Detection of a long-range correlation with an adaptive detrending method

Chang-Yong Lee*
Department of Industrial and Systems Engineering, Kongju National University, Kongju 314-701, South Korea
(Received 27 March 2012; published 30 July 2012)

We propose a methodology of estimating the scaling exponent for a long-range correlation in a nonstationary time series from the perspective of the regression analysis. By an adaptive degree determination of a regression polynomial, the proposed methodology is designed to properly remove various types of trends embedded in the nonstationary signal so that the scaling exponent can be estimated without artificial crossovers. To show the validity of the proposed methodology, we applied it to the detrended fluctuation analysis and tested it out against correlated data superimposed by various types of trends. It turned out that, unlike the conventional technique, our approach was capable of eliminating artificial crossovers. We also discuss the statistical characteristics of the proposed method with regard to the estimation of the scaling exponent.

DOI: 10.1103/PhysRevE.86.011135 PACS number(s): 02.50.Tt, 05.45.Tp, 05.40.−a

I. INTRODUCTION

The long-range correlation in diverse complex systems has been widely investigated [1] and is often quantified in terms of the scaling exponent estimated by using various methods [2]. In the case of a nonstationary signal, which is usually associated with trends, it is important to remove the trend embedded in the signal when the trend is due to external conditions, not to the intrinsic dynamics of the system. To this end, methods involving a process of elimination of the trend, referred to as detrending, have been proposed and applied to numerous fields. Examples of these methods are the detrended fluctuation analysis (DFA) [3,4] and the detrended moving average analysis (DMA) [5]. In addition, a method for the long-range correlation between different time series in the presence of nonstationarity, called the detrended cross-correlation analysis (DCCA) [6], was also introduced.

In the case of DFA, for example, the crossover is known to appear, and this results from nonconstant scaling exponents in scale. The crossover in DFA refers to a point at which two straight lines representing corresponding scaling exponents have different slopes. When the crossover exists, the value of the scaling exponent is not a constant but depends on scales. Thus, the crossover indicates different scaling behaviors depending on the range of scales and implies the correlation property of the signal differs in scales of time or space. The cause of the crossover has been extensively investigated [7–10]; in particular, it was shown that the appearance of the crossovers in DFA is closely related to the trend embedded in a time series [7,8].

In general, the cause of the crossover is twofold. One cause is the intrinsic property of the dynamics of the system, and the other is an artifact caused by the trend not being properly removed. When the existence of the crossover is related to the latter cause, the onset of the crossover reduces not only the scaling range but the accuracy of estimation. Thus, from a methodological perspective, the appearance of the crossover is an important issue in detecting the long-range correlation of a nonstationary signal.

In the conventional methods [3–6], a local trend is estimated by introducing a regression polynomial of a *predetermined* degree, irrespective of the scale (or box size). Generally speaking, as the subset size increases, the trend becomes more involved. In addition, the strength of the trend can be different from samples in a given subset size. Thus, it is intuitively legitimate to assume that different subset sizes and samples may require a regression polynomial with different degrees; furthermore, a polynomial of a higher degree may be necessary for a better fitting as the subset size increases. In this sense, it is desirable to have a systematic and consistent scheme for determining the degree of the regression polynomial that takes into account the possible dependence of the degree on the subset size as well as different samples.

In this paper, we investigate a way of removing the trend in a nonstationary signal from the perspective of the regression analysis [11] and propose an effective methodology of eliminating an artificial crossover systematically by the regression polynomial of different degrees depending on both the subset size and samples. The determination of the degree can be accomplished by utilizing a statistical hypothesis test with a predetermined significance level as the parameter. An underlying assumption of both the conventional and the proposed DFAs is that trend in a time series can be efficiently, if not entirely, removed by a regression analysis on the time series. The proposed DFA is designed to eliminate trend from a signal more effectively than the conventional DFA by an *adaptive* determination of the degree of the regression polynomial. The crux is an adaptive determination of the degree of a regression polynomial with respect to not only the subset size but different samples within a given subset size. The proposed DFA may work best for a polynomial trend. However, since any nonpolynomial trend can be approximated by a polynomial of a certain degree, the proposed method can be applied to, at least approximately if not exactly, any nonpolynomial trend. To illustrate the usefulness of the proposed methodology, in this paper, we applied the methodology to DFA. The proposed methodology, however, can be readily applicable to any method, such as DMA and DCCA, that adopts a regression polynomial as a means of removing nonstationarities.

