Heuristic segmentation of a nonstationary time series

Kensuke Fukuda,[1,2,3] H. Eugene Stanley,[2] and Luís A. Nunes Amaral[3]
[1]*NTT Network Innovation Laboratories, Tokyo 180-8585, Japan*
[2]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA*
[3]*Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA*
(Received 4 August 2003; published 25 February 2004)

Many phenomena, both natural and human influenced, give rise to signals whose statistical properties change under time translation, i.e., are nonstationary. For some practical purposes, a nonstationary time series can be seen as a concatenation of stationary segments. However, the exact segmentation of a nonstationary time series is a hard computational problem which cannot be solved exactly by existing methods. For this reason, heuristic methods have been proposed. Using one such method, it has been reported that for several cases of interest—e.g., heart beat data and Internet traffic fluctuations—the distribution of durations of these stationary segments decays with a power-law tail. A potential technical difficulty that has not been thoroughly investigated is that a nonstationary time series with a (scalefree) power-law distribution of stationary segments is harder to segment than other nonstationary time series because of the wider range of possible segment lengths. Here, we investigate the validity of a heuristic segmentation algorithm recently proposed by Bernaola-Galván *et al.* [Phys. Rev. Lett. **87**, 168105 (2001)] by systematically analyzing surrogate time series with different statistical properties. We find that if a given nonstationary time series has stationary periods whose length is distributed as a power law, the algorithm can split the time series into a set of stationary segments with the correct statistical properties. We also find that the estimated power-law exponent of the distribution of stationary-segment lengths is affected by (i) the minimum segment length and (ii) the ratio $R \equiv \sigma_\epsilon / \sigma_{\bar{x}}$, where $\sigma_{\bar{x}}$ is the standard deviation of the mean values of the segments and σ_ϵ is the standard deviation of the fluctuations within a segment. Furthermore, we determine that the performance of the algorithm is generally not affected by uncorrelated noise spikes or by *weak* long-range temporal correlations of the fluctuations within segments.

DOI: 10.1103/PhysRevE.69.021108 PACS number(s): 05.40.−a

I. INTRODUCTION

A stationary time series has statistical properties that do not change under time translation [1]. Interestingly, the time series that arise in a large number of phenomena in a broad range of areas—including physiologic systems, economic systems, vehicle traffic systems, and the Internet [2–11]—are nonstationary. Thus nonstationarity is a property common to both natural and human-influenced phenomena. For this reason, the statistical characterization of the nonstationarities in real-world time series is an important topic in many fields of research and numerous methods of characterizing nonstationary time series have been proposed [12].

One useful approach to quantifying a nonstationary time series is to view it as consisting of a number of time segments that are themselves stationary. The statistical properties of the segments (i) can help us better understand the overall nonstationarity of the time series and (ii) yield practical applications. For example, developing control algorithms for Internet traffic will be easier if we understand the statistical properties of these stationary segments [10].

In general, it is impossible to obtain the exact segmentation of a nonstationary time series because of the complexity of the calculation. An exact segmentation algorithm requires a computation time that scales as $O(N^N)$, where N is the number of points in the time series [13]. Hence, such an algorithm is not practical. For this reason, the segmentation of a real-world time series must accomplish a trade-off between the complexity of the calculation and the desired precision of the result.

Bernaola-Galván and co-workers recently proposed a heuristic segmentation algorithm [14] designed to characterize the stationary durations of heart beat time series. In this algorithm [14], the calculation cost is reduced by iteratively attempting to segment the time series into only *two* segments. A decision to cut the times series is made by evaluating a modified Student's *t*-test for the data in the two segments [15].

The application of this segmentation algorithm reveals that the distribution of the stationary durations in heart beat time series decays as a power law [14]. Intriguingly, a recent analysis of Internet traffic uncovered that the distribution of stationary durations in the fluctuation of the traffic flow density also follows a power-law dependence [16]. Because these signals have their origin in such diverse contexts, these findings suggest that the power-law decay of the distribution of the stationary period may be a common occurrence for complex time series. Thus, the correct implementation and interpretation of the results obtained by the segmentation algorithm is essential in understanding the dynamics of the system. In fact, there are many implementation issues concerning the segmentation algorithm of Ref. [14] that have not yet been addressed explicitly in the literature, especially those concerning the proper estimation of the value of the power-law tail's exponent in the cumulative distribution of stationary durations.

In this paper we systematically analyze different types of

surrogate time series to determine the scope of validity of the segmentation algorithm of Bernaola-Galván and co-workers [14]. In Sec. II, we briefly explain the segmentation algorithm. In Sec. III, we present results for the dependence of the exponent of the power-law tail on the minimum length of the segments in the distribution of the stationary durations. In Sec. IV, we consider the effect of the amplitude of the noise and the presence of spike-type noise. In Sec. V we consider long-ranged temporally correlated noise. Finally, in Sec. VI, we summarize our findings.

II. IMPLEMENTING THE SEGMENTATION ALGORITHM

A. The algorithm

To divide a nonstationary time series into stationary segments [14], we move a sliding pointer from left to right along the time series and, at each position of the pointer, compute the mean of the subset of the signal to the left of the pointer μ_{left} and to the right μ_{right}. For two samples of Gaussian distributed random numbers, the statistical significance of the difference between the means of the two samples, μ_1 and μ_2, is given by Student's *t*-test statistic [17],

$$t \equiv \left| \frac{\mu_1 - \mu_2}{S_D} \right|, \tag{1}$$

where

$$S_D = \left(\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2} \right)^{1/2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)^{1/2} \tag{2}$$

is the pooled variance [18], s_1, s_2 are the standard deviations of the two samples, and N_1 and N_2 are the number of points in the two samples.

Moving the pointer along our time series from left to right, we calculate t as a function of the position in the time series. We use the statistic t to quantify the difference between the means of the left-side and right-side time series. Larger t means that the values of the mean of both time series are more likely to be significantly different, making point $t_{\max}$, with the largest value of t, a good candidate as a cut point.

We then calculate the statistical significance $\mathcal{P}(t_{\max})$. Note that $\mathcal{P}(t_{\max})$ is not the standard Student's *t*-test [18] because we are not comparing independent samples. We could not obtain $\mathcal{P}(t_{\max})$ in a closed analytical form, so that $\mathcal{P}(t_{\max})$ is *numerically approximated* as [19]

$$\mathcal{P}(t_{\max}) \approx \{1 - I_{[\nu/(\nu+t_{\max}^2)]}(\delta\nu, \delta)\}^{\eta}, \tag{3}$$

where $\eta = 4.19 \ln N - 11.54$ and $\delta = 0.40$ are obtained from Monte Carlo simulations [14], N is the length of the time series to be split, $\nu = N-2$, and $I_x(a,b)$ is the incomplete beta function.

If the difference in mean is not statistically significant—i.e., if is smaller than a threshold (typically set to 0.95)—then the time series is not cut. If the difference in means between the left and right part of the time series is statistically significant, then the time series is cut into two segments as long as the means of the two new segments are significantly different from the means of the adjacent segments. If the time series is cut, we continue iterating the above procedure recursively on each segment until the obtained significance value is smaller than the threshold, or the length ℓ of the obtained segments is smaller than an imposed minimum segment length ℓ_0. We will see that the value of ℓ_0 is one of the parameters controlling the accuracy of the algorithm.

B. Surrogate time series

To investigate the validity of the algorithm, we analyze surrogate time series $x(t)$ generated by linking segments with different means. As described in Sec. I, the cumulative distribution of the stationary durations for some real-world time series is characterized by a power-law decay in the tail, so the probability of finding a segment of length larger than m, i.e., the cumulative distribution of segment lengths in our time series, is

$$P(>m) = \frac{\gamma+1}{m_0^{\gamma+1}} m^{-\gamma} \quad \text{for} \quad m \gg m_0, \tag{4}$$

where m_0 is the minimum length of a segment.

