A general statistical test for correlations in a finite-length time series

Jeffery A. Hanson and Haw Yang^{a)}

Department of Chemistry, University of California at Berkeley, Berkeley, California 94720, USA and
Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

(Received 27 March 2008; accepted 30 April 2008; published online 3 June 2008)

The statistical properties of the autocorrelation function from a time series composed of independently and identically distributed stochastic variables has been studied. Analytical expressions for the autocorrelation function's variance have been derived. It has been found that two common ways of calculating the autocorrelation, moving-average and Fourier transform, exhibit different uncertainty characteristics. For periodic time series, the Fourier transform method is preferred because it gives smaller uncertainties that are uniform through all time lags. Based on these analytical results, a statistically robust method has been proposed to test the existence of correlations in a time series. The statistical test is verified by computer simulations and an application to single-molecule fluorescence spectroscopy is discussed. © 2008 American Institute of Physics. [DOI: 10.1063/1.2931943]

I. INTRODUCTION

The dynamics of a system are usually characterized by the time correlation function of a variable q representing the system's state. If x is the experimental realization of q , then the time correlation function of interest is

$$C_{xx}(s) = \langle x(t)x(t+s) \rangle - \langle x \rangle^2, \quad (1)$$

where $\langle \cdots \rangle$ represents ensemble averaging. This time correlation can be recovered from a prolonged, time-averaged study of a single system if the system is ergodic and when the duration of observation T approaches infinity

$$C_{xx}(s) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)x(t+s)dt - \left[\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)dt \right]^2. \quad (2)$$

In practice, the correlation function is calculated from an experimentally recorded time series, $\{x_i\}$, $i=1, \dots, N$, with each data point discretely sampled at a time interval δt .

There are two common ways of calculating the empirical correlation function. The moving-average approach is appropriate for time series that are potentially aperiodic and is given by

$$\bar{C}_{xx}(s) = \frac{1}{N-j_s} \sum_{i=1}^{N-j_s} x_i x_{i+j_s} - \bar{x}^2, \quad (3)$$

where $j_s = s/\delta t$ and $\bar{x} = \sum_i x_i/N$. If the time series can be assumed to exhibit a periodicity of N , i.e., $x_i = x_{i+N}$, then the correlation function can be calculated using discrete Fourier transform (DFT) and its inverse operation (iDFT), $\bar{C}_{xx}(s) = \text{iDFT}\{\tilde{x}_i^* \tilde{x}_{i+j_s}\}$, where $\tilde{x}_i \equiv \text{DFT}\{x_i\}$ denotes the i th Fourier-transform element, and $*$ indicates complex conjugate. The Fourier transform method is equivalent to calculating

$$\bar{C}_{xx}^{\text{FT}}(s) = \frac{1}{N} \sum_{i=1}^N x_i x_{i+j_s} - \bar{x}^2. \quad (4)$$

Both Eqs. (3) and (4) can be considered as a finite-sample estimation of Eq. (2) and will contain uncertainties due to insufficient sampling of the time correlation function. This issue has been discussed by Bartlett for autoregression models, for which very general results have been obtained.¹ In the context of computer simulation, Zwanzig and Ailawadi² suggest that if the distribution in x is Gaussian and if the decay in correlation is approximately exponential, the statistical error due to finite correlation time decreases at a rate of $\sqrt{2\tau_0/T}$, where $\tau_0 \equiv 2 \int_0^\infty C_{xx}(s)^2 / C_{xx}(0)^2 ds$ and can be understood as the characteristic relaxation time of the system. Another way of looking at their results is to consider the number of samples along the correlation trace (in the form of $x_i x_{i+j_s}$) that are independent of each other, $n_{\text{ind}} = T/2\tau_0$. Therefore, the sampled trace converges to the true value with a rate $\propto 1/\sqrt{n_{\text{ind}}}$ as expected from standard statistical analysis.

