Research Article
The Optimal Bandwidth Parameter Selection in GPH Estimation

Weijie Zhou,1 Huihui Tao,1,2 Feifei Wang,3 and Weiqiang Pan2

1School of Wujinglian Economics, Changzhou University, Changzhou 213159, Jiangsu, China
2School of Business, Changzhou University, Changzhou 213164, Jiangsu, China
3Qu Qiubai School of Government, Changzhou University, Changzhou 213159, Jiangsu, China

Correspondence should be addressed to Weijie Zhou; wjzhou@cczu.edu.cn

Received 22 August 2021; Revised 3 November 2021; Accepted 11 November 2021; Published 3 December 2021

1. Introduction

Long-term memory widely exists in the fields of biology, medicine, geology, hydrology, climate, and social science fields [1–3]. It refers to the fact that observations depend on each other in a long term and the autocorrelation function of a sequence decays slowly. In a system with long memory, some important historical events influence the future in long time spans, which contribute to the formation of long memory. For example, it is shown that the big price rise and fall in stock markets and extreme high and low temperatures have an impact on the long memory of their corresponding sequences [4, 5]. According to the relationship between long memory and approximate entropy revealed by Pincus and Kalman, the stronger the long memory of the sequence is, the better its predictability will be [6]. In addition, Baillie found that if a time series has a long memory, it is difficult to characterize the internal structural features with short memory models, such as ARMA model. Additionally, the simulative and predictive accuracy of those models are comparatively low [7]. Therefore, the research on long memory is of great importance for the theoretical and practical applications [8, 9]. There are two types of long memory in time series. One is the persistent long memory, which means that the development trend of time series will keep in line with the current movement direction in future. The corresponding long memory parameter is $0 < d < 0.5$. Contrary to the persistent long memory, the other is anti-persistent long memory, indicating that the motion in the future will be opposite to that of present state, and its long memory parameter is $0 < d < 0.5$. It is generally believed that the British hydrologist Hurst was the first to study the long memory characteristics in a system. He used Hurst index ($H$) to depict the long memory strength of a time series [10]. The relationship between Hurst index and long memory parameters is $H = 0.5 + d$ [11, 12]. When $H \rightarrow 1$ (or $d \rightarrow 0.5$), it implies that persistent long memory of a time series is strong. When $H \rightarrow 0$ (or $d \rightarrow -0.5$), it indicates that the antipersistent long memory of a time series is strong. When $H \rightarrow 0.5$ (or $d \rightarrow 0$), the time series appears as random walk, suggesting its unpredictability in theory.

So far, there are no fewer than ten methods for calculating long memory parameters, which can be roughly divided into three categories. The first algorithm is estimated in the time domain, such as the Aggregated Variance, Differencing the Variance [13], Higuchi [14], R/S Analysis,
and Detrended Fluctuation Analysis (DFA) [15]. The second algorithm is the frequency-domain estimation, such as Whittle and Averaged Periodogram Estimation [16, 17]. The third algorithm is the wavelet-domain estimation methods, such as Wavelet Maximum Estimator and Wavelet-Based Estimation [18, 19]. In addition, based on the above algorithms, the researchers proposed many improved estimation methods, such as modified Rescaled Range [20], exact local Whittle method, and modified local Whittle estimation [21, 22]. However, as far as the time-domain estimation method is concerned, it is difficult to judge the significance of long memory since the statistic distribution of long memory parameters cannot be given. For the wavelet-domain estimation algorithm, the requirements of structural features in a sequence are often too harsh to correctly extract the modulus, which sometimes makes the results different from qualitative analysis. For the full-parameter estimation method in the frequency domain, it requires that the perturbed random item has a Gaussian distribution, such as Whittle estimation, and involves integral operation, which is difficult to meet in practice. For example, it is well known that the distribution curve of return time series in stock markets possesses a sharp peak and heavy tail feature [23, 24]. If the estimation methods in the time and the wavelet domains are regarded as a nonparametric method, the research of the semiparametric method is gradually developed as a compromise between full-parameter and nonparametric method in the frequency domain. Different from the previous method, the GPH proposed by Geweke and Porter-Hudak [25] has better advantages in semiparametric estimation, such as it reduces the normality requirement of random items in estimation and its statistical distribution of estimators is provided within a certain range. Based on the framework of the GPH method, several improved algorithms to estimate long memory of sequences were proposed [26, 27], which extend the application of GPH method to different long time memory sequences, simplify the concept, and improve the computational speed. Robinson establishes the asymptotic normality of the GPH estimator, and the results show that it is suitable for stationary and reversible Gaussian vector sequences [28]. Hurvich et al. established the asymptotic properties of the GPH estimation method and derived the expressions of asymptotic bias, variance, and mean square error of estimators, effectively evaluating the accuracy of asymptotic theory for the mean square error of finite sample size [29]. On this basis, Velasco generalized Robinson’s results, showing that, with sufficient data cones, the revised estimates of any $d$ (including nonstationary and irreversible processes) are consistent and asymptotically normally distributed [30]. In addition, Velasco proved the consistency of the logarithmic periodic graph regression estimates of the long memory parameters of the series when studying the long-range dependent linear time series and obtained the asymptotic distribution of the asymptotic periodic graph estimates of the long-range dependent time series under possibly non-Gaussian observations [31]. However, on the other hand, the new method requires programming in the operation, while the GPH algorithm has been implemented with menu-based operation on some metrology software, such as OX software and R software. Therefore, the GPH method is an indispensable method for estimating long memory in terms of the universality and maneuverability of application. However, there are three problems when using GPH to test the long memory of time series (financial data) [34–37]. Firstly, for the parameter $a$ of bandwidth $g(N) = N^a$ ($N$ is the sequence length), most of the studies mainly choose 0.5, 0.6, 0.7, and 0.8 or directly select $g(N) = N/2$ to estimate the long memory, which is formed by subjectivity of the authors. Secondly, we have no clear understanding of the action mechanism of how the bandwidth parameter $a$ influences the existence of long memory, persistence or antipersistence of long memory, and the accuracy of the estimated parameter $d$ of long memory, which are calculated by utilizing the GPH algorithm. Thirdly, as Jeong et al. [38] proposed, there is a common problem with GPH and other methods to estimate long-term memory of sequences, that is, few authors conducted a simulation analysis close to the actual sequence to test the accuracy of parameter $d$.