This paper is organized as follows. In Sec. II, we formalize the proposed methodology, called adaptive DFA (ADFA),

*clee@kongju.ac.kr

by explaining how the statistical hypothesis is applied to the adaptive determination of the regression polynomial in DFA. This section is followed by experimental results and a discussion of testing ADFA against correlated signals superimposed by trends. We also compare the results for both DFA and ADFA. The last section contains a summary and conclusion, including future studies.

II. ADAPTIVE DETRENDED FLUCTUATION ANALYSIS

For a given time series of $x_1,x_2,\ldots,x_N$, DFA first divides the time series into N/n subsets (or boxes) of an equal size n and calculates the accumulated time series $Y_n(i,k) = \sum_{j=1}^{(i-1)n+k} x_j$, where i ($i = 1,2,\ldots,N/n$) is the subset index and k ($k = 1,2,\ldots,n$) denotes the data points in each subset. With the accumulated time series $Y_n(i,k)$ of N/n subsets, DFA looks for scaling behavior of the fluctuation function $F(n)$ of $Y_n(i,k)$ as the box size n varies:

$$F(n) \propto n^{\beta}, \tag{1}$$

where β is the scaling exponent and $F(n)$ is given as

$$F^2(n) = \frac{n}{N}\sum_{i=1}^{N/n}\frac{1}{n}\sum_{k=1}^{n}[Y_n(i,k) - \hat{y}_{n,p}(i,k)]^2. \tag{2}$$

Here, $\hat{y}_{n,p}(i,k)$ is an estimate of a regression polynomial $y_{n,p}(i,k)$ of the degree p for $Y_n(i,k)$, evaluated at the kth value of the ith subset of the size n. Thus, $F^2(n)$ is an average fluctuation over N/n subsets of a given size n.

To demonstrate how to adaptively determine the degree p, we focus on a subset of the size n and drop the subset index i because the same scheme can be applied to the other subsets of the same size. With this convention, the regression polynomial $y_{n,p}(k)$ of the degree p fitted in the subset size n can be expressed as

$$y_{n,p}(k) = a_0 + a_1 z_k + a_2 z_k^2 + \cdots + a_p z_k^p + \epsilon_k, \tag{3}$$

where the a_r's ($r = 0,1,\ldots,p$) are unknown parameters to be estimated. In addition, it is usually assumed that the error $\epsilon_k \sim N(0,\sigma^2)$, where "$\sim$" hereafter stands for "is distributed as."

With a subset of size n, the total variance of $Y_n(k)$'s, called the total sum of squares (SST), can be decomposed into two parts:

$$\begin{aligned}&\sum_{k=1}^{n}[Y_n(k) - \bar{Y}]^2\\ &\quad= \sum_{k=1}^{n}[Y_n(k) - \hat{y}_{n,p}(k) + \hat{y}_{n,p}(k) - \bar{Y}]^2\\ &\quad= \sum_{k=1}^{n}[Y_n(k) - \hat{y}_{n,p}(k)]^2 + \sum_{k=1}^{n}[\hat{y}_{n,p}(k) - \bar{Y}]^2,\end{aligned} \tag{4}$$

where $\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_n(k)$ and the cross term vanishes. The vanishment of the cross term stems from estimating a_i in Eq. (3). The usual method of estimating a_i is the so-called method of least squares, with which the cross term can be shown to vanish by minimizing $\sum_k \epsilon_k^2$ with respect to a_i. See Ref. [11] for a detailed proof. The first term on the right hand side of Eq. (4) is called the sum of squares due to residual errors (SSE) X_{SSE} (or the residual sum of squares), and the second term is the sum of squares due to regression (SSR) X_{SSR}. SSR represents the amount of error that is removed by fitting the regression polynomial to the data, while SSE is the amount of error that still remains after the regression polynomial has been fitted. In this sense, SSE can be regarded as a pure fluctuation. Thus, from the perspective of the regression analysis, $F^2(n)$ is nothing but the average of X_{SSE}/n over the number of subsets N/n.