When the relaxation dynamics deviate significantly from simple exponential, the situation becomes more complicated. Lu and Vanden Bout have used computer simulations to characterize uncertainties in measuring correlation functions in the form of a stretched-exponential decay, $C(s) \propto \exp[-(s/\tau_s)^\beta]$.³ They report significant bias and large uncertainties in the fit τ_s and β parameters when the trajectory length is less than 100-fold longer than τ_s . Even when $T \sim 1000\tau_s$, there is still about 10% relative uncertainties with respect to these two parameters.

The aforementioned pioneering works on error bounds of a correlation function, however, it does not lead to an obvious way of making the most basic statement: Whether there is a correlation in a given time series. In fact, one might argue that a rigorous identification for the existence of correlation should precede any attempt to fit the correlation function to a physical model. Here the existence of autocorrelation is formulated as a statistical test problem, for which

^{a)}Electronic mail: hawyang@berkeley.edu.

statistically robust approaches of characterizing them are discussed. In finding a suitable solution to this problem, one is mindful of following criteria: (1) One should be able to make a statement based on a single time series alone, rather than from a collection of them. (2) The method should be conservative and robust enough to work on short time series. (3) Since in most cases the true probability distribution for the observables $\{x_i\}$ is unknown, the method should be general and independent of any presumed model in x_i (a model-free test). The next two sections describe a practical solution to this problem, followed by a demonstration applying the method to single-molecule spectroscopy.

II. THEORETICAL CONSIDERATIONS

Consider a sequence of variables, $\{x(t_i)\}$, $i=1, \dots, N$, acquired at a fix time interval, $\delta t \equiv t_{i+1} - t_i$, from an experiment or from a computer simulation. For brevity, $x(t_i)$ are denoted by x_i in subsequent discussions. The problem at hand is therefore to test the null hypothesis (H_0) that there is no correlation in the x_i time series versus the alternative hypothesis (H_A) that there is correlation. When H_0 is true, the x_i 's are independently and identically distributed (iid) random variables.

Consequently, under H_0 , the correlation estimator $\bar{C}_{xx}(s)$ in Eq. (3) [or $\bar{C}_{xx}^{\text{FT}}(s)$ in Eq. (4)] is also a random variable. For each time lag, the correlation estimation carries an uncertainty, $\Delta\bar{C}_{xx}^{\text{(CI)}}(s)$, where CI stands for confidence interval. As an example, one may claim with 95% confidence that a time trace has no correlation when zero is contained within the error bounds for more than 95% of the n_s correlation estimations, $\bar{C}_{xx}(s) \pm \Delta\bar{C}_{xx}^{(0.95)}(s)$.⁴ The remaining task is therefore to calculate $\Delta\bar{C}_{xx}^{\text{(CI)}}(s)$.

While the following discussions are general and do not

require a particular model for the distribution of x , it is nevertheless instructive to illustrate the results using specific examples. For this purpose, one considers the most commonly encountered probability density models, the Gaussian (normal) and Poisson distributions

$$f_{\text{gaus}}(x; \mu, \sigma_x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x-\mu)^2}{2\sigma_x^2}\right], \quad (5)$$

where μ and σ_x are the mean and standard deviation, respectively, and

$$f_{\text{poi}}(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad (6)$$

where λ is the average count within each time interval.

A. Potentially aperiodic time series—Correlation calculated by moving average

In most practical applications, there will be a sufficient number of time lags (n_s) included in calculating the correlation function ($n_s > 25$) such that its converging rate can be approximated by that of a Gaussian distribution (the central-limit theorem). The expectation value for the variance of $\bar{C}_{xx}(s)$ is (see the Appendix for details)

$$\text{var}\{\bar{C}_{xx}(s)\} = h_1(N) + h_2(N, j_s), \quad (7)$$

where

$$\begin{aligned} h_1(N) &= \frac{1}{N^3} E\{X^4\} - \frac{4}{N^3} E\{X^3\} E\{X\} - \frac{2N+3}{N^3} E\{X^2\}^2 \\ &\quad - \frac{4(N^2 - N - 3)}{N^3} E\{X^2\} E\{X\}^2 \\ &\quad + \frac{2(2N-3)(N+1)}{N^3} E\{X\}^4, \end{aligned} \quad (8)$$

$$h_2(N, j_s) = \begin{cases} \frac{1}{N-j_s} E\{X^2\}^2 + \frac{2(N-2j_s)^2}{(N-j_s)^2} E\{X^2\} E\{X\}^2 - \frac{3N-5j_s}{(N-j_s)^2} E\{X^2\} E\{X\}^2, & 1 \leq j_s < (N/2); \\ \frac{1}{N-j_s} [E\{X^2\}^2 - E\{X\}^4], & (N/2) \leq j_s < N. \end{cases} \quad (9)$$