In this paper, based on the ARFIMA $(1,d,1)$ process and some typical features of financial time series, we use the Monte Carlo method to test the impact of parameter $a$ on the existence of long memory, persistence or antipersistence of long memory, and estimation accuracy of the long memory parameter $d$, so as to give the optimal bandwidth in the GPH algorithm.

The structure of this paper is arranged as follows: Section 2 is the introduction of the GPH method. Section 3 gives the Monte Carlo simulation method and validation rules. Section 4 is the analysis of simulation results. The conclusion is summarized in Section 5.

2. The GPH Semiparametric Method of Long Memory Estimation

Many scholars’ studies [39] show that the basis of GPH semiparametric method is that the data process is a fractional white noise process. Therefore, the fractional white noise process $x_t$ satisfies $(1 - L)^d x_t = u_t$, where $u_t$ is a stationary process. If the $f_u(\lambda)$ is the spectral density function of $u_t$, then the spectral density function $f_x(\lambda)$ of $x_t$ can be expressed as

$$f_x(\lambda) = \left(4 \sin^2 \left(\frac{\lambda}{2}\right)\right)^{-d} f_u(\lambda),$$  \hspace{1cm} (1)

Discretizing the logarithmic form of equation (1),

$$\ln\{f_x(\lambda_j)\} = \ln\{f_u(0)\} - d \ln\{4 \sin^2(\lambda_j/2)\}$$
+ \ln\{f_u(\lambda_j)/f_u(0)\}, \hspace{1cm} (2)

where $\lambda_j = 2\pi j/N$, $j = 1, 2, \ldots, g(N)$, and $g(N) = N^a$. $N$ is the length of the sequence $\{x_t\}$, and $\lambda$ is called the harmonic frequency of the sample data. Geweke and Porter-Hudak proved that the last term $\ln\{f_u(\lambda)/f_u(0)\}$ in equation (2) is negligible or close to a constant in the sufficiently small
harmonic frequency coordinates. Therefore, the Ordinary Least Squares (OLS) algorithm can be performed on equation (2) to estimate the long memory parameter $d$.

Besides, when $d<0$, Geweke and Porter-Hudak illustrated that the estimator $\hat{d}$ of equation (2) has an approximate distribution:

$$
\hat{d} \sim N\left[ d, \frac{\pi^2}{6\sum_{j=1}^{g(N)} (z(j) - \bar{z}(j))^2} \right],
$$

where $z(j) = \ln\{4\sin^2(\lambda_j/2)\}$ and $\bar{z}(j) = \sum_{j=1}^{g(N)} z(j)/g(N)$. For $d \geq 0$, equation (3) is verified empirically, and its theoretical proof is still an open question. However, for actual sequences, it is difficult to know the true value of $d$ in advance. As the $d$ is estimated by equation (2), the existence of long memory can be validated by judging whether the estimator $\hat{d}$ is significantly different from 0. Ignoring the $\ln\{f_u(\lambda_j)/f_u(0)\}$ part, equation (2) is transformed as follows:

$$
\ln\{f_u(\lambda_j)\} = \ln\{f_u(0)\} - d \ln\{4\sin^2(\lambda_j/2)\} + \ln\left\{ \frac{I(\lambda_j)}{f_u(\lambda_j)} \right\}.
$$