The adaptive determination of p is designed by utilizing the probability distribution of the following statistics [12]:

$$F \equiv \frac{X_{\mathrm{SSR}}/p}{X_{\mathrm{SSE}}/(n-p-1)} \sim F(p,n-p-1). \tag{5}$$

That is, the ratio of SSR to SSE divided by its degrees of freedom is distributed as the Fisher-Snedecor distribution (or F distribution) with parameters p and $n-p-1$. The F statistics is used for the inference about the difference between SSR and SSE. The degree p is determined by using Eq. (5) as the statistics for a hypothesis test with a predetermined significance level in such a way that the regression polynomial of the degree p significantly contributes to SSR. More specifically, we find the highest p with which the corresponding test statistic F_0, the computed F value of Eq. (5), satisfies

$$F_0 > F_\alpha(p,n-p-1), \tag{6}$$

where $F_\alpha(p,n-p-1)$ is the critical F value with the significance level α, the only free parameter. That is, the terms of a degree higher than p do not contribute significantly to SSR. In this way, the proposed scheme systematically takes into account the possible dependence of the polynomial degree on both different subset sizes n and different subsets within a given size.

An effective implementation of the suggested methodology can be accomplished by rewriting the regression model of Eq. (3) in terms of the orthogonal polynomials.

$$y_{n,p}(k) = b_0\phi_0(z_k) + b_1\phi_1(z_k) + \cdots + b_p\phi_p(z_k) + \epsilon_k, \tag{7}$$

where $\phi_r(z_k)$ is a rth-degree polynomial in z_k and the polynomials are orthogonal over the z set:

$$\sum_{k=1}^{n}\phi_r(z_k)\phi_s(z_k) = 0 \quad \text{if} \quad r \neq s. \tag{8}$$

The orthogonal polynomials are especially useful when the z_k's are equally spaced, which is mostly satisfied in the case of a time series. In this case, the orthogonal polynomials are given recursively as (for $z_k = 1,2,\ldots,n$) [13]

$$\phi_0(z_k) = 1, \phi_1(z_k) = z_k - \frac{n+1}{2}$$

and, for $r \geqslant 1$,

$$\phi_{r+1}(z_k) = \phi_r(z_k)\phi_1(z_k) - \frac{r^2(n^2-r^2)}{4(4r^2-1)}\phi_{r-1}(z_k).$$

Because of the orthogonality, the least-squares estimate of coefficient b_r in Eq. (7) is solely contributed by ϕ_r, independent of the other ϕ_s's ($s \neq r$). Thus, each b_r can be estimated

independently from other b_s's ($s \neq r$) [11]. This implies that SSR contributed by ϕ_r is independently evaluated from ϕ_s's ($s \neq r$) and that it can be separately decided whether ϕ_r should be included in the resulting regression polynomial. When ϕ_r's ($r \leqslant s$) are included in the regression polynomial, Eq. (5) for each $\phi_r (r = 1,2,\ldots,s)$ becomes

$$F = \frac{X_{\mathrm{SSR},r}}{X_{\mathrm{SSE}}/(n-s-1)} \sim F(1,n-s-1), \tag{9}$$

where $X_{\mathrm{SSR},r}$ is SSR contributed by ϕ_r. With the above scheme and $b_0 = \bar{Y}$, we start with ϕ_1 and successively include an orthogonal polynomial of higher degree until F_0 evaluated by Eq. (9) indicates an insignificant contribution to SSR. That is, we stop adding ϕ_s when $F_0 < F_\alpha(1,n-s-1)$.

The usage of the orthogonal polynomials has the additional advantages of less time complexity and higher accuracy. The estimate of b_r in Eq. (7) is explicit so that no solutions of sets of simultaneous equations are needed, as is the case for Eq. (3). This also enables us to avoid any possible roundoff error, known as the ill-conditioned problem [14], which may occur and is susceptible to the matrix calculation required in the conventional method, especially for a high degree.

III. EXPERIMENTAL RESULTS AND DISCUSSION

We demonstrate an advantage of the proposed method (ADFA) by applying it to an artificially generated correlated signal, which is additively superimposed by various trends: quadratic, periodic, and power-law trends [7]. We also applied DFA to the superimposed signal for comparison. The results, as displayed in Fig. 1, show that the scaling behavior of $F(n)$ by using DFA at relatively small subset sizes n is different from that at large n's. As a result, there exist crossovers, marked by arrows, at which the fluctuation tends to increase by a different rate with increasing subset size. Considering that the trend is accumulated and becomes more dominant as the subset size increases, the regression polynomial of a fixed degree may not suffice to properly eliminate the trend. In contrast, no crossover is found when ADFA is applied to the superimposed signal, which implies that ADFA successfully removes the trends. Moreover, the estimated scaling exponents are consistent, within statistical errors, regardless of the type of the trend. This also supports the validity of the proposed method. In addition, the results for ADFA are insensitive to the significance level of $0.01 < \alpha < 0.1$ in the hypothesis test statistics [data not shown] [16].