Equations (7)–(9) are the first major result of this article which give an analytical expression for lag-time dependent variance for a correlation function of finite length. Note that the variance can be evaluated without an explicit knowledge of the underlying distribution in $\{x\}$. This is an important criterion of the solution because the distribution in $\{x\}$ is generally unknown.

In practice, the expectation values in Eqs. (8) and (9) can be calculated from the time series directly by averaging. For example, $E\{X\} \approx \bar{x} = (1/N) \sum_{i=1}^N x_i$, $E\{X^2\} \approx \bar{x}^2 = (1/N) \sum_{i=1}^N x_i^2$, and so forth. The standard deviation and therefore confidence

intervals can be readily computed using Eq. (7). For example, defining the standard deviation at a particular time lag s as

$$\sigma_{xx}(s) \equiv \sqrt{\text{var}\{\bar{C}_{xx}(s)\}}. \quad (10)$$

In the remainder of the article, an overbar indicates that the quantity (such as $\bar{\sigma}_{xx}$) is computed using sample estimates ($\bar{x}$, $\bar{x}^2$, etc) for the expectation value. Then $\Delta\bar{C}_{xx}^{(0.95)}(s)$ can be computed by $1.96\sigma_{xx}(s)$. Furthermore, as expected, Eq. (7) shows that when N is large, the correlation function converges to zero with a rate of $\sim N^{-1/2}$. Greater

FIG. 1. Comparison of the analytical expression for expected variance in Eq. (7) with computer simulations. (a) The expected standard deviation for $N=400$ as a function of the correlation time lag j_s . (b) Comparison of the analytical expression (solid line) with the standard deviation at each time lag from 10 000 simulated Poisson series ($\times$), in which each trace contains 400 elements. (c) A typical correlation trace $C_{xx}(s)$ from an $N=400$ time series (solid line) is seen to be bound by the theoretical standard deviation (dashed line). The inset shows a zoomed-in view for $J_s=150$ to $J_s=250$.

variances are expected at larger time lags where $(N-j_s)$ is small; the increase in variance also scales with $(N-j_s)^{-1/2}$.

To gain a better picture about the solution, one considers the behavior of a time series composed of $\{x_i\}$ from an iid Poisson distribution. Thus, Eqs. (8) and (9) become

$$h_1^{\text{poi}}(\lambda, N) = -\left(\frac{4\lambda^3}{N} + \frac{2\lambda^2}{N^2} - \frac{\lambda}{N^3}\right),$$

$$h_2^{\text{poi}}(\lambda, N, j_s) = \begin{cases} \frac{\lambda^2 + 6\lambda^3}{N-j_s} - \frac{2N\lambda^3}{(N-j_s)^2}, & 1 \leq j_s < (N/2); \\ \frac{\lambda^2 + 2\lambda^3}{N-j_s}, & (N/2) \leq j_s < N. \end{cases}$$

Similarly, the results for Gaussian-distributed variables are

$$h_1^{\text{gaus}}(\mu, \sigma_x, N) = -\left(\frac{4\mu^2\sigma_x^2}{N} + \frac{2\sigma_x^4}{N^2} - \frac{8\mu^2\sigma_x^2}{N^3}\right),$$

$$h_2^{\text{gaus}}(\lambda, N, j_s) = \begin{cases} \frac{\sigma_x^4 + 6\mu^2\sigma_x^2}{N-j_s} - \frac{2N\mu^2\sigma_x^2}{(N-j_s)^2}, & 1 \leq j_s < (N/2); \\ \frac{\sigma_x^4 + 2\mu^2\sigma_x^2}{N-j_s}, & (N/2) \leq j_s < N. \end{cases}$$