(4)

where $\ln\{f_u(\lambda_j)\}$ is the period gram, i.e., the square of the magnitude of the spectral density function. Porter-Hudak proved that $\ln\{f_u(\lambda_j)/f_u(0)\}$ obeyed the Gumbel distribution with a negative Euler constant, $-0.57721$, mean, and $\pi^2/6$ variance. Hence, equation (3) is further simplified to

$$
\ln\{f_u(\lambda_j)\} = \alpha - d \ln\{4\sin^2(\lambda_j/2)\} + e_j,
$$

(5)

where $\alpha = \ln\{f_u(0)\} + 0.57721$ and $e_j \sim N(0, \pi^2/6)$ under the large sample scenario. The $\hat{d}$ can be estimated by equation (5). Testing the existence of long memory in sequence can be judged as follows:

$$
H_0: d = 0; H_1: d \neq 0
$$

$$
T = \frac{-\hat{d}}{S(-\hat{d})} \sim t\{g(N) - 2\},
$$

(6)

where $S(-\hat{d}) = \sqrt{\pi^2/6\sum_{j=1}^{g(N)} (w(j) - \bar{w}(j))^2}$, $w(j) = \ln\{4\sin^2(\lambda_j/2)\}$, and $\bar{w}(j) = \sum_{j=1}^{g(N)} w(j)/g(N)$. When the sample size is large, the $t$ distribution approximates normal distribution and the statistical test of the estimator $\hat{d}$ by equation (6) can be approximately equal to that of equation (3). Setting a confidence level $\alpha$, we can check the existence of the long memory parameter $d$. In this paper, taking into account the robust characteristics of the GPH algorithm in estimating the long memory, we will let $\alpha = 0.1$. For a large sample, Agiakloglou et al. and Sowell mentioned that equation (5) can still estimate the long memory of the sequence, even if there are short-term components in the sequence, such as the ARFIMA process [39, 40]. In addition, Geweke and Porter-Hudak demonstrated the relationship between long memory parameter $d$ and the Hurst exponent $H$ by the structured method, i.e., $d = H - 0.5$.

3. Simulation Method and Validation Rules

3.1. Simulation Method. In empirical financial research, it is generally believed that the first-order model can adequately depict autocorrelation and fluctuation in financial time series [41]. Combined with the typical features of financial time series, such as sharp peak, heavy tail, asymmetric distribution, and long memory, this paper uses ARFIMA(1, d, 1) model with Skew Student’s t Distribution (SKST) to generate simulation data close to the actual sequence, so as to test the impact of bandwidth parameter $a$ of the GPH algorithm on long memory estimation. The ARFIMA (1, d, 1) model is expressed as

$$
(1 - \phi L)(1 - L)^9 y_t = (1 + \theta L)\epsilon_t,
$$

$$
\epsilon_t \sim \text{SKST}(x; \lambda, \nu),
$$

$$
\text{SKST}(x; \lambda, \nu) = \begin{cases} 
bc^{-\nu \frac{(x - a)^2}{2}} & x \leq \frac{-a}{\nu} \\
bc^{\frac{-\nu (x + a)^2}{2}} & x > \frac{-a}{\nu}
\end{cases}
$$

(7)

where $\lambda = 4\lambda c (\nu - 2/\nu - 1)$, $\nu = \Gamma(\nu + 1/2)\Gamma(\nu/2) \sqrt{\pi (\nu + 2)}$, $\lambda$ and $\nu$ are the skewness coefficient and the freedom degree of the biased student’s $t$ distribution. We randomly select $\lambda$ from (-3, 3) and set $\nu = 4$ in this paper. $\phi$ and $\theta$ are the autoregressive coefficient (AR) and the moving average coefficient (MA), respectively. It is found that most of the autoregressive and moving average coefficients in financial time series models with first order are between -1 and 1. Thus, $\phi$ and $\theta$ are taken from (-1, 1) randomly. We generate nine types of data with long memory parameters $d = -0.4, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3$, and 0.4 by equation (7). Given that the GPH algorithm is subject to sequence length in estimating long memory, we generate 5000 sequences with length $N = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 5000, 10000, 50000,$ and 50000 for each long memory parameter. Hence, there are $9 \times 14 \times 5000 = 630000$ sequences in total. Figure 1 shows one simulated sequence and its probability distribution. It can be seen that the simulated data has a sharp peak, a heavy tail, and asymmetric characteristics. When applying the GPH algorithm to estimate sequence having long memory, most of the literature suggests bandwidth parameters $0.2 < a < 0.5$, $0.5 < a < 0.5$, and $N^a < N/2$ [42, 43]. In order to fully understand the influence of different bandwidth parameters $a$ on the GPH estimation under different long memory parameters $d$ and sequence lengths, this paper takes $0.2 < a \leq \sup\{a: N^a \leq N/2\}$ and the discretization step size is taken as 0.01, where $\sup\{a: N^a \leq N/2\} = 0.9$. 


3.2. Validation Rules. It is mainly from three aspects to analyze the impact of the bandwidth parameter $a$ on testing of the long memory in sequences, including the existence of long memory, the persistence and antipersistence of long memory, and the calculation accuracy of long memory parameters with different sequence lengths. The validation of the first two aspects should be a progressive relationship. The long memory existence of a sequence is firstly tested. If a sequence has a long memory, we judge the persistence or antipersistence of the long memory. However, based on our simulation results, it is found that if we analyze the effect of bandwidth parameter $a$ on examining long memory as described above, a lot of useful information would be lost. The optimal range of the parameter $a$ derived from the existence judgment of long memory may be a very small interval, or even only a point. Hence, it is difficult to fully investigate the impact of different bandwidth parameters $a$ on the persistence or antipersistence judgment of long memory and estimation accuracy of long memory parameter $d$, which is not conducive to finding out the optimal bandwidth parameter $a$. To this end, we set the following rules to select the optimal bandwidth parameter $a$.

