PHYSICAL REVIEW E VOLUME 49, NUMBER 2 FEBRUARY 1994

Analyzing long-range correlations in finite sequences

N. Shnerb and E. Eisenberg
Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel
(Received 13 October 1993)

We study the fluctuations in the correlation exponent obtained for correlated and noncorrelated sequences by mapping them into a one-dimensional random-walk model. We investigate, both numerically and analytically, the widely used technique of averaging over overlapping samples. An explicit quantitative measure for the reduction of the sample-to-sample fluctuations in the exponent due to this process is given, and the limits for which the results obtained are reliable are discussed.

PACS number(s): 06.50.−x, 05.40.+j, 72.70.+m

I. INTRODUCTION

Recently, it was suggested that the existence of long-range correlations in one-dimensional strings can be checked by mapping them into a one-dimensional random-walk (RW) model, and studying the fluctuations of such walks [1-3]. This method has the advantage over the traditional power spectrum or direct calculations of the string correlation in that it yields high quality scaling data. Generally, this method can be applied to any source of information, e.g., DNA data, texts, music scores, pictures, etc. The usual averaging techniques involve the partition of the sequence to be analyzed into many, equal length, parts, and averaging the results over them. The exponent characterizing the power-law decay of the correlations is well defined for infinite sequences; however, in most practical cases the accuracy of the analysis is limited by the length of the relevant available sequence, e.g., the length of the text or the nucleotide chain in the DNA, which is typically of order $10^5 - 10^6$. Hence, in order to get a good average over long enough samples, it is usual to consider overlapping samples. It is therefore of importance to investigate the effect of such a procedure on the exponent calculated. Recently, Peng *et al.* [3] have used scaling arguments to estimate the accuracy of the exponent obtained through this method, and suggested that the error in the exponent scales as the number of the nonoverlapping samples. The purpose of this work is to calculate the exact effect of the overlapping samples procedure on the accuracy of the exponent and to show that this averaging procedure has only a limited effect on the reduction of the sample-to-sample fluctuations of the relevant exponent. We suggest a quantitative expression for the region in which this method is reliable.

II. DESCRIPTION OF THE GENERAL METHOD

Once a code is chosen, any given string of data is mapped into a sequence of numbers $u(i)$, the codification of the ith character. Following Peng *et al.* [1] we interpret these numbers as steps of a (one dimensional) RW. Define, then, the RW position $f(l)$ after l steps as

$$f(l) = \sum_{i=1}^{l} u(i), \tag{1}$$

and the difference d_l over a distance l by

$$d_l = f(l_0 + l) - f(l_0). \tag{2}$$

The mean square fluctuation of d_l is then

$$C(l) = \langle d_l^2 \rangle - \langle d_l \rangle^2, \tag{3}$$

where the average is taken over different l_0. $C(l)$ is expected, in the limit of long-range correlations, to have the scaling form $C(l) \sim l^{2\alpha}$, where α is the exponent describing the power-law decay of the correlations.

As has been emphasized by Peng *et al.* [1], we have

$$C(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} \langle u(i)u(j) \rangle. \tag{4}$$

One therefore sees that this method involves averaging over many correlation functions, and hence it is more reliable than direct calculations. The value $1/2$ for α implies that the correlation function decays exponentially, i.e., the text is not correlated after some typical length. On the other hand, $\alpha = 1$ corresponds to the scale-invariant $1/f$ noise, also called the maximal complexity limit [4].

Usually, it is common to average over all possible values of l_0 in order to get the maximal number of samples. The question of the reliability of this method was raised recently by Peng *et al.* [3], who have applied a theoretical argument to derive a scaling relation for $\Delta\alpha$ as a function of the sequence length N and the sample length l, namely,

$$\Delta\alpha(l, N) \sim \left(\frac{l}{N}\right)^{1/2}. \tag{5}$$

In this work we give an explicit measure to the sample-to-sample fluctuations in $C(l)$ in terms of l, N, p, and β, where p is a parameter which describes the amount of overlap, and β is determined by the distribution function of each elementary segment. We find that this expression indeed has the scaling form of Eq. (5), with a prefactor which depends solely on p and β. This prefactor is a bounded function of the degree of overlap, and thus even in the limit of maximal overlapping samples the resulting reduction of the sample-to-sample fluctuations leads to reliable results only in a limited region.