We generate time series with a power-law distribution of segment lengths by the following procedure.

(1) Draw from the interval $[m_0, +\infty]$ a sequence of segment lengths $\{m_i\}$ with distribution given by Eq. (4).

(2) Draw from the interval [0,1] a sequence of mean time series values $\{\overline{x_i}\}$ with uniform probability.

(3) Draw from the interval $[-\sqrt{3}\sigma_\epsilon, \sqrt{3}\sigma_\epsilon]$ a sequence of fluctuation values $\{\epsilon_i(k_i)\}$, for $k_i = 1, \ldots, m_i$, with uniform probability.

The resulting time series is given by

$$x(t) = \overline{x_i} + \epsilon_i(k_i), \tag{5}$$

where i is such that

$$\sum_{j=0}^{i-1} m_j < t \leqslant \sum_{j=0}^{i} m_j \tag{6}$$

and k_i is such that

$$k_i = t - \sum_{j=0}^{i-1} m_j . \tag{7}$$

To quantify the level of the noise, we define the ratio

$$R \equiv \frac{\sigma_\epsilon}{\sigma_{\overline{x}}}, \tag{8}$$

where $\sigma_{\overline{x}}$ is the standard deviation of the mean of the segments and σ_ϵ is the standard deviation of the fluctuations within a segment. For $\overline{x_i}$ uniformly distributed in the interval [0,1], $\sigma_{\overline{x}} = 0.3$.

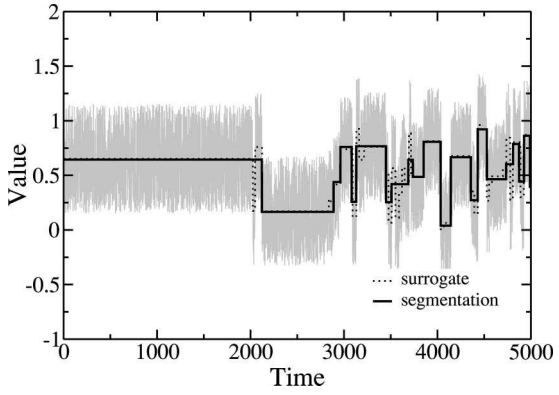

FIG. 1. Surrogate time series constructed according to the procedure described in Sec. II B with parameters $\gamma=1.0$, $m_0=20$, and $R=1$. The time series is drawn in light gray, and the stationary segments are represented by a dashed gray line. The black line displays the output of the segmentation algorithm for $\ell_0=50$. It is visually apparent that the black line provides an adequate coarse-grained description of the surrogate time series. Note also that the algorithm cannot extract the smallest segments because of the restrictions on resolution for $m<\ell_0$.

For each set of parameters (γ, m_0, R) we generate ten time series, each with 50 000 data points. Note that knowing *a priori* the value of m_0 in a real-world time series is unlikely, but in order to test the algorithm in a consistent way, we consider in the following section $m_0 \geqslant 20$ because that is the resolution limit for the algorithm (see also Appendix A).

Figure 1 displays a surrogate time series, the corresponding stationary durations, and the result of the segmentation algorithm with $\ell_0=50$ and $m_0=20$. The segments obtained by the segmentation algorithm do not exactly match the stationary segments in the surrogate time series but the figure strongly suggests that the algorithm provides the correct coarse-grained description of the time series. As expected, the segmentation algorithm cannot extract segments with length $m<\ell_0$.

III. ACCURACY OF THE SEGMENTATION ALGORITHM

A. Dependence on ℓ_0 and m_0

Figure 2(a) displays the cumulative distribution of segment lengths, which is the probability of finding a segment with length larger than ℓ for a surrogate time series split for different values of ℓ_0. The cumulative distributions of the length of the stationary segments cut by the segmentation algorithm for surrogate time series are well fit by a power-law decay. For all cases, we find

$$P_{\ell_0,m_0}(>\ell)\sim \ell^{-\hat{\gamma}(\ell_0,m_0)} \tag{9}$$

with the same exponent value $\hat{\gamma}\approx 1.0$ [20], indicating that the segmentation algorithm splits the nonstationary time series into segments with the correct asymptotic statistical properties. However, the range of scales for which we observe a power-law decay with $\hat{\gamma}\approx\gamma$ depends strongly on the selection of ℓ_0.

For ℓ greater than about $5\ell_0$, the tails of the distributions are consistent with a power-law decay with $\hat{\gamma}\approx\gamma=1.0$. Additionally, all $P(>\ell)$ track the surrogate data for $\ell>1000$, i.e., the algorithm correctly identifies the large segments independently of the selection of ℓ_0. For $\ell<5\ell_0$, the distributions are not consistent with a power-law decay. The origin of this behavior lies in the fact that (i) for $\ell_0=20=O(m_0)$, there are not enough data points to reliably perform Student's *t*-test, so one cannot reasonably expect any statistical procedure to be able to extract those short segments and (ii) for $\ell_0 \gg m_0$, one fails to extract the stationary segments with m

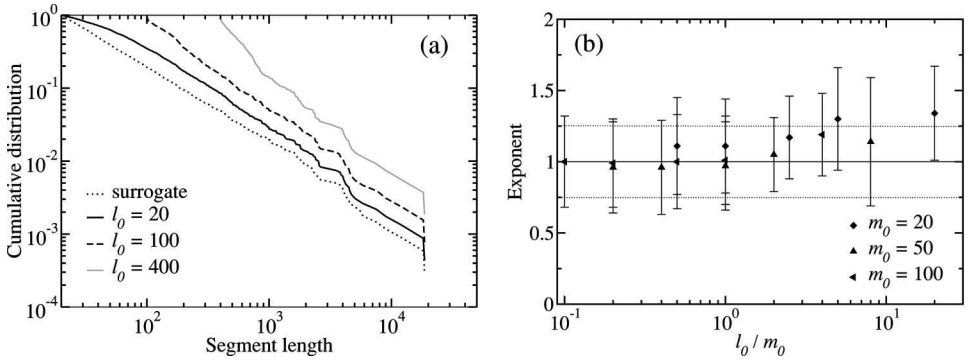

FIG. 2. (a) Cumulative distribution $P(>\ell)$ of segment lengths larger than ℓ for surrogate time series with 50 000 data points; $\gamma=1.0$, $m_0=20$, and $R=1$. The dotted line indicates the input segment length distribution for the surrogate time series. The slope of the line is 1.0 for $20<m<4000$. The other curves show $P(>\ell)$ for different values of ℓ_0. For $\ell_0=20$, the curve is well described in the range $100<\ell<4000$ by a power law with $\hat{\gamma}\approx 1.0$. The distribution does not decay as a power law for $\ell<100$ due to the fact that the segmentation algorithm cannot split a time series with a number of points insufficient to perform Student's *t*-test. For $\ell_0=400$, $P(>\ell)$ decays as a power law for $1000<\ell<8000$. The segmentation algorithm correctly splits all segments in the surrogate time series with length $\ell>1000$. (b) Dependence of $\hat{\gamma}$ on ℓ_0 and m_0. The mean and the standard deviation of the exponent for the original time series is 1.05 ± 0.24, shown by the black solid and dotted lines. The error bars show the standard deviation of the estimates $\hat{\gamma}$. For $\ell_0/m_0<5$, we find $\hat{\gamma}\approx 1$. Thus, the algorithm accurately estimates exponents in this region. However, the values of the exponent are close to 1.3 for $\ell_0/m_0>10$, meaning that $m_0 \ll \ell_0$ leads to an overestimation of γ.