Rule 1. Based on Monte Carlo simulation and the GPH algorithm, the optimal bandwidth parameter set on the existence test, persistence or antipersistence judgment of long memory, and estimation accuracy of long memory parameter $d$ are recorded as $A_1$, $A_2$, and $A_3$.

Rule 2. According to the different requirements on testing long memory of sequences by the GPH algorithm, three related subsets of optimal parameters $a$ are constructed, that is, $A_1$, $A_1 \cap A_2$, and $A_1 \cap A_2 \cap A_3$. $A_1 \cap A_2$ denotes the optimal parameter set which satisfies the existence test and persistence or antipersistence judgment of long memory synchronously. $A_1 \cap A_2 \cap A_3$ represents the optimal parameter $a$ set which satisfies the existence test, persistence or antipersistence judgment of long memory, and estimation accuracy of the long memory parameter. Based on the above definition, the judging accuracy for the existence test under parameter $a$ belonging to $A_1$ set is higher than other sets. The same meaning is suitable for $A_1 \cap A_2$ and $A_1 \cap A_2 \cap A_3$ sets.

Rule 3. Given a time series, we assume that the probability of its long memory and no long memory is equal, i.e., 0.5. And, for time series with long memory, the probability of different long memory parameters $d$ is equal.

3.2.1. The Existence Test of Long Memory. When $d \neq 0$, we set $P_{xzd}(d,a) = 1/5000 \text{Num}_{T_{d,a} > t_{a2}(g(N) - 2)}$. Additionally, when $d = 0$, we assume that $P_{xzd}(0,a) = 1/5000 \text{Num}_{T_{d,a} < t_{a2}(g(N) - 2)}$. $P_{xzc}(a) = 1/8 \sum_{a \neq 0} P_{xzd}(d,a)$, $P_{xzf}(a) = P_{xzd}(0,a)$, and $P_{xzzl}(a) = P_{xzc}(a) + P_{xzf}(a)/2$. $T_{d,a}$ is the $t$ statistic of $d$ under the bandwidth parameter $a$. $\text{Num}_{T_{d,a} > t_{a2}(g(N) - 2)}$ represents the number of estimated values $d$ rejecting the null hypothesis under the bandwidth parameter $a$, which is equivalent to the number of $d \neq 0$; similarly, $\text{Num}_{T_{d,a} < t_{a2}(g(N) - 2)}$ represents the number of estimated values $d$ that do not reject the null hypothesis under the bandwidth parameter $a$, which is equivalent to the number of $d = 0$. $P_{xzd}(d,a)$ is the judging accuracy for the existence test of long memory using the GPH algorithm under bandwidth parameter $a$ and long memory parameter $d$. Obviously, the closer the $P_{xzd}(d,a)$ approaches to 1, the more accurate the GPH algorithm is. $P_{xzd}(0,a)$ is employed for measuring the accuracy of the existence test on sequences with no long memory. In order to comprehensively judge the ability of the GPH algorithm to estimate sequences with different long memory, we give the mean of $P_{xzd}(d,a)$ ($d \neq 0$), i.e., $P_{xzc}(a)$. As $d = 0$, we set $P_{xzf}(a) = P_{xzd}(0,a)$ for comparative analysis. In the actual analysis, for a sequence $\{x_t\}$, it is impossible to know whether it has long memory in advance. Therefore, the $P_{xzzl}(a)$ is constructed to test whether the long memory of
sequence exists or not under different bandwidth parameters $a$.

3.2.2. The Persistence or Antipersistence Judgment of Long Memory. As $d < 0$ and $d > 0$ denote the persistence and antipersistence of long memory parameter, respectively, we construct the judging accuracies $Pzc(a)$ and $Pfc(a)$ to study the impact of the GPH algorithm with different parameters $a$ on the long memory test.

Setting $Pc(d, a) = \text{Num}(d, a)/5000$,

$$Pzc(a) = \frac{1}{4} \{Pc(0.1, a) + Pc(0.2, a) + Pc(0.3, a) + Pc(0.4, a)\},$$

$$Pfc(a) = \frac{1}{4} \{Pc(-0.4, a) + Pc(-0.3, a) + Pc(-0.2, a) + Pc(-0.1, a)\},$$

$$Pcszl(a) = \frac{Pzc(a) + Pfc(a)}{2},$$

where $\text{Num}(d, a)$ denotes the number of estimated $\tilde{d}$ and actual $d$ with the same plus or minus sign under the bandwidth parameter $a$. $Pc(d, a)$ is the judging accuracy of persistence or antipersistence for long memory under long memory parameter $d$ and bandwidth parameter $a$. $Pzc(a)$ and $Pfc(a)$ are judging accuracy of persistence and antipersistence long memory, respectively. $Pcszl(a)$ is the comprehensive judging accuracy of persistence and antipersistence long memory.