The manner in which the standard deviation $\sigma_{xx}(s)$ varies with the time lag s can be assessed by plotting $\sigma_{xx}(s)$ against the time lag index j_s . The results for Poisson distribution are displayed in Fig. 1(a), which shows that the standard deviation does not vary smoothly with time lag, nor does it increase in a monotonic manner. In particular, there is a sharp dip at $j_s=N/2$ —an interesting feature not obvious from the approximate expressions discussed in previous works.^{1,2} A closer examination of the derivation (cf. the Appendix) reveals that the dip is due to changes in the relative weighing

in the higher-order moments, e.g., $E\{X_i^2 X_j^2\}$ and $E\{X_i X_j^2 X_k\}$, when the time lag index exceeds $N/2$ [cf. Eqs. (A2) and (A3)].

This point is further illustrated by comparing the analytical results in Eqs. (7) and (10) with computer simulations. For this purpose, 10 000 traces of 400 iid Poisson variables with $\lambda=20$ were simulated using MATLAB (v7.3.0 with the Statistics Toolbox). The autocorrelation function of each trace was then calculated using Eq. (3). At each time lag index j_s , the standard deviation of $\bar{C}_{xx}(j_s)$ was calculated from the 10 000 traces. The results are displayed in Fig. 1(b), showing quantitative agreement between the analytical $\sigma_{xx}(s)$ and converged simulations.

For Poisson time series of finite length, the correlation $\bar{C}_{xx}(s)$ tends to deviate from zero, particularly at larger time lags. An example is shown in Fig. 1(c). In this example, there are 154 out of 399 ($\sim 38.6\%$) correlation points that show a nonvanishing correlation amplitude exceeding 1 standard deviation, close to what one would expect for the error rates using one standard deviation to reject a null hypothesis. However, the nonuniform variance at different time lags and the possible extra correlation among different j_s introduced by the moving-average calculation for finite N make it less than ideal to use moving-average calculated correlation and its variance for tests of null correlation.

B. Periodic time series—Correlation calculated by discrete Fourier transform

Following the same procedure as that used for deriving $\text{var}\{\bar{C}_{xx}\}$, the variance for Fourier-transform calculated correlation function can be found to be

$$\begin{aligned} \text{var}\{\bar{C}_{xx}^{\text{FT}}(s)\} &= \frac{1}{N^3} E\{X^4\} - \frac{4}{N^3} E\{X^3\} E\{X\} \\ &\quad + \frac{(N-3)(N+1)}{N^3} E\{X^2\}^2 \\ &\quad - \frac{2(N^2-2N-6)}{N^3} E\{X^2\} E\{X\}^2 \\ &\quad + \frac{N^2-2N-6}{N^3} E\{X\}^4. \end{aligned} \quad (11)$$

Equation (11) is the second major result of this article. An important feature seen in the analytical expression is that the variance is uniform through all time lags (cf. Fig. 2). As expected, the variance decreases at a rate of $\sim N^{-1}$ for large N . Furthermore, $\bar{C}_{xx}^{\text{FT}}(s)$ generally exhibits an expected variance that is smaller than the variance of $\bar{C}_{xx}(s)$, suggesting that the Fourier-transform method is a preferred approach for testing the existence of correlation in a time series. In applications, Eq. (11) can be generally used to quantify the statistical noise in any time correlation functions, even for those that exhibit nonvanishing correlations. Before introducing a statistical test based on $\bar{C}_{xx}^{\text{FT}}(s)$ the expected variance for Poisson and Gaussian distributions are included as follows:

FIG. 2. Comparison of the analytical expression for expected variance in Eq. (11) with computer simulations. (a) Comparison of the analytical expression (solid line) with the standard deviation at each time lag from 10 000 simulated Poisson series ($\times$), in which each trace contains 400 elements. (b) A typical correlation trace $C_{xx}^{\text{FT}}(s)$ from an $N=400$ time series (solid line) is seen to be bound by the theoretical standard deviation (dashed line). In this particular plot, 71 out of 199 time lags ($\sim 35.7\%$) exceed the 1 standard deviation bound.