TABLE I. Estimated exponent $\hat{\gamma}$, as defined by Eq. (5) for $\gamma=1.0$. The mean and the standard deviation of the exponents are calculated for the ranges indicated inside parentheses using Eq. (B1). The column labeled "input" presents exponent estimates obtained from the segment lengths used to generate the surrogate time series. We find $\hat{\gamma}\approx\gamma=1.0$ for $\ell_0<5\ m_0$.

	ℓ_0					
m_0	10	20	50	100	400	Input
20	1.1 ± 0.3	1.1 ± 0.3	1.2 ± 0.3	1.3 ± 0.4	1.3 ± 0.3	1.1 ± 0.3
	$(\ell>20)$	$(\ell>20)$	$(\ell>50)$	$(\ell>110)$	$(\ell>1000)$	
50	1.0 ± 0.3	1.0 ± 0.3	1.0 ± 0.3	1.1 ± 0.3	1.1 ± 0.5	1.0 ± 0.2
	$(\ell>50)$	$(\ell>50)$	$(\ell>100)$	$(\ell>110)$	$(\ell>1000)$	
100	1.0 ± 0.3	1.0 ± 0.3	1.0 ± 0.3	1.0 ± 0.3	1.2 ± 0.3	1.1 ± 0.3
	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>1000)$	

$<\ell_0$. The reason for the latter is that the value of ℓ_0 is in this case considerably larger than the length of the shortest segments in the time series, so the algorithm is forced to merge a number of short segments into longer ones with length greater than ℓ_0. This process gives rise to a deficit in the number of segments with length smaller than ℓ_0. Hence the slower initial decay of $P(>\ell)$.

Table I shows the mean and standard deviation of the estimated exponent value $\hat{\gamma}$ calculated from surrogate time series for several values of m_0 and ℓ_0 (see Appendix B for details on how to estimate $\hat{\gamma}$). Our results indicate that $\hat{\gamma}$ depends on both m_0 and ℓ_0: If $\ell_0\gg m_0$, $\hat{\gamma}$ overestimates γ, while if $\ell_0\approx 0(m_0)$, the algorithm correctly estimates the value of the exponent γ; cf. Fig. 2(b).

B. Dependence on γ

Next, we focus on the dependency of the accuracy of the segmentation algorithm on the value of γ. Figure 3(a) displays the cumulative distribution of segment lengths for surrogate time series generated with $\gamma=2.0$. A challenge for the segmentation algorithm is that Eq. (4) indicates that the probability of finding segments with length shorter than $4m_0$ is 90% for $\gamma=2.0$, which can lead to the "aggregation" of several consecutive segments of small length into a single longer segment.

As we found for $\gamma=1$, the tails of the distributions follow power-law decays for large ℓ, showing that the algorithm yields segments with the proper statistical properties. In Fig. 3(b), we show the dependence of $\hat{\gamma}$ on ℓ_0 and m_0 for $\gamma=2.0$. We find that for $\ell_0/m_0<4$ the algorithm extracts segments with distributions of lengths that decay in the tail as power laws with exponents that are quite close to 2.0.

In Tables II and III, we report the values of $\hat{\gamma}$ for $\gamma=2.0$ and $\gamma=3.0$, respectively. We find that for small m_0 and large γ, one overestimates γ. Thus, we surmise that for $\gamma>3.0$, it becomes impractical to estimate γ accurately, except for extremely long time series. This fact is not as serious a limitation as one may think because for large γ it is always difficult to judge whether a distribution decays in the tail as an exponential or as a power law with a large exponent.

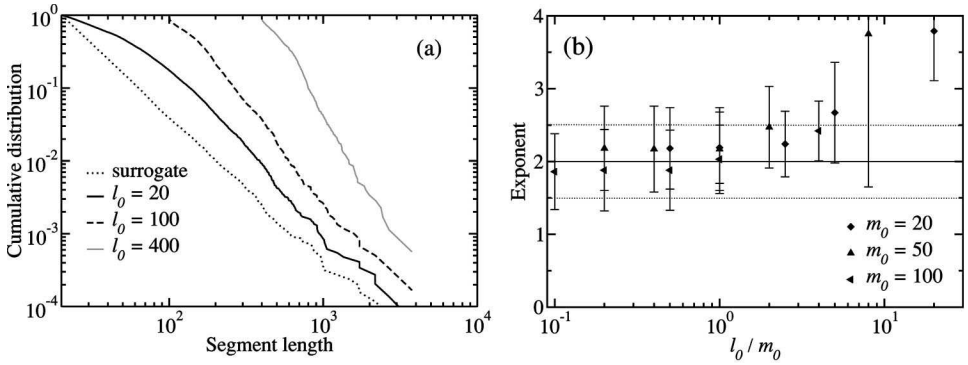

FIG. 3. (a) Cumulative distribution of segment lengths for $\gamma=2.0$, $m_0=20$, and $R=1$. For $\ell_0=20$ and $100<\ell<1000$, the exponent $\hat{\gamma}$ of the power law is close to γ. For $\ell_0=100$, we also find $\hat{\gamma}\approx\gamma$. However, for $\ell_0=400$, the algorithm fails to split the time series correctly for $\ell<1000$. Moreover, note that even though the exponent estimate is correct, the algorithm yields segments that are longer than the ones in the surrogate time series. (b) Dependence of $\hat{\gamma}$ on ℓ_0 and m_0 for $\gamma=2.0$. For $\ell_0/m_0<4$, the algorithm yields segments with the correct statistical properties.

TABLE II. Estimated exponents $\hat{\gamma}$ for $\gamma=2.0$. The mean and the standard deviation of the exponents are calculated for the ranges indicated inside parentheses using Eq. (B1). The column labeled ''input'' presents exponent estimates obtained from the segment lengths used to generate the surrogate time series. We find $\hat{\gamma}\approx\gamma=2.0$ for $\ell_0<4\ m_0$.

	ℓ_0					
m_0	10	20	50	100	400	Input
20	2.2 ± 0.6	2.2 ± 0.5	2.2 ± 0.5	2.4 ± 0.7	3.8 ± 0.7	2.2 ± 0.1
	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>200)$	$(\ell>800)$	
50	2.2 ± 0.6	2.2 ± 0.6	2.2 ± 0.6	2.5 ± 0.6	3.8 ± 2.1	2.1 ± 0.4
	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>200)$	$(\ell>1000)$	
100	1.9 ± 0.5	1.9 ± 0.6	1.9 ± 0.6	2.0 ± 0.5	2.4 ± 0.4	2.0 ± 0.4
	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>200)$	$(\ell>1000)$	

IV. ROBUSTNESS OF THE ALGORITHM WITH REGARD TO NOISE

A. Amplitude of fluctuations around a segment's mean

Another factor that may affect the performance of the segmentation algorithm of Bernaola-Galván and co-workers is the amplitude of the fluctuations within a segment. It is plausible that greater noise amplitudes will increase the difficulty in identifying the boundaries of neighboring segments. Thus, we next analyze the effect of the amplitude of the noise for surrogate time series.

Figure 4 demonstrates that for large R, the segmentation algorithm yields few short segments. This result arises from the concatenation of neighboring segments with means that become statistically indistinguishable due to the large value of σ_ϵ.

We show in Fig. 5 the cumulative distributions of segment lengths for $\gamma=1.0$, $m_0=20$, and for different values of R. For large ℓ and $R\leqslant3$, we find $\hat{\gamma}\approx\gamma$, while for $R=4$ the algorithm becomes ineffective at extracting the stationary segments in the time series. It is visually apparent that for $R=4.0$ the fluctuations within a segment are so much larger than the jumps between the means of the stationary segments that the segmentation becomes unable to parse the different segments; cf. Fig 4(d).