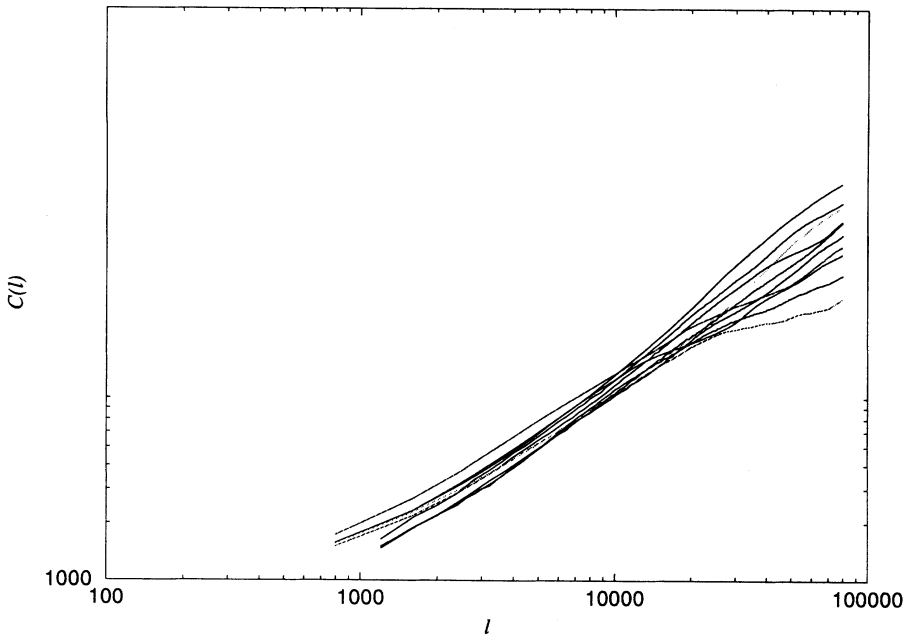

FIG. 1. $C(l)$ as a function of l, for ten different samples of 1 200 000 random ± 1 bits, where the sample size is 80 000 and the average is taken over overlapping samples.

III. FLUCTUATIONS IN THE EXPONENT — NUMERICAL EVIDENCE

In order to investigate the limits in which the use of overlapping samples is helpful, we have applied this method to sequences of noncorrelated random numbers. We have generated ten different sequences of 1.2×10^6 random ± 1 bits and then partitioned them into 1000 overlapping samples of length 80 000. The results are presented in Fig. 1. It is evident that the accuracy of the measurement of α, which is proportional to the slope of the curves presented in this figure, is very low for large l. However, the curves obtained by this method are smoother than the curves obtained by performing the average with nonoverlapping samples only (Fig. 2). This effect of the overlapping technique, namely the smoothness of the curve, has nothing to do with the reliability of the measurement of α, as implied by Fig. 1.

IV. ANALYTICAL APPROACH

Let us investigate a simple model for comparing the averaging over overlapping and nonoverlapping samples. In our model we consider a finite sequence of N noncorrelated random numbers $u(j)$, and test the correlations of samples of length l (we assume N to be an integer multiple of l). We then partition the sequence into $S = \frac{N}{l/2}$ segments, and define