$$\text{var}\{\bar{C}_{xx}^{\text{FT}}(s)\}_{\text{poi}} = \frac{\lambda^2}{N} - \frac{2\lambda^2}{N^2} + \frac{\lambda}{N^3}, \quad (12)$$

$$\text{var}\{\bar{C}_{xx}^{\text{FT}}(s)\}_{\text{gaus}} = \frac{\sigma_x^4}{N} - \frac{2\sigma_x^4}{N^2} + \frac{8\mu^2\sigma_x^2}{N^3}. \quad (13)$$

III. A STATISTICAL TEST FOR THE EXISTENCE OF CORRELATION

The preceding discussion indicates that when H_0 is true (no correlation), $\bar{C}_{xx}^{\text{FT}}(s)$ has a Gaussian distribution with mean zero and standard deviation $\sigma_{xx}^{\text{FT}} = \sqrt{\text{var}\{\bar{C}_{xx}^{\text{FT}}(s)\}}$. When there is correlation, one expects to observe significant deviation of the mean from zero, in particular for small j_s 's. These considerations lead one to propose a statistical test for the existence of correlation, described as follows: The null hypothesis that there is no correlation in a time series of arbitrary distribution is rejected with a false-positive error rate of α when the test statistic (Z_N) exceeds the critical value c_α of confidence level $(1 - \alpha)$,

$$Z_N = \left| \frac{1}{n_t} \sum_{j_s=1}^{n_t} \bar{C}_{xx}^{\text{FT}}(j_s) \right| > c_{1-\alpha}, \quad (14)$$

where n_t is number of time lags included in the test. Equation (14) is the third major result of this work. The critical region can be calculated using

$$c_{1-\alpha} = \sqrt{2} \text{Erfc}^{-1}(\alpha) \bar{\sigma}_{xx}^{\text{FT}}, \quad (15)$$

where $\bar{\sigma}_{xx}^{\text{FT}} = \sigma_{xx}^{\text{FT}} / \sqrt{n_t} = \sqrt{\text{var}\{\bar{C}_{xx}^{\text{FT}}(s)\} / n_t}$ that can be calculated using Eq. (11), and $\text{Erfc}^{-1}(\alpha)$ is the inverse complementary error function which can be computed numerically. For example, the confidence intervals for $\alpha=0.31$, 0.1, and 0.05 are $\bar{\sigma}_{xx}^{\text{FT}}$, $1.64\bar{\sigma}_{xx}^{\text{FT}}$, and $1.96\bar{\sigma}_{xx}^{\text{FT}}$, respectively.

The proposed test was verified using computer simulations. In order to realistically reflect conditions in an actual application, the expected standard deviation $\bar{\sigma}_{xx}^{\text{FT}}$ [cf. Eq. (11)] was calculated using the empirical mean and higher moments calculated from the simulation. The results are tabulated in Table I. It can be seen that the proposed test performs very well, even for short time series. The computer simulations suggest that the criterion in Eq. (14) can be regarded as a conservative test for the existence of correlation. For the results summarized in the table, 25 time lags ($n_t=25$) were used. A larger sampling size in n_t , while decreasing statistical noise in the test may reduce the power of the test. Indeed, the general characteristics of the false-negative rate and the power of the test are not known at this point; they very likely depend on the specific type of correlation that the data exhibit.

IV. APPLICATION TO SINGLE-MOLECULE SPECTROSCOPY

Uncertainties arising from finite-length experiments are a general concern in time-dependent fluorescence single-molecule experiments. In these experiments the lengths of time traces are limited by irreversible photochemical reactions of the chromophores that render them nonemissive (for recent reviews see Refs. 5–7). The issue is further aggravated by pervasive detection noise (in this case, Poisson counting statistics), not readily appreciable in bulk experiments or computer simulations, which has been shown to be important in the interpretation of single-molecule experiments (see also recent reviews in Refs. 8–10).