In Fig. 6, we show $\hat{\gamma}$ for different values of R. For $\gamma=1.0$, we estimate $\hat{\gamma}\approx\gamma$ for $0<R<4$. In contrast, for $\gamma=2.0$, we estimate $\hat{\gamma}\approx\gamma$ only for $0<R<1.5$. When R increases, the segmentation algorithm is unable to cut the segments because the greater amplitude of the fluctuations inside a segment decreases the significance of the differences between regions of the time series. This effect yields very large segments, which results in very small estimates of $\hat{\gamma}$. This effect is even stronger for $\gamma=3.0$, for which we find $\hat{\gamma}\approx\gamma$ only for $0<R<0.6$.

B. Spike noise

In many data-collection situations, one obtains data with spike noise. This type of noise is typically due to instrumentation failure or due to deficiencies of the algorithm used for preprocessing the data, and in many situations it may be impossible to fully clean the data of such noise. Due to its ubiquity, it is important to quantify the performance of the segmentation algorithm for signals with spike noise. Thus, we next analyze the effect of spike noise on the performance of the segmentation algorithm.

We generate surrogate time series as before and then for each t replace, with probability p, the original value of $x(t)$

TABLE III. Estimated exponents $\hat{\gamma}$ for $\gamma=3.0$. The mean and the standard deviation of the exponents are calculated for the ranges indicated inside parentheses using Eq. (B1). The column labeled ''input'' presents exponent estimates obtained from the segment lengths used to generate the surrogate time series. For this value of γ, one finds that a small m_0 leads to a clear underestimation of γ. Note that the standard deviation of $\hat{\gamma}$ becomes larger, indicating the difficulty in obtaining $\hat{\gamma}$ accurately. Also noteworthy is the fact that because γ is so large, the range of segment lengths m drawn becomes much reduced. This implies that if one sets $\ell_0=400$ one is unable to properly estimate γ.

	ℓ_0					
m_0	10	20	50	100	400	Input
20	2.4 ± 0.5	2.3 ± 0.5	2.5 ± 0.5	2.7 ± 0.5	4.3 ± 1.4	3.1 ± 0.5
	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>200)$	$(\ell>1000)$	
50	3.0 ± 0.8	3.0 ± 0.8	3.0 ± 0.9	3.4 ± 0.6	5.9 ± 1.3	3.0 ± 0.6
	$(\ell>100)$	$(\ell>100)$	$(\ell>100)$	$(\ell>200)$	$(\ell>1000)$	
100	2.8 ± 0.9	2.8 ± 0.8	2.8 ± 0.9	2.8 ± 0.8	5.2 ± 1.7	2.9 ± 0.7
	$(\ell>200)$	$(\ell>200)$	$(\ell>200)$	$(\ell>200)$	$(\ell>1000)$	

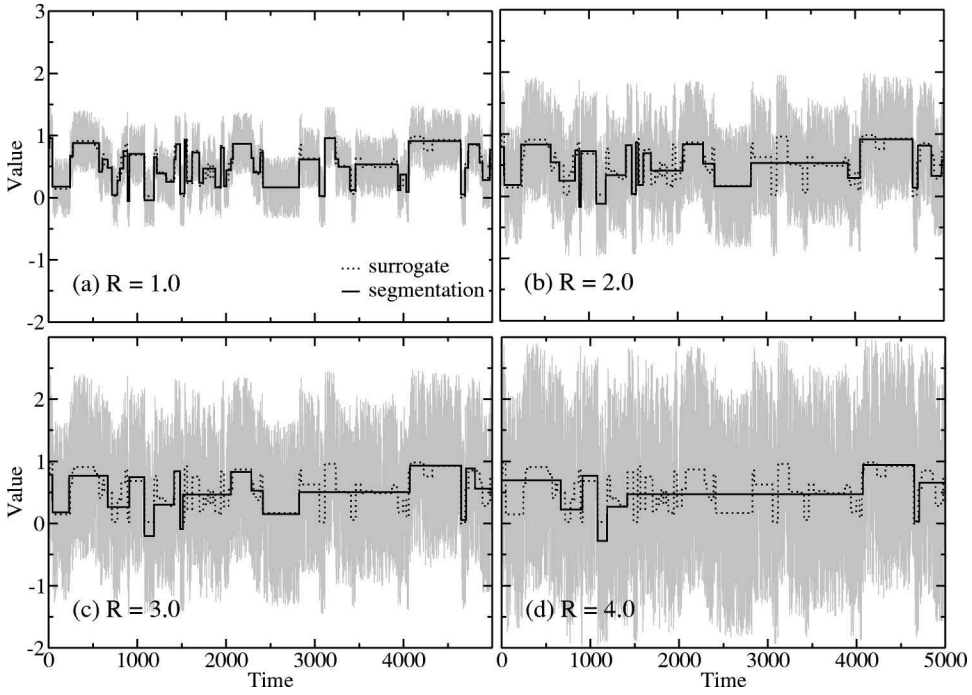

FIG. 4. Surrogate time series for different amplitudes of the fluctuations within the segments. The surrogate time series were generated with the same sequence of random numbers and with $\gamma=1.0$, $m_0=20$, and (a) $R=1.0$, (b) $R=2.0$, (c) $R=3.0$, and (d) $R=4.0$. The segmentation was performed with $\ell_0=20$. It is visually apparent that the segmentation algorithm yields longer segments as R increases. This fact arises from the fact that the statistical test cannot distinguish two neighboring segments whose difference in means is much smaller than σ_ϵ.

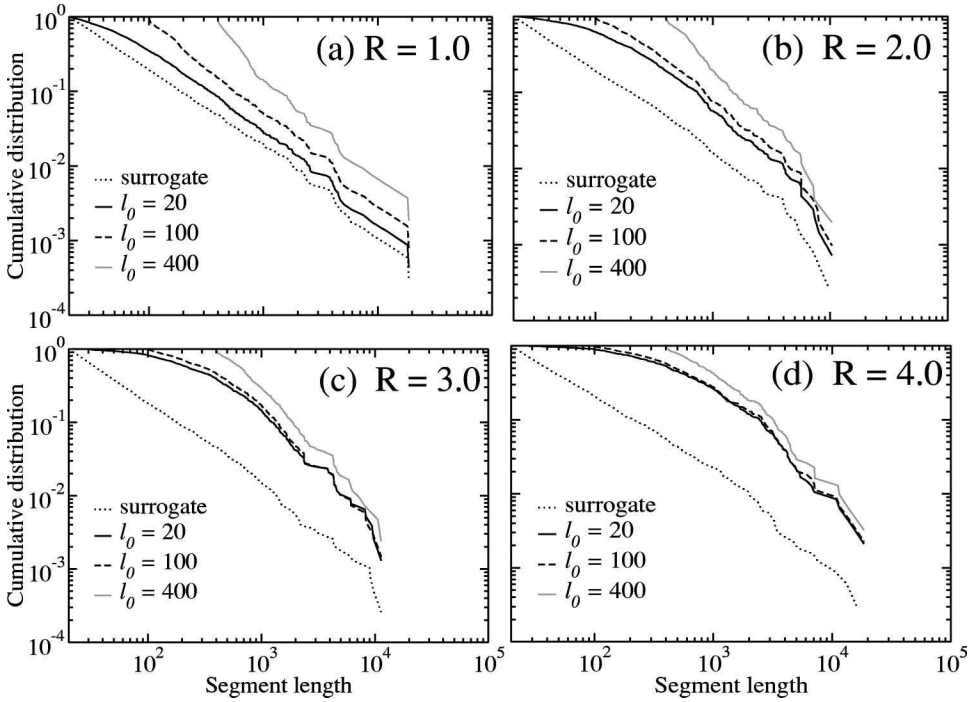

FIG. 5. Cumulative distribution of segment length obtained with the segmentation algorithm for different amplitudes of the noise. As in Fig. 1, the time series analyzed were generated with parameter for $\gamma=1.0$, $m_0=20$, and (a) $R=1.0$, (b) $R=2.0$, (c) $R=3.0$, and (d) $R=4.0$. For $R\leqslant 3$, the tails of the distributions decay as power laws. For $R=4.0$, it is difficult to discriminate whether our the tails of distribution conform to exponential or power-law decays. Note that as R increases, the dependence of the functional form of the distributions on ℓ_0 decreases appreciably.