The inset of Fig. 1 depicts the dependence of the degree of the regression polynomial in ADFA on the subset size n. As we expect, the degree is not a constant but varies in n, and a polynomial of a higher degree is adopted as n increases. Furthermore, the degree also varies in subsets for a given subset size as the error bars indicate. These findings imply that the adaptive determination of the degree can properly take into account the contribution from the trends in not only the different subset sizes but also different subsets within the same size.

In general, the higher the degree p of the regression polynomial is, the better the polynomial fits to the data. As a result, SSR in Eq. (4) becomes larger for a higher degree p. Given that SST is a constant for given data, the observed

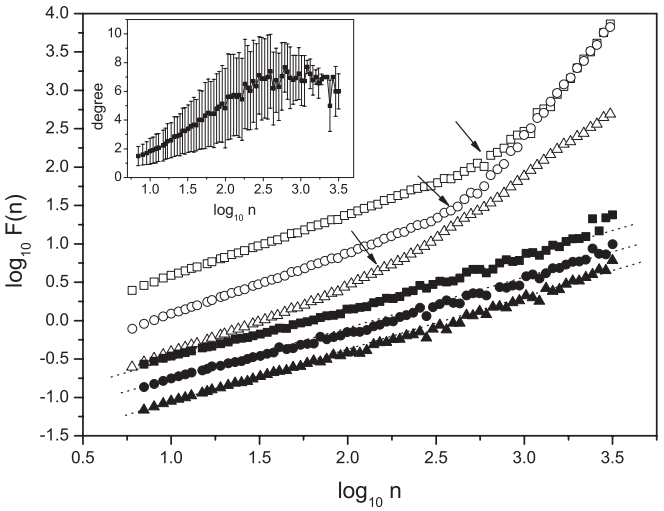

FIG. 1. Log_{10}-log_{10} plots of the fluctuation $F(n)$ vs the subset size n for a correlated signal, superimposed additively by various trends. The correlated signals are generated by using the method proposed in Ref. [15]. For DFA, $F(n)$ vs n is plotted for the quadratic (open squares), periodic (open circles), and power-law (open triangles) trends together with the arrows indicating the crossover points. For ADFA, $F(n)$ vs n is plotted for the same superimposed trends: quadratic (solid squares), periodic (solid circles), and power-law (solid triangles) trends. By using ADFA with a significance level $\alpha = 0.1$, we obtained the scaling exponents, estimated by the least-squares fit (dotted lines), $\beta = 0.68, 0.67, 0.68$ for the quadratic, periodic, and power-law trends, respectively. The plots for periodic and power-law trends are vertically shifted for clarity. The inset shows a plot of the average degree of the regression polynomial over the N/n subsets vs n, where ADFA is applied to the correlated signal superimposed by the quadratic trend. The error bars are twice the estimated sample standard deviations from the N/n subsets for each n.

tendency of a higher p for large n implies that SSE by ADFA becomes smaller than that by DFA as n increases. With $F(n)$ being nothing but the average of SSE over subsets, we can infer that the difference between $F(n)$'s obtained by DFA and ADFA becomes considerable as n increases. This, in turn, implies that estimated β by using ADFA is smaller than that by using DFA.

To empirically verify the above assertion, we estimate the scaling exponents β for the correlated time series of known correlation exponents γ [15] by using both ADFA and DFA. As shown in the inset of Fig. 2, not only is $F(n)$ by ADFA smaller than that by DFA, but the disparity becomes marked as n increases. Thus, we expect that the scaling exponent β obtained by ADFA is smaller than that by DFA, as shown in Fig. 2. Equally important, we find that the scaling exponents of ADFA is smaller than those of DFA by an approximately constant amount regardless of γ's as the estimated slope ($\approx$0.47), being close to 0.5, indicates. This suggests that the disparity is consistent and thus not due to any irregularity of the proposed method.

