale used in the estimation and the sequence of filters $(a_m)_{1\leq m\leq M}$ are defined by

$$a_i^m = \begin{cases} a_j, & \text{if } i=jm \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } i=0,\ldots,ml+1 \quad \text{(A7)}$$

One immediately sees that

$$E(S_N(k,a^m)) = m^{Hk}E(S_N(k,a)). \quad \text{(A8)}$$

By estimating $E(S_N(k,a^m))$ by $S_N(k,a^m)$, an estimator of H can be deduced from a simple linear regression of $\{\ln S_N(k,a^m)\}_{1\leq m\leq M}$ on $\{k\ln m\}_{1\leq m\leq M}$ (for more details, see Refs. [19–21,40]). In particular, a Daubechies wavelet with four vanishing moments (db4) is usually chosen as the filter a and M is set to 5 empirically [40].

[1] C. Granger and R. Joyeux, J. Time Ser. Anal. **1**, 15 (1980).
[2] B. B. Mandelbrot and J. R. Wallis, Water Resour. Res. **5**, 321 (1969).
[3] M. S. Keshner, Proc. IEEE **70**, 212 (1982).
[4] S. Peleg, J. Naor, R. Hartley, and D. Avnir, IEEE Trans. Pattern Anal. Mach. Intell. **PAMI-6**, 518 (1984).
[5] T. Lundahl, W. J. Ohley, S. M. Kay, and R. Siffert, IEEE Trans. Med. Imaging **5**, 152 (1986).
[6] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[7] S. Roche, D. Bicout, E. Macia, and E. Kats, Phys. Rev. Lett. **91**, 228101 (2003).
[8] V. Maxim, L. Sendur, J. Fadili, J. Suckling, R. Gould, R. Howard, and E. Bullmore, Neuroimage **25**, 141 (2005).
[9] A. M. Wink, F. Bernard, R. Salvador, E. Bullmore, and J. Suckling, Neurobiol. Aging **27**, 1395 (2006).
[10] P. Embrechts and M. Maejima, *Selfsimilar Processes* (Princeton University Press, Princeton, NJ, 2002).
[11] J. Beran, *Statistics for Long-Memory Processes* (Chapman & Hall, New York, 1994).
[12] H. D. Jeong, J. S. Lee, D. McNickle, and K. Pawlikowski, Simul. Model. Pract. Theory **15**, 1173 (2007).
[13] M. S. Taqqu, V. Teverovsky, and W. Willinger, Fractals **3**, 785 (1995).
[14] T. Higuchi, Physica D **31**, 277 (1988).
[15] B. B. Mandelbrot and J. R. Wallis, Water Resour. Res. **5**, 242 (1969).
[16] J. Geweke and S. Porter-Hudak, J. Time Ser. Anal. **4**, 221 (1983).
[17] P. Whittle, Ark. Mat. **2**, 423 (1953).
[18] G. W. Wornell and A. V. Oppenheim, IEEE Trans. Signal Process. **40**, 611 (1992).
[19] J. Istas and G. Lang, Ann. I.H.P. Probab. Stat. **33**, 407 (1997).
[20] J. F. Coeurjolly, J. Stat. Software **5**, 1 (2000).
[21] J. T. Kent and A. T. A. Wood, J. R. Stat. Soc. Ser. B (Methodol.) **59**, 679 (1997).
[22] E. McCoy and A. Walden, J. Comput. Graph. Stat. **5**, 26 (1996).
[23] P. Abry and D. Veitch, IEEE Trans. Inf. Theory **44**, 2 (1998).
[24] B. Audit, E. Bacry, J. F. Muzy, and A. Arneodo, IEEE Trans. Inf. Theory **48**, 2938 (2002).
[25] H. Kettani and J. Gubner, *Proceedings of the 27th Annual IEEE Conference on Local Computer Networks* (IEEE Computer Society, Washington, D.C., 2002), p. 160.
[26] J. Mielniczuk and P. Wojdyłło, Comput. Stat. Data Anal. **51**, 4510 (2007).
[27] B. B. Mandelbrot and J. W. Van Ness, SIAM Rev. **10**, 422 (1968).
[28] Y. Sinai, Theory Probab. Appl. **21**, 64 (1976).
[29] C. Hurvich, R. Deo, and J. Brodsky, J. Time Ser. Anal. **19**, 19 (1998).
[30] P. Robinson, Ann. Stat. **23**, 1048 (1995).
[31] J. Davidson and P. Sibbertsen, Econ. Lett. **102**, 83 (2009).
[32] B. Pilgram and D. T. Kaplan, Physica D **114**, 108 (1998).
[33] P. Flandrin, IEEE Trans. Inf. Theory **35**, 197 (1989).
[34] G. M. Raymond, D. B. Percival, and J. B. Bassingthwaighte, Physica A **322**, 169 (2003).
[35] K. Shimotsu and P. C. B. Phillips, J. Time Ser. Anal. **23**, 57 (2002).
[36] D. W. K. Andrews and P. Guggenberger, Econometrica **71**, 675 (2003).
[37] E. Moulines and P. Soulier, Ann. Stat. **27**, 1415 (1999).
[38] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques* (Cambridge University Press, Cambridge, England, 1993).
[39] G. Wornell, IEEE Trans. Inf. Theory **36**, 859 (1990).
[40] J. Coeurjolly, Stat. Inference Stoch. Process. **4**, 199 (2001).
[41] J. Bardet, G. Lang, G. Oppenheim, A. Philippe, and M. Taqqu, in *Theory and Applications of Long-Range Dependence*, edited by P. Doukhan, M. Taqqu, and G. Oppenheim (MA-:Birkhäuser, Boston, MA, 2003), p. 579.
[42] R. B. Davies and D. S. Harte, Biometrika **74**, 95 (1987).
[43] A. Wood and G. Chan, J. Comput. Graph. Stat. **3**, 409 (1994).
[44] O. Toussoun, *C. i. d. Géographie: Mémoire sur l'Histoire du Nil* (Imprimerie de l'Institut fran ais d'Archéologie Orientale, Cairo, 1925).
[45] http://lib.stat.cmu.edu/S/beran.
[46] H. E. Hurst, Trans. ASCE **116**, 770 (1951).
[47] D. Mohr, Ph.D. thesis, Princeton University, 1981.
[48] H. Graf, Ph.D. thesis, Swiss Federal Institute of Technology Zurich, Zürich, 1983.
[49] E. McCoy, Ph.D. dissertation, Imperial College, 1994.
[50] W. Rea, L. Oxley, M. Reale, and J. Brown, e-print arXiv:0901.0762v1.
[51] W. Palma, *Long-Memory Time Series: Theory and Methods* (Wiley-Interscience, Hoboken, New Jersey, 2007).