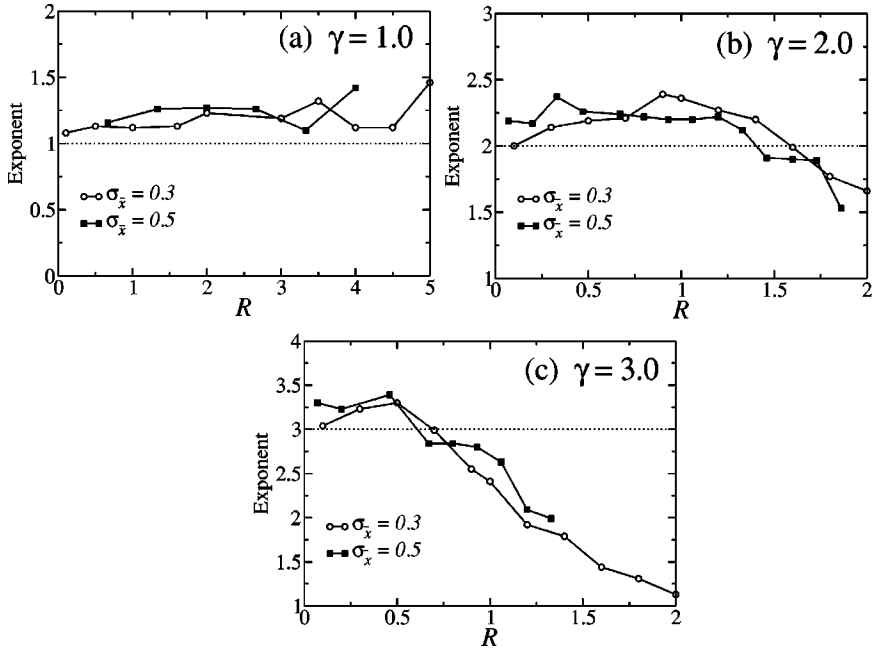

FIG. 6. Dependence of $\hat{\gamma}$ on the ratio $R=\sigma_\epsilon/\sigma_{\bar{x}}$ for $m_0=20$, and three distinct values of γ: (a) $\gamma=1.0$, (b) $\gamma=2.0$, and (c) $\gamma=3.0$. The segmentation was performed with $\ell_0=20$. The different curves in each plot correspond to different values of $\sigma_{\bar{x}}$. For $\gamma=1.0$, we find $\hat{\gamma}=1.2$ for $0<R<4.0$, indicating that the algorithm is robust against increases in R. For $\gamma>1.0$, we find that the impact of an increasing R on the performance of the algorithm becomes more and more marked. Specifically, for $\gamma=2.0$, we find $\hat{\gamma}\approx\gamma$ for $R<1.5$, while for $\gamma=3.0$, we find $\hat{\gamma}\approx\gamma$ for $R<0.6$.

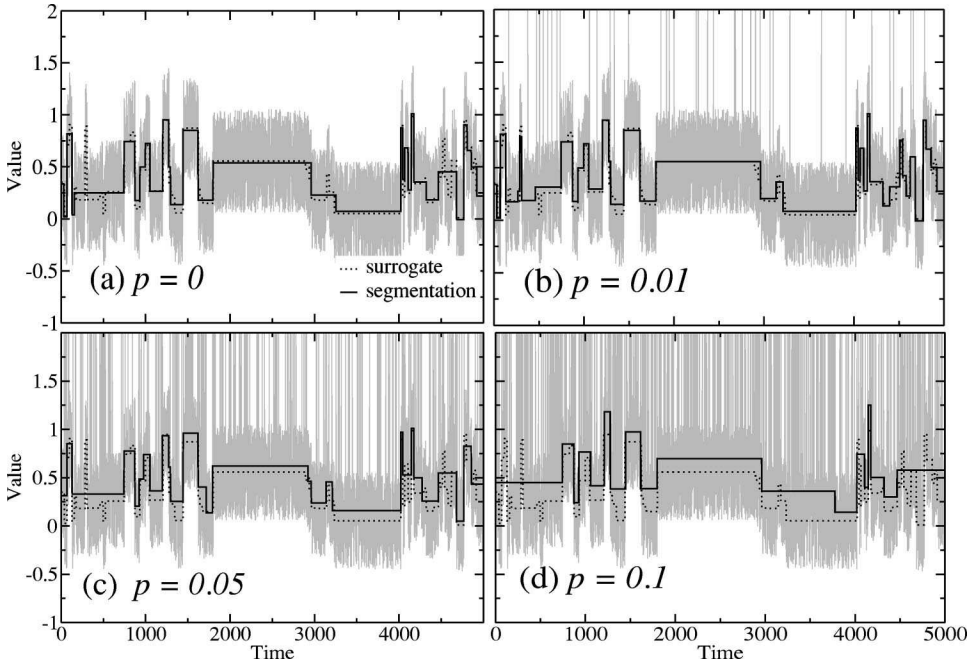

FIG. 7. Surrogate time series with uncorrelated spike noise for $m_0=20$, $\sigma_\epsilon=0.3$, $\gamma=1.0$, and (a) $p=0$, (b) $p=0.01$, (c) $p=0.05$, and (d) $p=0.1$. The segmentation was performed with $\ell_0=20$. For all values of p considered, the application of the segmentation algorithm yields segments that, in a coarse-grained way, match well the segments in the surrogate data.

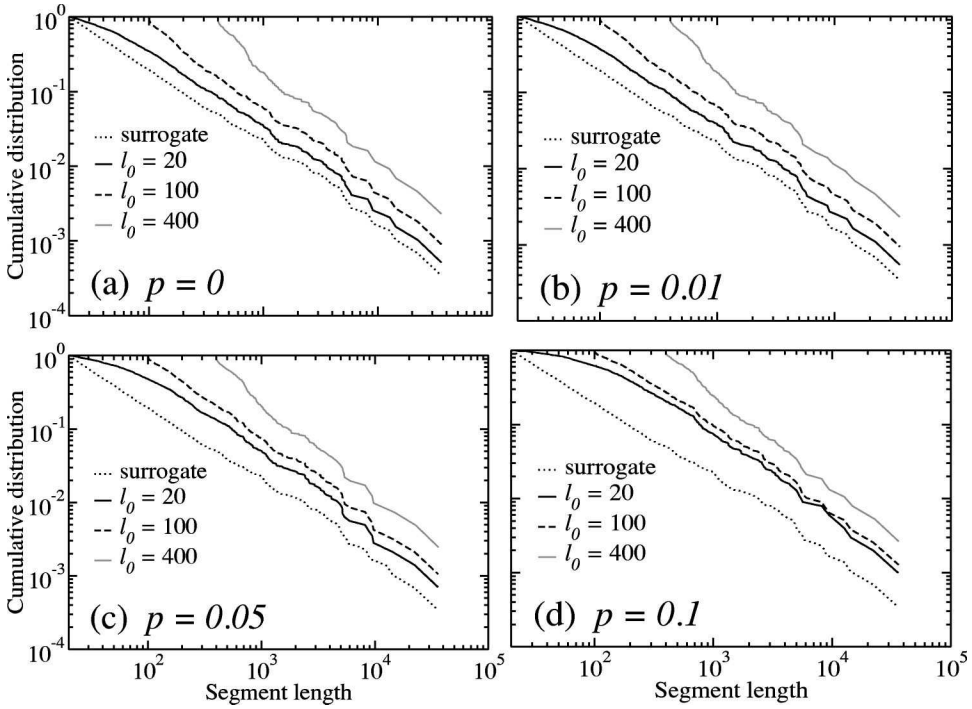

FIG. 8. Cumulative distribution of segment lengths for surrogate time series with $\gamma=1.0$, $m_0=20$, $R=1$, and for different densities p of the spike noise: (a) $p=0.0$, (b) $p=0.01$, (c) $p=0.05$, and (d) $p=0.1$. The value of $x(t)$ is set at 2.0 for a spike. It is visually apparent that even for $p=0.1$ all the distributions decay as power laws for $\ell>200$ and that the slopes in the log-log plots are similar to the slope of the input distributions, i.e., the segmentation algorithm yields segments with the correct statistical properties even in the presence of strong spike noise.