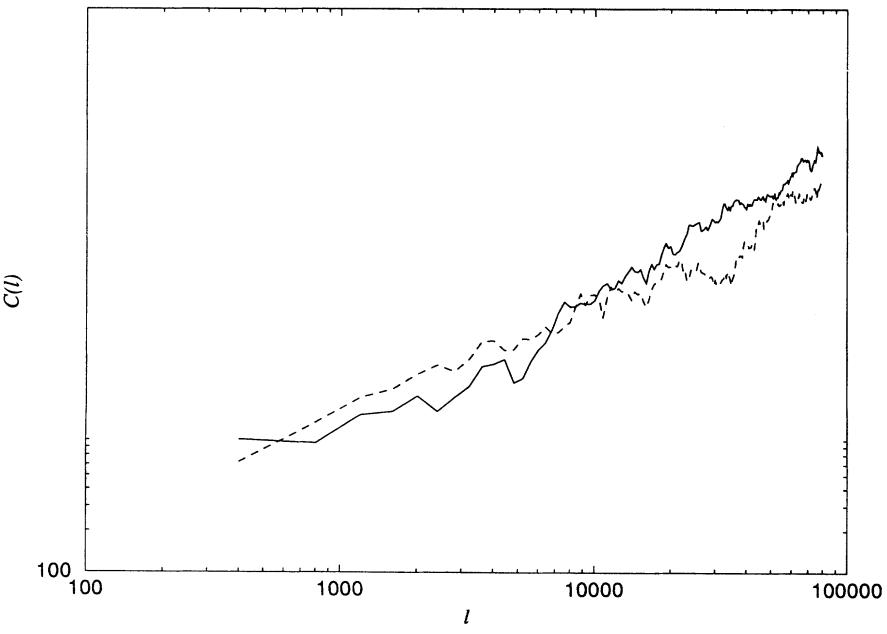

FIG. 2. $C(l)$ as a function of l, for two different samples of 1 200 000 random ± 1 bits, where the sample size is 80 000 and the average is taken over nonoverlapping samples only.

$$r_i = \sum_{j=(i-1)l/2+1}^{il/2} u(j), \tag{6}$$

where $1 \le i \le S$. When $u(j)$ are symmetric random variables, i.e., $\langle u(j) \rangle = 0$, so are the variables r_i. Consider now the averaging process in which the samples taken consist of the $S/2$ segments $i, i+1$ where $1 \le i \le S$ and i is odd, and therefore do not overlap. We compare the fluctuations of the correlation function $C(l)$ obtained by this process (denoted as case 1) with the results of the second option where the samples taken consist of the $S-1$ segments $i, i+1$ where $1 \le i \le S-1$ without any other constraints (case 2).

The statistical estimate of $\langle d_l \rangle$ and $\langle d_l^2 \rangle$ is given, for the first (nonoverlapping samples) method by

$$D = \frac{2}{S} \sum_{i \text{ even}}^{S-1} r_i + r_{i+1}, \quad D_2 = \frac{2}{S} \sum_{i \text{ even}}^{S-1} (r_i + r_{i+1})^2, \tag{7}$$

respectively. The correlation function C is then estimated statistically by $c = D_2 - D^2$, and its variance is given by $\Delta c^2 = \langle c^2 \rangle - \langle c \rangle^2$, where $\langle\,\rangle$ refers to averaging with respect to the probability distribution. For example, the average of the correlation function c is given by

$$\langle c \rangle = \left\langle \frac{2}{S} \sum_{i \text{ even}}^{S-1} r_i^2 + 2 r_i r_{i+1} + r_{i+1}^2 - \left(\frac{2}{S}\right)^2 \sum_{i,j \text{ even}}^{S-1} r_i r_j + r_i r_{j+1} + r_{i+1} r_j + r_{i+1} r_{j+1} \right\rangle. \tag{8}$$