Correlations in fluorescence signals are often used to test whether favorable experimental conditions have been achieved. Ideally, photon emission from a single chromophore should obey Poisson statistics and should be uncorrelated; however, one can imagine situations under which this would not hold. For example, the orientation randomization usually assumed in interpreting data from Förster resonance energy transfer (FRET) experiments is valid only when interactions between the dye and the biomolecule are

TABLE I. False-positive rate (probability) for the statistical test of correlation in a time series characterized by computer simulations. The statistics were obtained from 10 000 simulations of iid Poisson random variables with $\lambda=20$. Numbers in parentheses were obtained using the exact variance in Eq. (12) with the exact λ .

	$N=200$	$N=400$	$N=800$	$N=1600$	$N=3200$
$1 \times \bar{\sigma}_{xx}^{\text{FT}}$	0.31(0.30)	0.32(0.32)	0.31(0.31)	0.32(0.32)	0.31(0.31)
$1.64 \times \bar{\sigma}_{xx}^{\text{FT}}$	0.05(0.06)	0.08(0.08)	0.09(0.09)	0.10(0.10)	0.10(0.10)
$1.96 \times \bar{\sigma}_{xx}^{\text{FT}}$	0.02(0.03)	0.03(0.03)	0.04(0.04)	0.05(0.05)	0.05(0.05)

averaged out during each distance measurement, typically on the millisecond time scale. However, it is conceivable that an organic dye—which typically contains aromatic rings—could interact with the hydrophobic moieties in a biological macromolecule such that the orientation of the dye may be transiently restricted. Control experiments examining the emission intensity fluctuations from a single dye-biomolecule conjugate are usually carried out for this purpose; a time trace showing no autocorrelation usually indicates freely diffusing, randomly oriented dye molecules. These important controls, while not always documented in published reports, are usually evaluated by visual inspection and subjective interpretation. Therefore, an unbiased method for quantitative assessment of the presence of correlation in short, shot-noise limited single-molecule fluorescence data will help to advance the field.

We have applied the newly developed method for testing the existence of correlation in a time series to single-molecule fluorescence data. Mutant *E. coli* adenylate kinase (AK) molecules were labeled with a single chromophore (Alexa Fluor 555/Invitrogen) and immobilized on a passivated quartz cover slip, as described previously.¹¹ Single molecule experiments were performed on a confocal microscope using epi-illumination with a solid-state diode pumped, frequency-doubled Nd:yttrium-aluminum-garnet laser.¹² In order to test the orientational freedom of the dye attached to AK on the experimental time scale ($\sim$ ms), a Pockel's cell was used to rotate the orientation of the polarized excitation source by 90° at a period of 1 kHz.¹¹ Since the absorption of photons by a dye molecule is dependent on the relative orientation of the transition dipole moment and the polarization of the excitation source, a dye molecule transiently immobilized on the protein with a lifetime of milliseconds or greater will have a strong feature in its intensity autocorrelation at the oscillation frequency of the Pockel's cell. However, if the dye is freely rotating with respect to the immobilized protein on timescales faster than the Pockel's cell alternation frequency, the autocorrelation function should be a Gaussian with zero mean. A sample trajectory collected on a single AK molecule with a Pockel's cell oscillation frequency of 1 kHz is displayed in Fig. 3(a). This trajectory illustrates the challenges underlying the quantitative analysis of autocorrelation functions from single-molecule fluorescence data: they are short and inherently noisy. A Fourier-transform autocorrelation of this experimental trajectory is presented in Fig. 3(b). Error bars were calculated according to the model-independent Eq. (11) and have been plotted to 95% confidence. The newly developed statistical test for the existence of correlation [Eq. (14), $Z_N = 1.0 \times 10^{-2}$ with $N = 2062$ compared to the critical value $c_{0.05} = 3.0 \times 10^{-2}$ with $n_t = 25$] indicates that this trajectory has no correlation within 95% confidence. The lack of correlation in this trajectory reveals that there are no transient interactions between the dye and the protein on the millisecond time scale and that these types of interactions do not contribute to the millisecond dynamics that are measured in FRET experiments.