by $x(t)=2$. The effect of this procedure is illustrated in Fig. 7 for four distinct values of p. The figure also suggests that the segmentation algorithm yields a good coarse-grained description of the surrogate time series for p as large as 0.1. This result suggests that the algorithm is robust to the existence of uncorrelated spike noise in the data (Fig. 8).

V. CORRELATED NOISE

In this section, we investigate the effect of long-range correlations in the fluctuations around the segment's mean on the performance of the segmentation algorithm. This study is particularly important because real-world time series often display long-range power-law decaying correlations.

A. Segmentation of correlated noise with no segments

We generate temporally-correlated noise whose power spectrum decays as $S(f)\sim f^{-\beta}$ [21]. These surrogate time series consists of 60 000 points with mean zero. In Fig. 9, we display the cumulative distribution of segment lengths for time series generated with $\beta=0.3$, 0.5, and 1.0. For uncorrelated time series—i.e., $\beta=0$—we confirmed a single segment, as expected. For small but nonzero β, the algorithm

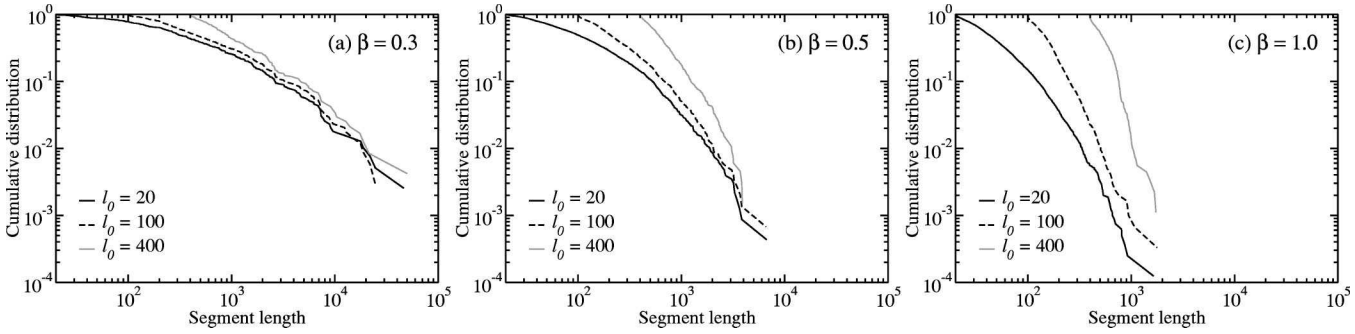

FIG. 9. Segmentation of Gaussian distributed long-range correlated noise with power spectrum $S(f)\sim f^{-\beta}$ for (a) $\beta=0.3$, (b) $\beta=0.5$, and (c) $\beta=1.0$. We generate time series with 60 000 data points according to the modified Fourier filtering method of Ref. [21]. We also consider the case $\beta=0$ for which we find a single segment in the time series. It is visually apparent that as β increases, which indicates an increase in the strength of the correlations, the distribution $P(>\ell)$ also decays more rapidly. Notably, the decay is never consistent with a power-law dependence.

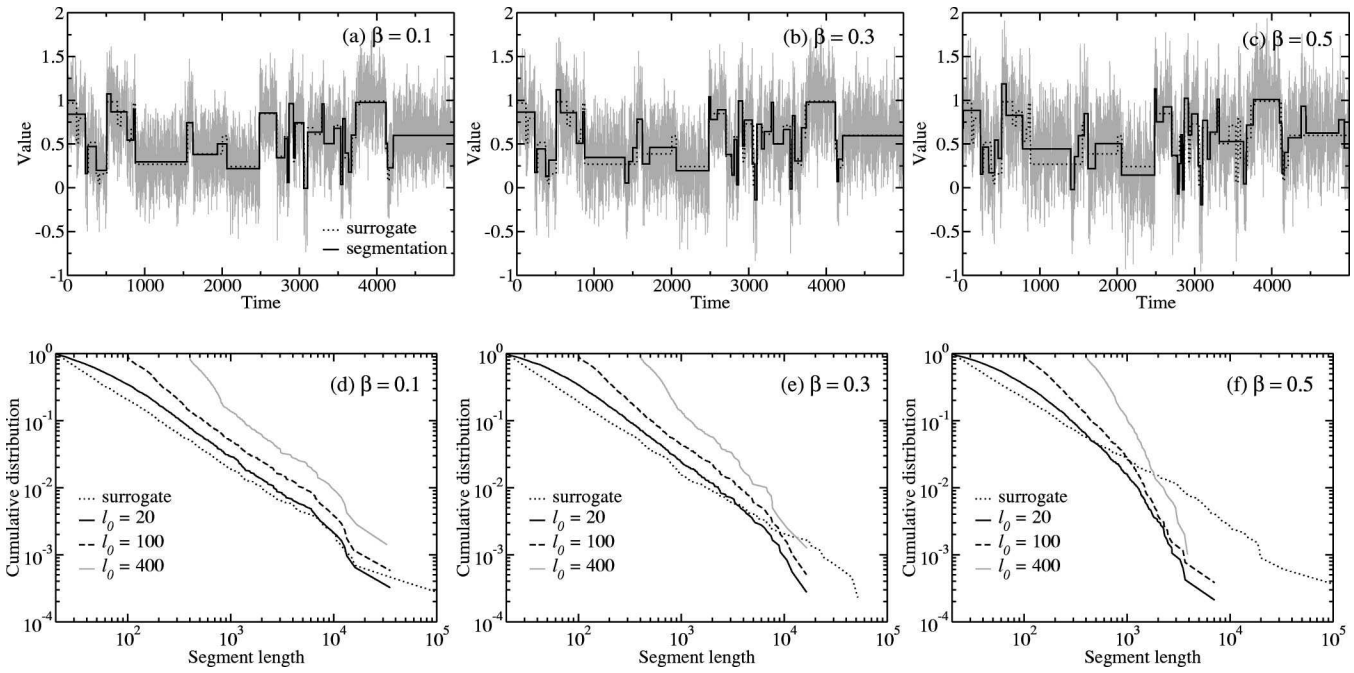

FIG. 10. Segmentation of surrogate time series obtained according to the procedure described in Sec. II B but with $\{\epsilon_i(k_i)\}$ having long-range correlations. The fluctuations $\{\epsilon_i(k_i)\}$ were generated with the modified Fourier filtering method of Ref. [21]. We show surrogate time series obtained with $\gamma=1.0$, $m_0=20$, $R=1$ and power spectra exponents: (a) $\beta=0.1$, (b) $\beta=0.3$, and (c) $\beta=0.5$. The segmentation was performed with $\ell_0=20$. For $\beta=0.1$ and 0.3, the results of the segmentation algorithm closely track the segments in the surrogate time series. For $\beta=0.5$, it is visually apparent that short segments are identified correctly while long segments are cut multiple times, indicating that the algorithm ''judges'' the noise within a segment as a nonstationary time series. This is to be expected for large β since the correlated noise in a segment is not stationary. Cumulative distributions of segment length for (d) $\beta=0.1$, (e) $\beta=0.3$, and (f) $\beta=0.5$. The distributions confirm the visual impression obtained from (a)–(c).

still identifies some very long segments, and one finds a slow decaying distribution of segment lengths. As the value of β increases, so does the strength of the correlations resulting in nonstationary time series with regions of distinct means, which the segmentation algorithm is able to identify. Thus, as β increases the distribution of segment lengths decays more rapidly.