3.2.3. The Estimation Accuracy of the Long Memory Parameter. According to the simulation results, if the error rate $1/5000 \sum_{j=1}^{5000} |d - d_j|/|d|$ is used to verify the precision of the GPH algorithm, it is found that there are several orders of magnitude between big and small $|d|$. It is not conducive to finding out the optimal range of bandwidth parameter $a$. In this paper, some rules are made as follows. If the estimated $\tilde{d}$ falls within neighborhood of the truth value $d$, i.e., $U(d, \delta)$, it is considered that the estimated $\tilde{d}$ by the GPH algorithm is valid under the accuracy $\delta$. A basic selection principle of $\delta$ is that the neighborhood $U(d, \delta)$ of different parameters $d$ does not overlap with each other. The estimation efficiency of the GPH algorithm under long memory parameter $d$ and bandwidth parameter $a$ is defined as follows:

$$Pyx(d, a) = \frac{\text{Num}(d - \delta \leq \tilde{d} < d + \delta)}{5000}$$

$$Pyxm(a) = \frac{1}{9} \sum Pyx(d, a).$$

$\text{Num}(d - \delta \leq \tilde{d} < d + \delta)$ refers to the number of the estimated parameter $\tilde{d}$ falling in the neighborhood $U(d, \delta)$. $Pyxm(a)$ is the average estimation efficiency of the GPH algorithm with different long memory parameters. The larger the $Pyxm(a)$ is, the higher the estimation efficiency of the GPH algorithm under the estimation accuracy $\delta$ is. Generally, the smaller the choice of $\delta$, the farther the distance between neighborhoods $U(d, \delta)$ of different parameters $d$ and the fewer the number of estimated $\tilde{d}$ falling into $U(d, \delta)$, so that the discrimination degree of $Pyx(d, a)$ may decline, which is not conducive to finding out the optimal range of bandwidth parameter $a$. According to the long memory parameter $d$ in Monte Carlos simulation, we set $\delta = 0.05$ in this paper.

4. Result Analysis

4.1. The Existence Judgment of Long Memory. As can be seen in Figure 2, the judging accuracy of the existence test on long memory increases gradually in the bandwidth parameter $[0.2, \sup[a: N^a \leq N/2]]$, while the judging accuracy on that of no long memory decreases gradually. To be more specific, within the range $[0.2, 0.4]$ of the bandwidth parameter $a$, no matter how long the sequence is (simulated length in this paper), if the sequence has a long memory, the judging accuracy is less than 0.3. When $a \rightarrow 0.2$, the judging accuracy is only 0.1. However, if a sequence has no long memory, the judging accuracy of the GPH algorithm can approximately reach 0.9. It is impossible to know whether the long memory of the time series exists in advance, so it is difficult to distinguish the long memory from the no long memory reasonably by the GPH algorithm with bandwidth parameter $a \in [0.2, 0.4]$. In the range $[0.8, \sup[a: N^a \leq N/2]]$ of bandwidth parameter $a$, for sequences with different lengths, if the sequence has long memory, the judging accuracy of the GPH algorithm is about 0.9. For sequences with no long memory, the judging accuracy is less than 0.2. Hence, with bandwidth parameter $a \in [0.8, \sup[a: N^a \leq N/2]]$, it is not suitable to distinguish the long memory from no long memory. Further analysis shows that it is not suitable to use the GPH method to estimate the existence of long memory in time series on the bandwidth parameter $a \in [0.4, 0.5] \cup [0.7, 0.8]$. According to the judging accuracy curves of the existence test on long memory and no long memory, the intersection of two curves is $a = 0.6$. And, at the left side of the point, the judging accuracy of the no long memory is low, and on the right side of the point, the judging accuracy of the long memory is also low. Therefore, the judging accuracy of the no long memory and the long memory reaches a relatively high level at this point, which is beneficial to estimating the long memory by the GPH algorithm. In addition, the longer the time series length is, the higher the judging accuracy corresponding to
the point is. In order to find out the optimal range of bandwidth parameter \( a \), we plot the judging accuracy curves \( P_{xzzl}(a) \) in Figure 3. When the sequence length is 2000 or more, it is seen that the \( P_{xzzl}(a) \) is more than 0.75 around \( a \approx 0.6 \).

4.2. The Persistence or Antipersistence Judgment of Long Memory. As seen in Figure 4, for time series with different lengths, the judging accuracy curves of persistence or antipersistence for long memory present the same parabola shape. Within the bandwidth parameter \( a \in [0.5, 0.7] \), the judging accuracy reaches a relatively high value. Figure 5 gives the comprehensive judging accuracy \( P_{cszl}(a) \) of persistence and antipersistence. Within the bandwidth parameter \( a \in [0.5, 0.7] \), it is seen that the judging accuracy increases gradually with the increase in the time series length, and when the length of the sequence is above 2000, the judging accuracy \( P_{cszl}(a) \) is over 0.9.

