THE JOURNAL OF CHEMICAL PHYSICS **128**, 214101 (2008)

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

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

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

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

In practice, the correlation function is calculated from an experimentally recorded time series, $\{x_i\}$, $i=1,\ldots,N$, 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

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

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), $\bar{C}_{xx}(s) = \mathrm{iDFT}[\bar{x}_i^*\bar{x}_j]_s$, where $\bar{x}_j = \mathrm{DFT}[x_i]_j$ denotes the ith Fourier-transform element, and $*$ indicates complex conjugate. The Fourier transform method is equivalent to calculating

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

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 Ailawadi[2] 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 $\tau_0 = 2\int_0^\infty C_{xx}(s)^2/C_{xx}(0)^2 ds$ 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 $x_i x_{i+s}$) that are independent of each other, $n_{\mathrm{ind}} = T/2\tau_0$. Therefore, the sampled trace converges to the true value with a rate $\propto 1/\sqrt{n_{\mathrm{ind}}}$ as expected from standard statistical analysis.

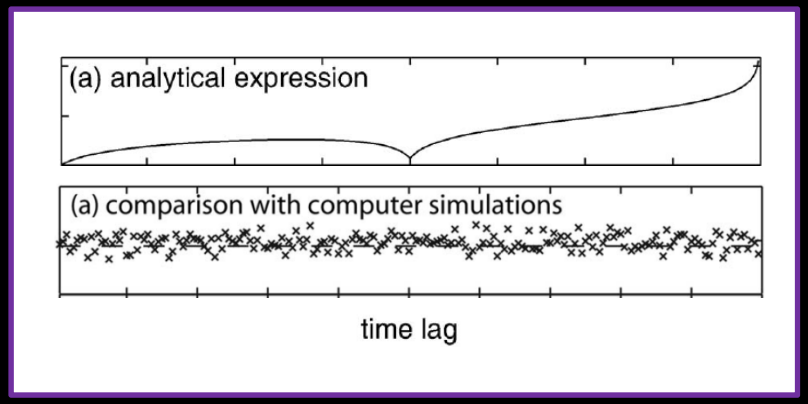
When the relaxation dynamics deviate significantly from simple exponential, the situation becomes more complicated. Lu and Vanden Bout have used computer simulations to characterize uncertainties in measuring correlation functions in the form of a stretched-exponential decay, $C(s) \propto \exp[-(s/\tau_s)^\beta]$.[3] 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.

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

Jeffery A. Hanson and Haw Yang



Def. Time correation function

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

Emperical corr. fn. w. discrete time series $\{x_i\}, i = 1, \ldots, N$

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

The variance

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

Covariance

$$\mathrm{cov}(X,Y) = E\left[(X - E(X))(Y - E(Y)) \right] = E[XY] - E[X]E[Y]$$

variance of the sum

$$\mathrm{var}\{X + Y\} = \mathrm{var}\{X\} + \mathrm{var}\{Y\} - 2\,\mathrm{cov}\{X, Y\}$$

The variance

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

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

for $1 < j_s < N/2$,

$$= \frac{1}{N-j_s} E\left\{X_i^2 X_j^2\right\} + \frac{2(N-2j_s)}{(N-j_s)^2} E\left\{X_i X_j^2 X_k\right\} + \left[1 - \frac{3N-5j_s}{(N-j_s)^2}\right] E\left\{X_i X_j X_k X_l\right\} - E\left\{x_i\right\}^4$$

$$= \frac{1}{N-j_s} E\left\{X^2\right\}^2 + \frac{2(N-2j_s)}{(N-j_s)} E\left\{X^2\right\} E\left\{X\right\}^2 - \frac{3N-5j_s}{(N-j_s)^2} E\left\{X\right\}^4$$

for $N/2 < j_s < N$,

$$= \frac{1}{N-j_s}\left[E\left\{X^2\right\}^2 - E\left\{X\right\}^4\right]$$

The variance

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

$$\boxed{\operatorname{var}\left\{\bar{x}^2\right\}} = \operatorname{var}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2\right\} = E\left\{\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^4\right\} - E\left\{\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2\right\}^2$$

$$= \frac{1}{N^3}E\left\{X^4\right\} + \frac{4(N-1)}{N^3}E\left\{X^3\right\}E\left\{X\right\} + \frac{4(N-1)(N-3)}{N^3}E\left\{X^2\right\}\left\{X\right\}^2 - \frac{4N^2-10N+6}{N^3}E\left\{X\right\}^4$$

$$\boxed{\operatorname{cov}\left\{\frac{1}{N-j_s}\sum_{i=1}^{N-j_s} x_i x_{i+j_s}, \bar{x}^2\right\}}$$

$$= E\left\{\left[\frac{1}{N-j_s}\sum_{i=1}^{N-j_s} x_i x_{i+j_s}\right]\left[\bar{x}^2\right]\right\} - E\left\{\frac{1}{N-j_s}\sum_{i=1}^{N-j_s} x_i x_{i+j_s}\right\}E\left\{\bar{x}^2\right\}$$

$$= \frac{2}{N^2}E\left\{X^3\right\}E\left\{X\right\} + \frac{2}{N^2}E\left\{X^2\right\}^2 + \frac{4N-10}{N^2}E\left\{X\right\}^2 E\left\{X^2\right\} - \frac{4N-6}{N^2}E\left\{X\right\}^4$$