We have not attempted to determine the functional form of $P(>\ell)$ as a function of β. The important result to retain from this portion of our analysis is that correlations lead to nonstationarity of the time series, which in turn result in there being regions with different means that the segmentation algorithm is able to identify. Notably, the decay of $P(>\ell)$ is never consistent with a power-law dependence for the values of β considered.

B. Segmentation of a time series with segments with different means and correlated fluctuations

A question prompted by the results presented in Sec. V A is: ''What happens if the time series has correlated fluctuations superimposed on a nonstationary sequence of segments?'' In order to answer this question, we analyze surrogate time series obtained according to the process described in Sec. II B, but in which $\epsilon_i(k_i)$ has long-range correlations.

We show in Figs. 10(a)–10(c) typical surrogate time series generated with $\gamma=1.0$, $m_0=20$, $R=1.0$, $\ell_0=20$, and different temporal correlations: (a) $\beta=0.1$, (b) $\beta=0.3$, and (c) $\beta=0.5$. The figure suggests that the segmentation algorithm can correctly parse the short segments but that long-segments get cut multiple times, especially for $\beta=0.5$. This result is to be expected because the strong correlations in the noise lead to marked changes in the mean.

In Figs. 10(d)–10(f), we plot the cumulative distributions of segment lengths for $\beta=0.1$, 0.3, and 0.5. The data confirm quantitatively the visual impression given by Figs. 10(a)–10(c)—i.e., longer segments get cut multiple times. In particular, for $\beta=0.5$ the distributions clearly deviate from the a power-law decay, independent of the selection of ℓ_0. Note that this fact should not be interpreted as a shortcoming of the algorithm; for large β, correlated noise inside a long segment *is* nonstationary, so that the algorithm is cutting a *nonstationary signal*.

VI. DISCUSSION

In this paper we analyzed nonstationary surrogate time series with different statistical properties in order to investigate the validity of the segmentation algorithm of Bernaola-Galván and co-workers [14]. Our results demonstrate that this heuristic segmentation algorithm can be extremely effective in determining the stationary regions in a time series provided that a few conditions are fulfilled. First, one must have enough data points in the time series to yield a large number of segment lengths, otherwise one will not be able to

reach the asymptotic regime of the tail of the distribution of segment lengths.

Second, the ratio of the amplitude of the fluctuations within a segment to the typical jump between the means of the stationary segments must be relatively small (less than about 0.6) in order for one to trust the output of the segmentation algorithm. This concern contrasts with the case of spike noise in the data which affects the performance of the segmentation algorithm only weakly.

Finally, if there are long-range temporal correlations of the fluctuations around the mean of the segment, then the segmentation algorithm correctly cuts the time series into the stationary segments for small β. However, for $\beta > 0.3$ the fluctuations inside long segments become nonstationary, which results in the algorithm detecting many ''stationary'' durations inside these long segments.

Our analysis provides a number of clear guidelines for using the segmentation algorithm of Bernaola-Galván *et al.* [14] effectively.

(1) One must perform the segmentation for a number of different values of ℓ_0 in order to identify the region for which the tails of the distributions of segment lengths reach the asymptotic scaling behavior. (Note: If $\hat{\gamma}$ is large, then the estimation error can be quite considerable, especially if m_0 is small.)

(2) One must calculate the ratio between the standard deviation of the mean value of the segment and the standard deviation of the fluctuations within a segment after performing the segmentation. If the $R > 0.6$, then there is the possibility that $\hat{\gamma}$ is considerably underestimating γ.

ACKNOWLEDGMENTS

We thank S. Havlin, P. Ch. Ivanov, and especially P. Bernaola-Galván for discussions. We thank NIH/NCRR (P41 RR13622) and NSF for support. L.A.N.A. acknowledges a Searle Leadership Fund Award.

APPENDIX A: PERFORMANCE OF THE ALGORITHM FOR FIXED SEGMENT LENGTHS

In order to quantify the dependence of the performance of the segmentation algorithm on the parameter ℓ_0 and to identify the algorithm's length resolution, we analyze time series comprising segments of fixed length. We concatenate segments with constant length m_0 with alternating means of 0.0 and 1.0. We then add fluctuations to those segments with a standard deviation of 0.3. We define the fraction of successfully split segments,

$$Q \equiv \frac{\#\text{ of segments correctly identified by the algorithm}}{\#\text{ of segments in surrogate data}}, \tag{A1}$$

where $Q = 1.0$ corresponds to perfect segmentation.

In Fig. 11, we plot Q as a function of m_0 for segmentations performed with different values of ℓ_0 . For $m_0 < 20$, the segmentation algorithm does not yield the correct segments in the surrogate time series even though the segment's means are quite different. This result suggests that the resolution of the segmentation algorithm is ≈ 20. We also find that for $\ell_0 = 5$, the algorithm splits the time series into too many segments. These results suggest that for optimal performance $10 < \ell_0 \leq m_0$.

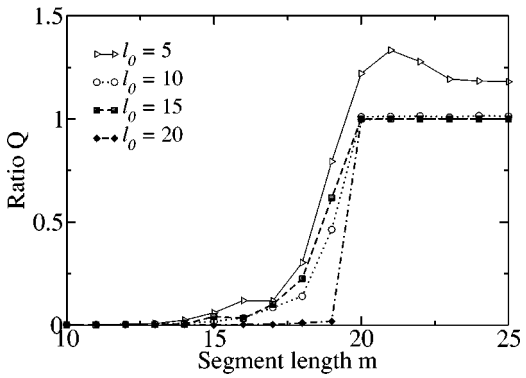

FIG. 11. Algorithm's resolution with regard to ℓ_0 and m. We generate surrogate time series which are constructed by the alternated concatenation of two types of segments with length m_0: one with mean 0.0 and standard deviation 0.3 and another with mean 1.0 and standard deviation 0.3. We plot Q, which is defined in Eq. (A1), as a function of m_0 for different values of ℓ_0. Our results indicate that the segmentation algorithm works best when $m_0 > 20$ and $10 < \ell_0 \leq m_0$.

APPENDIX B: ESTIMATION OF $\hat{\gamma}$

To estimate the exponent $\hat{\gamma}$ characterizing the power-law decay of the tail of $P(>\ell)$, we first calculate local estimates for segment lengths around ℓ_n,

$$\hat{\gamma}(\ell_n) \equiv \frac{\ln[P(>\ell_{n+\tau})] - \ln[P(>\ell_n)]}{\ln(\ell_{n+\tau}) - \ln(\ell_n)}. \tag{B1}$$

This expression is a generalization of the Hill estimator [22], for which $\tau = 1$. In our analysis, we have used $\tau = 5$. In Fig. 12 we present the local estimates $\hat{\gamma}(\ell_n)$ for the data shown in Fig. 2(a). The black lines indicate the values of the exponent for segment lengths around ℓ, while the dashed lines indicate the values of the exponent of the power law in the distribution of the surrogate time series.

We then estimate $\hat{\gamma}$ by calculating the mean of all local estimates in the region where $\hat{\gamma}(\ell_n)$ is approximately constant. That is, we omit the values in the regions corresponding to the initial exponential decay of the distribution and those corresponding to the truncation due to finite sample size.