Using the fact that $\langle r_i \rangle = 0$ and defining the quantity $\sigma^2 = \langle r_i^2 \rangle$, one can express $\langle c \rangle$ in terms of S and σ; under the assumption that there are no correlations in the original sequence, the only contributions for c arise from the terms proportional to even powers of r. For the nonoverlapping method it is then evident that the two terms

$$\sum_{i,j \text{ even}}^{S-1} r_i r_{j+1} + r_{i+1} r_j$$

do not contribute to the average since i, j are even but $i+1, j+1$ are odd, hence they cannot coincide. With this, the result is

$$\langle c \rangle = 2\sigma^2 \left(1 - \frac{2}{S}\right). \tag{9}$$

The same process of calculation for these quantities holds for case 2 with two differences; one must replace any occurrence of $\frac{2}{S}$ by $\frac{1}{S-1}$ (the inverse of the number of samples), and there are contributions to the average from the two terms which vanish in case 1. With this,

$$\langle c_{\text{overlap}} \rangle = 2\sigma^2 \left(1 - \frac{2}{S-1}\right). \tag{10}$$

In order to measure the fluctuations of the correlation function in terms of the moments of r, one should take the average of Δc^2 with the above rules. To the leading order in $\frac{1}{S}$, the results are

$$\frac{\Delta c^2_{\text{nonoverlap}}}{\Delta c^2_{\text{overlap}}} = \frac{S}{S-1} \left(1 + \frac{\sigma^4}{\langle r^4 \rangle}\right). \tag{11}$$

One sees that the overlapping method *reduces* (for the case $S \gg 1$) the sample-to-sample fluctuations, but this reduction is limited by the ratio $\beta = \frac{\sigma^4}{\langle r^4 \rangle}$. This ratio depends on the details of the distribution function of r; for the Gaussian distribution $\beta = 1/3$ and for the Poisson distribution $\beta = 1/4$. The relation between the two methods turns out to be of order unity.

The generalization of these calculations, with partition of each segment into two parts, to the case of partition into p parts, can be made. For the general case the results are

$$\Delta c^2_{\text{nonoverlap}} = \frac{1}{S}[\sigma^4(2p^3 - 3p^2) + p^2\langle r^4 \rangle], \tag{12}$$

$$\Delta c^2_{\text{overlap}} = \frac{1}{S-p+1}[\tfrac{1}{3}\sigma^4(4p^3 - 9p^2 + 2p) + p^2\langle r^4 \rangle]. \tag{13}$$

For the case of the Gaussian distribution (with $S \gg p$)

$$\frac{\Delta c^2_{\text{overlap}}}{\Delta c^2_{\text{nonoverlap}}} = \frac{2}{3} + \frac{1}{3p^2}, \tag{14}$$

i.e., the minimal ratio is $\frac{2}{3}$. This is also the result obtained for the common case where the average is performed using all possible values of l_0, such that $p = l \gg 1$, and the bits obey a *discrete* ± 1 distribution, for which $\sigma^4 = \langle r^4 \rangle$.

In view of this one sees that the accuracy in the measurement of $\langle c \rangle$ in case 2 (with overlap) is better than in case 1 (with only nonoverlapping samples); this, however, has only a limited effect. The relation between these cases is bounded by β, a quantity which is determined by the distribution function of the segments involved.

V. DISCUSSION

As we have shown, there is only a limited effect in the use of overlapping samples. According to Eq. (12) and the definition of Δc^2, it is evident that the effect of the overlapping average is a rescaling of the sample length, i.e., if, for the nonoverlapping case, one gets reliable results when the sequence of N numbers is partitioned into samples with length l such that $S = N/l$, with the overlapping method the sequence can be partitioned into segments of length $lf(p, \beta)$, where f is determined by Eqs. (12) and (13), without any loss of accuracy. However, as one sees, for most common cases f is bounded by 1.5, and the usage of the overlap method beyond this limit is problematic.

[1] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Nature (London) **356**, 168 (1992).

[2] A. Schenkel, J. Zhang, and Y.C. Zhang, Fractals **1**, 47 (1993).

[3] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, M. Simons, and H.E. Stanley, Phys. Rev. E **47**, 3730 (1993).

[4] Y.C. Zhang, J. Phys. (France) I **1**, 971 (1991).