The disparity can be further investigated in terms of the probability distribution of a random sample. Consider a randomly generated (thus uncorrelated) time series $x_1, x_2, \ldots, x_n$ of a mean μ and a variance σ^2. Then, an accumulated

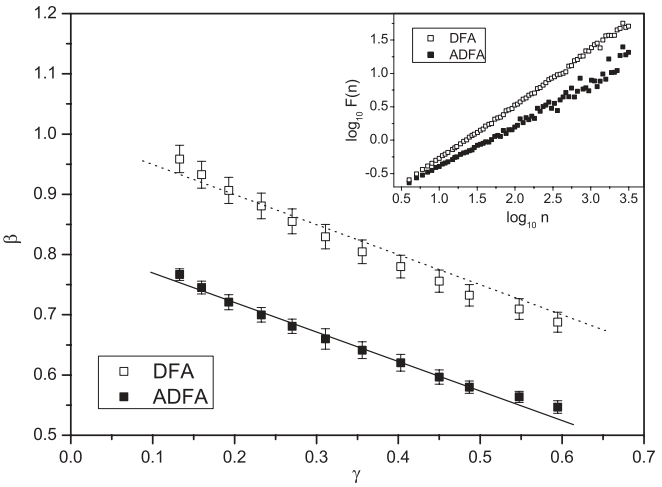

FIG. 2. Plots of the scaling exponent β vs the correlation exponent γ [15] obtained by DFA (open squares) and ADFA (solid squares) with the correlated data of various γ's. The solid line is a least-squares fit of the results of ADFA being $\beta \approx 0.81 - 0.47\gamma$, and the dotted line represents the theoretical relation between β and γ for DFA, $\beta = 1 - \gamma/2$ [17]. The correlated signals are generated by using the method proposed in Ref. [15], and the error bars are the standard deviation of 10 independent trials. The inset shows the $\log_{10}$-$\log_{10}$ plots of $F(n)$ vs n obtained by DFA (open squares) and ADFA (solid squares) when $\gamma = 0.65$.

value $Y_n(k) = \sum_{i=1}^{k} x_i$ of x_i's is approximately distributed as $N(k\mu, k\sigma^2)$ by the central limit theorem. From the theory of the regression analysis [11] and the definition of the chi-square distribution [12], it follows that

$$\sum_{k=1}^{n} \frac{[Y_n(k) - \hat{y}_{n,p}(k)]^2}{k\sigma^2} \sim \chi^2(n - p - 1), \tag{10}$$

where $\chi^2(n - p - 1)$ denotes a chi-square distribution with $n-p-1$ degrees of freedom. By making an approximation of $k \approx \langle k \rangle = n/2$, Eq. (10) becomes

$$Y^2_{n,p} \equiv \frac{2}{n\sigma^2} \sum_{k=1}^{n} [Y_n(k) - \hat{y}_{n,p}(k)]^2 \sim \chi^2(n - p - 1). \tag{11}$$

With Eq. (11), we obtain an analytic expression of $F^2(n)$ in Eq. (2) for an uncorrelated signal as

$$F^2(n) = \frac{\sigma^2}{2} \langle Y^2_{n,p} \rangle \approx \frac{\sigma^2}{2}(n - p - 1), \tag{12}$$

where $\langle \cdot \rangle$ represents an average over N/n subsets and we have used the fact that the expectation value of a random variable being distributed as $\chi^2(n-p-1)$ is $n-p-1$. From this, we find that the fluctuation $F^2(n)$ depends not only on the subset size n but also on the degree p.

In particular, when p ($\ll n$) is independent of n, which is the case for DFA, $F(n) \propto n^{1/2}$, so that the scaling exponent $\beta_0 \approx 1/2$ for an uncorrelated signal [3]. For the case of ADFA, on the contrary, p varies in samples of a given subset size n and tends to increase in n, as shown, for instance, in the inset of Fig. 1. This suggests that $F(n)$ by ADFA is expected to be smaller than that by DFA. Furthermore, due to the increasing tendency of p as n, the difference in $F(n)$ between DFA and ADFA becomes prominent as n increases. Thus, the scaling exponent β_0 for uncorrelated data obtained by ADFA should be less than $1/2$, although it is highly unlikely to find analytically the numerical expression of β_0 for ADFA due mainly to the nontrivial dependence of p on both the n and N/n subsets.

Nevertheless, similar to the case of DFA in which $1/2 < \beta < 1$ indicates a long-range correlation, the existence of a long-range correlation in a time series can be tested by using ADFA. That is, we can infer that a long-range correlation exists in a signal when the scaling exponent β is in the range of $\beta_0 < \beta < \beta_0 + 1/2$, where β_0 is the exponent for a randomly shuffled (thus uncorrelated) version of the original time series.