4.3. Estimation of Long Memory Parameters. In Figure 6, for the time series with different lengths and long memory parameters \( d \), the estimation accuracy curves with long memory parameters \( d \) exhibit a similar shape. Within the bandwidth parameter \( a \in [0.55, 0.7] \), the accuracy reaches a relatively high value. In Figure 7, the average estimation accuracy \( P_{yxm}(a) \) of the GPH algorithm is given. Within the bandwidth parameter \( a \in [0.55, 0.7] \), it can be seen that with the increase in the time series length, the average estimation accuracy increases gradually under \( \delta = 0.05 \), which indicates that the probability of the estimated value \( \hat{d} \) falling into the neighborhood \( U(d, \delta) \) increases with the increase in the sequence length and the GPH algorithm is effective.
In order to make the optimal parameters \( a \) suitable for three branches of the long memory test together, ten bandwidth parameters \( a \) corresponding to the high judging accuracy of the GPH estimation under different sequence lengths are recorded as the optimal bandwidth parameter range, as seen in Table 1.

According to Table 1, we use intersection which refers to the common part of different sets to find out the optimal bandwidth parameter \( a \) ranges in several scenarios. Without considering the sequence length, we can select \([0.59, 0.62]\) as the optimal bandwidth range of the GPH algorithm for existence test and persistence or antipersistence judgment of...
long memory, while the optimal bandwidth parameters of the estimation accuracy on long memory parameter $d$ belong to $[0.6, 0.67]$. By taking the intersection of the optimal bandwidth parameters for the existence test, persistence or antipersistence judgment of long memory, and the estimation accuracy of long memory parameter, the following conclusions can be drawn:

① If the GPH algorithm is only needed to estimate the existence test of long memory, the desirable range of bandwidth parameters is $[0.59, 0.62]$. 

![Figure 6: Estimation accuracy of the long memory parameter.](image1)

![Figure 7: Average estimation accuracy of long memory parameter $d$.](image2)
If the existence test and persistence or antipersistence judgment of long memory are estimated synchronously, the desirable range is [0.59, 0.61].

In addition to the above two aspects of long memory, if the estimation accuracy of the long memory parameter is further required, the desirable bandwidth parameter range can vary in [0.6, 0.61]. Taking into account the sequence length, we can apply the desirable bandwidth parameters range of [0.58, 0.61] to the existence test, persistence or antipersistence judgment of long memory, and the estimation accuracy of long memory parameter with sequence length below 1000. Given the operational convenience, a = 0.6 is recommended as the optimal bandwidth parameter for estimating the long memory by the GPH algorithm.

Table 2 provides the calculation accuracy of the GPH algorithm for estimating long memory with the bandwidth parameter a = 0.6. \( P(A_1) \) \((A_1 = 0.6)\) denotes the calculation accuracy of the GPH algorithm for the existence test of long memory, which is equal to \( P_{\text{czzl}}(a) \). \( P(A_1 \cap A_2) \) \((A_1 \cap A_2 = 0.6)\) is the calculation accuracy for satisfying the existence test and persistence or antipersistence judgment of long memory synchronously.

The calculation step of \( P(A_1 \cap A_2) \) is to take the ratio of the minimum number among the existence test and persistence or antipersistence judgment of long memory to the total simulation number, which is similar to the calculation of \( P_{\text{czzl}}(a) \), i.e.,

\[
P_{\text{Ac}}(d,a) = \min\{\text{Num}\{|T_{d,a}| > t_{a/2}(g(N)-2), \text{Num}(d,a)\}\}/5000
\]

\[
P_{\text{Acz}}(a) = \frac{1}{4}\{P_{\text{Ac}}(0.1,a) + P_{\text{Ac}}(0.2,a) + P_{\text{Ac}}(0.3,a) + P_{\text{Ac}}(0.4,a)\}
\]

\[
P_{\text{Afz}}(a) = \frac{1}{4}\{P_{\text{Ac}}(-0.4,a) + P_{\text{Ac}}(-0.3,a) + P_{\text{Ac}}(-0.2,a) + P_{\text{Ac}}(-0.1,a)\}
\]

\[
P(A_1 \cap A_2) = \frac{P_{\text{Acz}}(a) + P_{\text{Afz}}(a)}{2}
\]


\[
P(A_1 \cap A_2 \cap A_3) \quad (A_1 \cap A_2 \cap A_3 = 0.6) \text{ is the calculation accuracy for satisfying the existence test, persistence or antipersistence judgment of long memory, and the estimation accuracy of long memory parameter synchronously. The calculation process of } P(A_1 \cap A_2 \cap A_3) \text{ is}
\]

\[
P_{AYx}(d, a) = \frac{\min\{\text{Num}[|T_{d,a} > t_{\alpha/2} (g(N) - 2)|], \text{Num}(d, a), \text{Num}(d - \delta < \hat{d} < d + \delta)\}}{5000}
\]

\[
P_{AYx}(0, a) = \frac{\min\{\text{Num}[|T_{d,a} < t_{\alpha/2} (g(N) - 2)|], \text{Num}(0 - \delta < \hat{d} < 0 + \delta)\}}{5000}
\]

\[
P(A_1 \cap A_2 \cap A_3) = \frac{1}{9} \sum P_{AYx}(d, a).
\]

As seen from Table 2, with the increase in the sequence length, the \( P(A_1) \), \( P(A_1 \cap A_2) \), and \( P(A_1 \cap A_2 \cap A_3) \) are gradually enlarged. When the sequence length is more than 700, \( P(A_1) \) exceeds 0.7, and \( P(A_1 \cap A_2) \) is over 0.7 when the sequence length is more than 1000. However, for \( P(A_1 \cap A_2 \cap A_3) \), when the sequence length is 5000, its value is only 0.4822, which is mainly caused by the poor accuracy of the GPH algorithm in estimating the long memory parameter. In Figure 7, when the sequence length is short, such as 300, the judging accuracy is only 0.2 under the estimation accuracy \( \delta = 0.05 \), implying that about 80% of estimated parameter \( d \) falls outside the neighborhood of true value \( d \). The result is consistent with the conclusion in [26]. Therefore, the GPH algorithm has certain defects when estimating long memory parameters. Only when the sequence length is over 10000, the estimation result is effective.