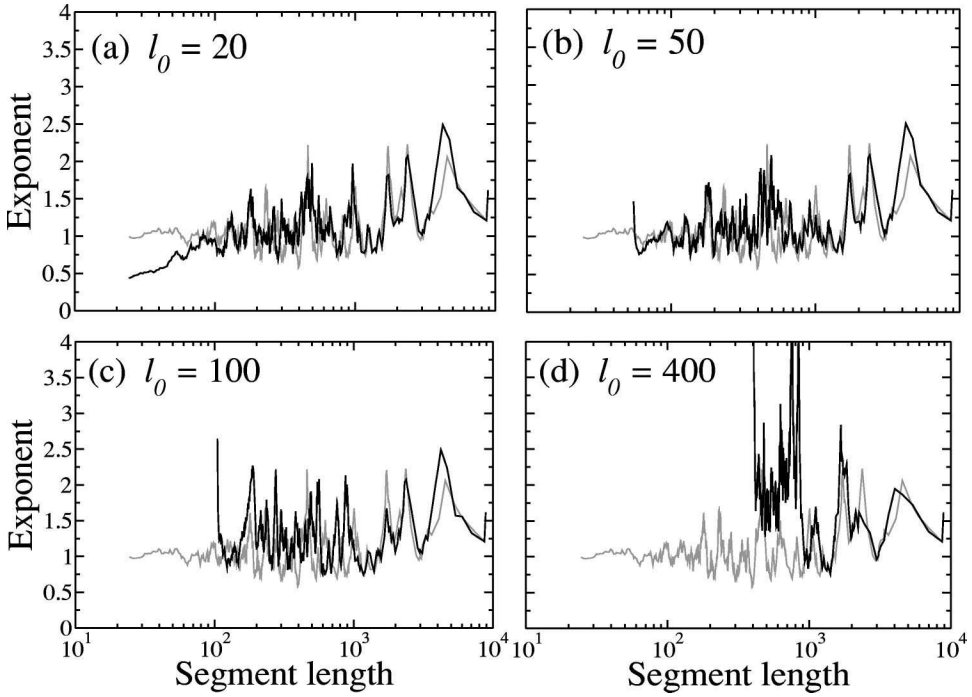

FIG. 12. Local estimate of the exponent $\hat{\gamma}$ characterizing the power-law decay of the distribution $P(>\ell)$ for (a) $\ell_0=20$, (b) $\ell_0=50$, (c) $\ell_0=100$, and (d) $\ell_0=400$. These results are obtained for surrogate time series generated with parameters $\gamma=1.0$, $m_0=20$, and $R=1.0$. The local estimate of $\hat{\gamma}$ is calculated using Eq. (B1). We indicate the local estimate of $\hat{\gamma}$ by the full black line and the local estimate of the exponent in the surrogate data γ by the dotted gray line. For $\ell_0=20$ and $\ell>100$, $\hat{\gamma}(\ell_n)$ closely tracks $\gamma(\ell_n)$ and both curves have averages close to one. For $\ell_0=100$ and 400, the segmentation results track the properties of the surrogate data only for $\ell>\ell_0$, while for smaller ℓ the segmentation results overestimate the value of γ.

[1] R. L. Stratonovich, *Topics in the Theory of Random Noise* (Gordon and Breach, New York, 1981), Vol. 1.

[2] P.Ch. Ivanov, M.G. Rosenblum, C.-K. Peng, J.E. Mietus, S. Havlin, and H.E. Stanley, Nature (London) **383**, 323 (1996); P.Ch. Ivanov, L.A.N. Amaral, A.L. Goldberger, S. Havlin, M.G. Rosenblum, Z.R. Struzik, and H.E. Stanley, *ibid.* **391**, 461 (1999).

[3] A. Bunde, S. Havlin, J. Kantelhardt, T. Penzel, J.-H. Peter, and K. Voigt, Phys. Rev. Lett. **85**, 3736 (2000).

[4] A.L. Goldberger, L.A.N. Amaral, J.M. Hausdorff, P.Ch. Ivanov, C.-K. Peng, and H.E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **99**, 2466 (2002).

[5] H.E. Stanley, L.A.N. Amaral, S.V. Buldyrev, P. Gopikrishnan, V. Plerou, and M.A. Salinger, Proc. Natl. Acad. Sci. U.S.A. **99**, 2561 (2002).

[6] T. Musha and H. Higuchi, J. Appl. Phys. **15**, 1271 (1976).

[7] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, IEEE/ACM Trans. Netw. **2**, 1 (1995).

[8] V. Paxson and S. Floyd, IEEE/ACM Trans. Netw. **3**, 226 (1996).

[9] M. Crovella and A. Bestavros, IEEE/ACM Trans. Netw. **5**, 835 (1997).

[10] M. Takayasu, H. Takayaus, and T. Sato, Physica A **233**, 924 (1996); M. Takayasu, H. Takayasu, and K. Fukuda, *ibid.* **277**, 248 (2000).

[11] *Unsolved Problems of Noise*, edited by D. Abbott and L. B. Kish (Melville, New York, 1999).

[12] C.-K. Peng, S. Havlin, H.E. Stanley, and A.L. Goldberger, Chaos **5**, 82 (1995); Z.R. Struzik, Fractals **8**, 163 (2000).

[13] Consider a time series with N points, then the problem of segmenting this time series in up to N segments is equivalent to the problem of placing each of N balls in one of N boxes. This is a standard combinatorial problems, and we know that there are N^N distinct possibilities. Hence the problem of finding which of those N^N possibilities is optimal will necessarily take a time that scales as $O(N^N)$.

[14] P. Bernaola-Galván, P.Ch. Ivanov, L.A.N. Amaral, and H.E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001).

[15] Student's *t*-test is used to test for differences in means when variances are assumed to be equal. The assumptions used in deriving the statistics of t are (i) both samples are Gaussian distributed with equal variance and (ii) the samples are drawn independently and at random from each populations. The surrogate time series we analyze in this paper are drawn from Gaussian or uniform distributions, so the first assumption is well verified in this case. However, in other situation such as the analysis of financial time series [P. Gopikrishnan, M. Meyer, L.A.N. Amaral, and H.E. Stanley, Eur. Phys. J. B **3**, 139 (1998)], the distribution is quite different from a Gaussian. Cases such as this suggest the need for a distribution-free statistical test [B. Efron and G. Gong, Am. Stat. **37**, 36 (1983)] to be incorporated into the segmentation algorithm discussed here.

[16] K. Fukuda, L. A. N. Amaral, and H. E. Stanley (unpublished).

[17] W. Feller, *An Introduction to Probability Theory and Its Application*, 2nd ed. (Wiley, New York, 1971), Vol. 1.

[18] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1988).

[19] The expression presented in Eq. (3) was obtained according to the following method. We generated a large number of sequences comprising N uncorrelated random numbers, first following a uniform distribution in the interval [0, 1], and then following a Gaussian distribution with mean zero and standard deviation one. We then used the segmentation algorithm on these sequences of random number and calculated the probability of obtaining a given value of $t_{\max}$. Finally, we fitted that data to the expression presented in Eq. (3) and estimated the parameters η and δ.

[20] We use the notation $\hat{\gamma}$ as the abbreviation form of $\hat{\gamma}(\ell_0, m_0)$ when the context is clear. Also, we use $P(>\ell)$ instead of $P_{\ell_0,m_0}(>\ell)$.

[21] H.A. Makse, S. Havlin, M. Schwartz, and H.E. Stanley, Phys. Rev. E **53**, 5445 (1996).

[22] B.M. Hill, Ann. Stat. **3**, 1163 (1975).