FIG. 3. Autocorrelation analysis of an experimental single-molecule fluorescence trajectory. (a) A representative single-molecule fluorescence trajectory. Irreversible photobleaching is indicated by a vertical dashed line. Trajectory is binned at 2.5 ms. (b) Fourier-transform autocorrelation analysis of the single-molecule trajectory. The statistical error bounds at 95% confidence level were calculated using $1.96 \times \sqrt{\text{var}\{C_{xx}^{\text{FT}}(s)\}}$, Eq. (11). The statistical test for the existence of correlation, Eq. (14), gives a test statistic of $Z_N = 1.0 \times 10^{-2}$ compared to the critical value $c_{0.05} = 3.0 \times 10^{-2}$, indicating within 95% confidence that this trajectory has no appreciable intensity autocorrelation. Only the first 25 lag times are included in the confidence testing. Fourier-transform autocorrelation analysis was performed with a bin size of 100 μ s.

V. CONCLUDING REMARKS

Here we have presented analytical expressions for the variance in autocorrelation functions as well as a statistical test for the presence of correlation in a time series of finite length. It was found that the Fourier-transform method of calculating the autocorrelation is preferred since it gives a constant error at all time lags. These results allow one to quantitatively access the uncertainty in an experimentally measured time series even for exceptionally noisy or short data. This is particularly useful in the field of single-molecule fluorescence experiments where shot noise is pervasive and irreversible photobleaching of the probes limits the length of each experiment. As an illustration of the application, this test is used to critically examine whether previously acquired experimental data exhibit dye-protein interactions on a millisecond time scale. The test results confirmed that with 95% confidence, there are no detectable dye-protein interactions, indicating that any dynamics measured in FRET experiments do not originate from transient orientational restriction of the dyes. The method described in this article is general because it does not require an explicit knowledge of the form of the correlation or the underlying distribution in each data element; it is expected to have wider applications beyond single-molecule spectroscopy.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation and the National Institute of Health. H.Y. is an Alfred P. Sloan Fellow.

APPENDIX: DERIVATION OF Eq. (7)

The variance for Eq. (3) is

$$\text{var}\{\bar{C}_{xx}(s)\} = \text{var}\left\{\frac{1}{N-j_s} \sum_{i=1}^{N-j_s} x_i x_{i+j_s}\right\} + \text{var}\{\bar{x}^2\} - 2\text{cov}\left\{\frac{1}{N-j_s} \sum_{i=1}^{N-j_s} x_i x_{i+j_s}, \bar{x}^2\right\}. \quad (\text{A1})$$

Note that the first and the second terms are not mutually independent. Let $E\{\dots\}$ denote the statistical expectation value (as opposed to ensemble averaging) for the random variable contained in the curly brackets. For Poisson variables [cf. Eq. (6)], the first four moments are $E\{X_i\}=\lambda$, $E\{X_i^2\}=\lambda^2+\lambda$, $E\{X_i^3\}=\lambda^3+3\lambda^2+\lambda$, and $E\{X_i^4\}=\lambda^4+6\lambda^3+7\lambda^2+\lambda$. To evaluate the variance, one expands it in terms of the expectation value expression and regroups the expansion according to the powers in X_i so that the expectation values can be calculated. For example, the first term in Eq. (A1) is evaluated by

$$\text{var}\left\{\frac{1}{N-j_s} \sum_{i=1}^{N-j_s} x_i x_{i+j_s}\right\} = E\left\{\left(\frac{1}{N-j_s} \sum_{i=1}^{N-j_s} x_i x_{i+j_s}\right)^2\right\} - E\left\{\left(\frac{1}{N-j_s} \sum_{i=1}^{N-j_s} x_i x_{i+j_s}\right)\right\}^2.$$