5. Conclusions

In this paper, we use the Monte Carlo simulation method to generate long memory sequences with different lengths by using the ARFIMA \((1, d, 1)\) process, so as to study the impact of the GPH algorithm on existence test, persistence or antipersistence judgment of long memory, and the estimation accuracy of long memory parameter. Within the bandwidth parameter \( a \in [0.5, 0.7] \), for the time series with different lengths, the judging accuracy of the GPH algorithm for the existence test, persistence or antipersistence judgment of long memory, and the estimation accuracy of long memory parameter all reaches a relatively high level. \( a = 0.6 \) can be selected as the optimal bandwidth parameter in application. With the length of time series increasing from 100 to 50000, the accuracy rate of the GPH algorithm for estimating the existence test of long memory increases from 0.5612 to 0.8786. The calculation accuracy of the GPH algorithm for persistence or antipersistence judgment of long memory is from 0.4697 to 0.8673. The calculation accuracy for satisfying the existence test and persistence or antipersistence judgment of long memory is from 0.0623 to 0.6624. The rules used in the analysis of long memory estimation by the GPH algorithm are gradually discussed from the experimental results. It is a practical and novel method, which can be used as a reference for other methods in testing the long memory.

Data Availability

The data used in this study are available on request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant nos. 71701024 and 82073339) and the National Social Science Fund (nos. 208ZD128, 20CRK018, 20CTJ015, and 21BJY007).

References

[1] X. Yu, R. Ghasemizadeh, I. Y. Padilla, D. Kaeli, and A. Alshawabkeh, “Patterns of temporal scaling of groundwater level fluctuation,” Journal of Hydrology, vol. 536, pp. 485–495, 2016.
[2] Z. Q. Jiang, W. J. Xie, and W. X. Zhou, “Multifractal analysis of financial markets,” Reports on Progress in Physics, vol. 82, no. 12, pp. 1361–6633, 2019.
[3] W. Zhou, Z. Wang, and H. Guo, “Modelling volatility recurrence intervals in the Chinese commodity futures market,” Physica A: Statistical Mechanics and its Applications, vol. 457, pp. 514–525, 2016.
[4] D. Grech and G. Pamula, “The local Hurst exponent of the financial time series in the vicinity of crashes on the polish stock exchange market,” Physica A, vol. 387, no. 16-17, pp. 4299–4308, 2008.
[5] N. M. Yuan, Z. T. Fu, H. Q. Li, and J. Mao, “Effect of extreme value loss on long-term correlated time series,” Theoretical and Applied Climatology, vol. 109, pp. 133–140, 2012.
[6] S. Pincus and R. E. Kalman, “Irregularity, volatility, risk, and financial market time series,” Proceedings of the National Academy of Sciences, vol. 101, no. 38, pp. 13709–13714, 2004.
[7] R. T. Baillie, “Long memory processes and fractional integration in econometrics,” Journal of Econometrics, vol. 73, no. 1, pp. 5–59, 1996.
[8] J. Arteche and J. Orbe, "A strategy for optimal bandwidth selection in local whittle estimation," *Econometrics and Statistics*, vol. 4, no. 4, pp. 3–17, 2017.

[9] D. Jonathan, "Multivariate models with long memory dependence in conditional correlation and volatility," *Journal of Empirical Finance*, vol. 48, pp. 162–180, 2018.

[10] H. E. Hurst, "Long term storage capacity of reservoirs," *Transactions American Society of Civil Engineers*, vol. 116, no. 76, pp. 770–808, 1951.

[11] L. Kristoufek, "Mixed-correlated ARFIMA processes for power-law cross-correlations," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 24, pp. 6484–6493, 2013.

[12] M. A. Hauser and E. Reschenhofer, "Estimation of the fractionally differencing parameter with the R/S method," *Computational Statistics & Data Analysis*, vol. 20, no. 5, pp. 569–579, 1995.

[13] V. Teverovsky and M. Taqqu, "Testing for long-range dependence in the presence of shifting means or a slowly declining trend, using a variance-type estimator," *Journal of Time Series Analysis*, vol. 18, no. 3, pp. 279–304, 1997.

[14] T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277–283, 1988.

[15] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E*, vol. 49, no. 2, pp. 1685–1689, 1994.

[16] P. Whittle, "The analysis of multiple stationary time series," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 15, no. 1, pp. 125–139, 1953.