IV. SUMMARY AND CONCLUSION

In this paper, we proposed a methodology of estimating the scaling exponent for a long-range correlation in a nonstationary time series by properly removing any trend embedded in the time series. The proposed methodology was designed to determine adaptively the degree of a regression polynomial in terms of the statistical hypothesis test with a significance level as the parameter. The adaptively determined degree varied from not only different subset sizes but different subsets within a given subset size.

To demonstrate the usefulness of the proposed methodology, we applied it to DFA and tested the implemented DFA, an adaptive DFA (ADFA), out against correlated data superimposed by various types of trends. It turned out that ADFA could estimate the scaling exponent without artificial crossovers, in contrast to the conventional technique. This result showed that ADFA was capable of eliminating an artificial crossover systematically by the regression polynomial of different degrees depending on both the subset size and subsets. We also analytically discussed the statistical characteristics of ADFA for an uncorrelated signal.

This study is the first step in the estimation of the scale exponent from the perspective of the adaptive determination of the polynomial degree. Any variation of the proposed methodology would be interesting and merit further investigation. Although, in this paper, we used the F distribution as the test statistic, other test statistics, such as the sample coefficient of determination ($r^2 \equiv X_{\mathrm{SSR}}/X_{\mathrm{SST}}$), could be an alternative and deserve to be investigated. In addition, the proposed methodology can be readily implemented to DMA and DCCA, in addition to DFA, which adopt a regression polynomial as a means of removing nonstationarities. Further studies regarding not only the methodology itself but also practical application would be valuable and support the advantage of the adaptive determination of the polynomial degree in estimating the scaling exponent.

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant No. 2011-0026677).

[1] For a review, see, for example, M. Sahimi, Rev. Mod. Phys. **65**, 1393 (1993); A. Bunde and S. Havlin, *Fractals in Science* (Springer, Berlin, 1994).
[2] For a review of the methods, see, for example, B. Malamud and D. Turcotte, J. Stat. Plann. Inference **80**, 173 (1999); M. Taqqu, V. Teverovsky, and W. Willinger, Fractals **3**, 785 (1995).
[3] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[4] C.-k. Peng, S. Havlin, H. E. Stanley, and A. Goldberger, Chaos **5**, 82 (1995).
[5] S. Arianos, A. Carbone, and C. Türk, Phys. Rev. E **84**, 046113 (2011).
[6] B. Podobnik and H. E. Stanley, Phys. Rev. Lett. **100**, 084102 (2008).
[7] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley, Phys. Rev. E **64**, 011114 (2001).
[8] Z. Chen, P. C. Ivanov, K. Hu, and H. E. Stanley, Phys. Rev. E **65**, 041107 (2002).
[9] G. Rangarajan and M. Ding, Phys. Rev. E **61**, 004991 (2000).
[10] C. Heneghan and G. McDarby, Phys. Rev. E **62**, 006103 (2000).
[11] For a general reference for the regression analysis, see, for example, G. Seber and A. Lee, *Linear Regression Analysis* (Wiley, Hoboken, NJ, 2003).
[12] R. V. Hogg, A. Craig, and J. W. McKean, *Introduction to Mathematical Statistics* (Prentice Hall, New York, 2005).
[13] N. Johnson and F. Leone, *Statistics and Experimental Design in Engineering and the Physical Sciences*, Vol. 1, Chap. 12 (Wiley, New York, 1977).
[14] J. Todd, J. Res. Natl. Bur. Stand. Sec. B (US) **65**, 19 (1961).
[15] H. A. Makse, S. Havlin, M. Schwartz, and H. E. Stanley, Phys. Rev. E **53**, 5445 (1996).
[16] In the statistical inference, especially in hypothesis testing, the most commonly used levels of significance α are $0.01 \leqslant \alpha \leqslant 0.10$. We have tried a few α's and found that the results are independent of α within the statistical error.
[17] It is known [15] that the correlation function $C(n) = \langle x_i x_{i+n} \rangle \propto n^{-\gamma}$ exhibits a long-range correlation for $0 < \gamma < 1$. Given that the scaling exponent of $1/2 < \beta < 1$ indicates a long-range correlation for DFA, the relation between β and γ is $\beta = 1 - \gamma/2$.