Further reduction of the earlier expression depends on the time lag j_s . To continue the derivation for $1 \leq j_s < (N/2)$, one has

$$\begin{aligned} & \frac{1}{N-j_s} E\{X_i^2 X_j^2\} + \frac{2(N-2j_s)}{(N-j_s)^2} E\{X_i X_j^2 X_k\} \\ & + \left[1 - \frac{3N-5j_s}{(N-j_s)^2}\right] E\{X_i X_j X_k X_l\} - E\{X_i\}^4 \\ & = \frac{1}{N-j_s} E\{X^2\}^2 + \frac{2(N-2j_s)}{(N-j_s)^2} E\{X^2\} E\{X\}^2 \\ & - \frac{3N-5j_s}{(N-j_s)^2} E\{X\}^4, \end{aligned} \quad (\text{A2})$$

where X denotes random variables from some probability distribution $f(X)$ and the i, j, k, l subscripts specify that the X 's are different. For $(N/2) \leq j_s < N$, the terms are

$$\begin{aligned} & \frac{1}{N-j_s} E\{X_i^2 X_j^2\} + \left[1 - \frac{1}{N-j_s}\right] E\{X_i X_j X_k X_l\} - E\{X_i\}^4 \\ & = \frac{1}{N-j_s} [E\{X^2\}^2 - E\{X\}^4]. \end{aligned} \quad (\text{A3})$$

Following the same procedure, the second and the third terms in Eq. (A1) become

$$\begin{aligned} \text{var}\{\bar{x}^2\} &= \text{var}\left\{\left(\frac{1}{N} \sum_{i=1}^N x_i\right)^2\right\} \\ &= E\left\{\left(\frac{1}{N} \sum_{i=1}^N x_i\right)^4\right\} - E\left\{\left(\frac{1}{N} \sum_{i=1}^N x_i\right)^2\right\}^2 \\ &= \frac{1}{N^3} E\{X^4\} + \frac{4(N-1)}{N^3} E\{X^3\} E\{X\} \\ &+ \frac{2N-3}{N^3} E\{X^2\}^2 + \frac{4(N-1)(N-3)}{N^3} E\{X^2\} E\{X\}^2 \\ &- \frac{4N^2-10N+6}{N^3} E\{X\}^4 \end{aligned} \quad (\text{A4})$$

and

$$\begin{aligned} \text{cov}\left\{\frac{1}{n_s} \sum_{i=1}^{n_s} x_i x_{i+j_s}, \bar{x}^2\right\} &= E\left\{\left[\frac{1}{n_s} \sum_{i=1}^{n_s} x_i x_{i+j_s}\right] [\bar{x}^2]\right\} - E\left\{\frac{1}{n_s} \sum_{i=1}^{n_s} x_i x_{i+j_s}\right\} E\{\bar{x}^2\} \\ &= \frac{2}{N^2} E\{X^3\} E\{X\} + \frac{2}{N^2} E\{X^2\}^2 + \frac{4N-10}{N^2} E\{X\}^2 E\{X^2\} \\ &- \frac{4N-6}{N^2} E\{X\}^4. \end{aligned} \quad (\text{A5})$$

Replacing Eqs. (A2)–(A5) into Eq. (A1) gives Eq. (7).

¹M. S. Bartlett, J. R. Stat. Soc. Ser. B (Methodol.) **8**, 27 (1946).

²R. Zwanzig and N. K. Ailawadi, *Phys. Rev.* **182**, 280 (1969).

³C. Y. Lu and D. A. Vanden Bout, *J. Chem. Phys.* **125**, 124701 (2006).

⁴G. M. Jenkins and D. G. Watts, *Spectral Analysis and Its Applications* (Holden-Day, San Francisco, 1968).

⁵P. Tinnefeld and M. Sauer, *Angew. Chem. Int. Ed.* **44**, 2642 (2005).

⁶X. Michalet, S. Weiss, and M. Jager, *Chem. Rev.* **106**, 1785 (2006).

⁷W. E. Moerner, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 12596 (2007).

⁸Y. J. Jung, E. Barkai, and R. J. Silbey, *J. Chem. Phys.* **117**, 10980 (2002).

⁹M. Lippitz, F. Kulzer, and M. Orrit, *ChemPhysChem* **6**, 770 (2005).

¹⁰D. S. Talaga, *Curr. Opin. Colloid Interface Sci.* **12**, 285 (2007).

¹¹J. A. Hanson, K. Duderstadt, L. P. Watkins, S. Bhattacharyya, J. Brokaw, J.-W. Chu, and H. Yang, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18055 (2007).

¹²L. P. Watkins, H. Chang, and H. Yang, *J. Phys. Chem.* **110**, 5191 (2006).