[17] I. Lobato and P. M. Robinson, "Averaged periodogram estimation of long memory," *Journal of Econometrics*, vol. 73, no. 1, pp. 303–324, 1996.

[18] Y. Wang, "Minimax estimation via wavelets for indirect long-memory data," *Journal of Statistical Planning and Inference*, vol. 64, no. 1, pp. 45–55, 1997.

[19] P. Abry and D. Veitch, "Wavelet analysis of long-range-dependent traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 2–15, 1998.

[20] A. W. Lo, "Long term memory in stock market prices," *Econometrica*, vol. 59, no. 5, pp. 1279–1313, 1991.

[21] K. Shimotsu and P. C. B. Peter, "Exact local whittle estimation of fractional integration," *Annals of Statistics*, vol. 33, pp. 1890–1933, 2005.

[22] J. Hou and P. Perron, "Modified local whittle estimator for long memory processes in the presence of low frequency (and other) contaminations," *Journal of Econometrics*, vol. 182, no. 2, pp. 309–328, 2014.

[23] M. Bernardi, A. Marootti, and L. Petrella, "Multiple risk measures for multivariate dynamic heavy-tailed models," *Journal of Empirical Finance*, vol. 43, pp. 1–32, 2017.

[24] W. J. Zhou, J. Pan, and X. L. Wu, "Forecasting the realized volatility of CSI 300," *Physica A*, vol. 531, no. 1, Article ID 121799, 2019.

[25] J. Geweke and S. Porter-Hudak, "The estimation and application of long memory time series models," *Journal of Time Series Analysis*, vol. 4, no. 4, pp. 221–238, 1983.

[26] D. W. K. Andrews and P. Guggenberger, "A bias-reduced log-periodogram regression estimator for the long-memory parameter," *Econometrica*, vol. 71, no. 2, pp. 675–712, 2003.

[27] V. A. Reisen, C. Lévy-Leduc, and M. S. Taqqu, "AnM-estimator for the long-memory parameter," *Journal of Statistical Planning and Inference*, vol. 187, pp. 44–55, 2017.

[28] P. M. Robinson, "Log-periodogram regression of time series with long range dependence," *Annals of Statistics*, vol. 23, no. 3, pp. 1048–1072, 1995.

[29] C. M. Hurvich, R. Deo, and J. Brodsky, "The mean squared error of Geweke and Porter-hudak’s estimator of the memory parameter of a long-memory time series," *Journal of Time Series Analysis*, vol. 19, no. 1, pp. 19–46, 1998.

[30] C. Velasco, "Non-stationary log-periodogram regression," *Journal of Econometrics*, vol. 91, no. 2, pp. 325–371, 1999.

[31] C. Velasco, "Non-gaussian log-periodogram regression," *Econometric Theory*, vol. 16, no. 1, pp. 44–79, 2000.

[32] L. Charfeddine and D. Guégan, "Breaks or long memory behavior: an empirical investigation," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 22, pp. 5712–5726, 2012.

[33] D. Zhang and Q. Ji, "Further evidence on the debate of oil-gas price decoupling: a long memory approach," *Energy Policy*, vol. 113, pp. 68–75, 2018.

[34] F. X. Diebold and A. Inoue, "Long memory and regime switching," *Journal of Econometrics*, vol. 105, no. 1, pp. 131–159, 2001.

[35] P. Sibbertsen, "Log-periodogram estimation of the memory parameter of a long-memory process under trend," *Statistics & Probability Letters*, vol. 61, no. 3, pp. 261–268, 2003.

[36] F. F. Molinares, V. A. Reisen, and F. Cribari-Neto, "Robust estimation in long-memory processes under additive outliers," *Journal of Statistical Planning and Inference*, vol. 139, no. 8, pp. 2511–2525, 2009.

[37] L. Deng, "The distribution characteristic and bias of semi-parametric estimator based on an ARFIMA (p, d, q) process," *Statistical Research*, vol. 27, no. 9, pp. 97–104, 2010.

[38] H. D. J. Jeong, J. S. R. Lee, D. McNickle, and K. Pawlikowski, "Comparison of various estimators in simulated FGN," *Simulation Modelling Practice and Theory*, vol. 15, no. 9, pp. 1173–1191, 2007.

[39] C. Agiakloglou, P. Newbold, and M. Wohar, "Bias in an estimator of the fractional difference parameter," *Journal of Time Series Analysis*, vol. 14, no. 3, pp. 235–246, 1993.

[40] F. Sowell, "Maximum likelihood estimation of stationary univariate fractionally integrated time series models," *Journal of Econometrics*, vol. 53, no. 1-3, pp. 165–188, 1992.

[41] C. G. Lamoureux and W. D. Lastrapes, "Forecasting stock-return variance toward an understanding of stochastic implied volatilities," *Review of Financial Studies*, vol. 6, no. 2, pp. 293–326, 1993.

[42] R. Weron, "Estimating long range dependence: finite sample properties and confidence intervals," *Physica A*, vol. 312, no. 2, pp. 285–299, 2002.

[43] J. Elder and A. Serletis, "Long memory in energy futures prices," *Review of Financial Economics*, vol. 17, no. 2, pp. 146–155, 2008.