The variance

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

Result 1

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

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

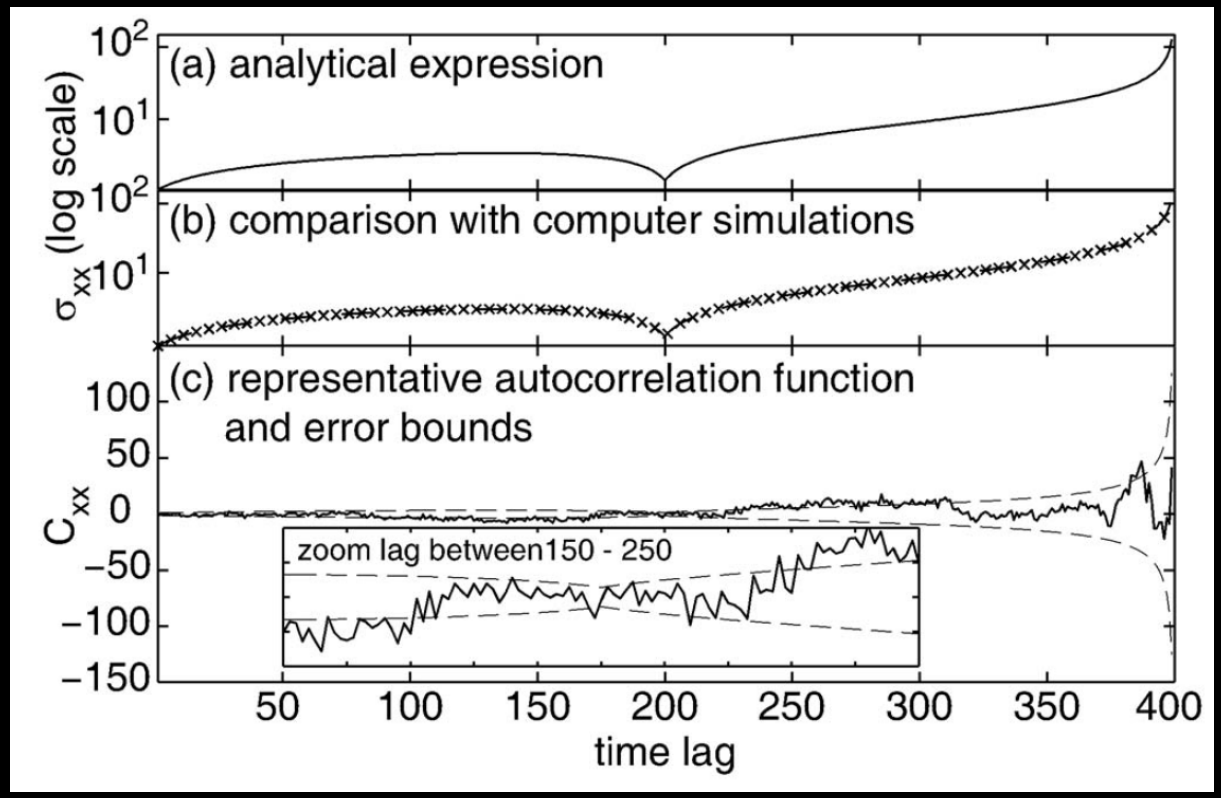
$$h_2(N, j_s) = \begin{cases} \dfrac{1}{N-j_s}E\{X^2\}^2 + \dfrac{2(N-2j_s)^2}{(N-j_s)^2}E\{X^2\}E\{X\}^2 - \dfrac{3N-5j_s}{(N-j_s)^2}E\{X^2\}E\{X\}^2, & 1 \le j_s < (N/2); \\ \dfrac{1}{N-j_s}[E\{X^2\}^2 - E\{X\}^4], & (N/2) \le j_s < N. \end{cases}$$

(a) std from N=400
(b) 10,000 simulated Poisson series (400-long)
(c) an example of corr. fn. from simulation

For the Poisson series

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

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

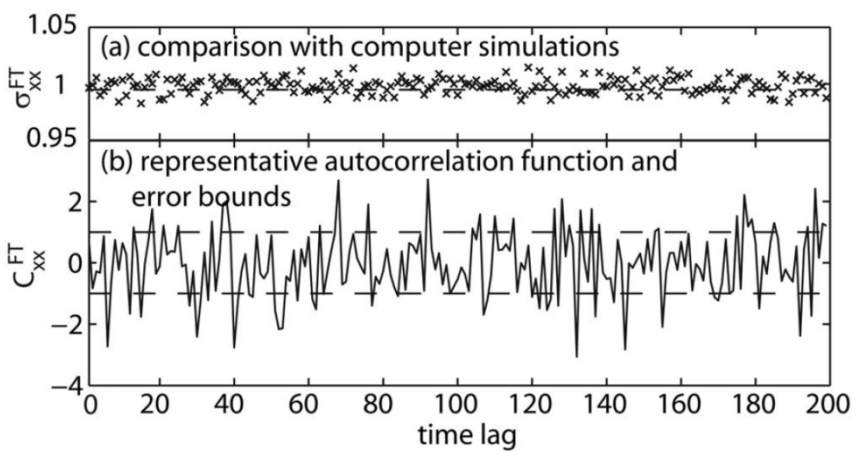


Fourier transform of the correation function

$$\bar{C}_{xx}(s) = iDFT\left\{\tilde{x}_i^* \tilde{x}_j\right\}_s$$

$$\tilde{x}_j = DFT\left\{x_i\right\}_j$$

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



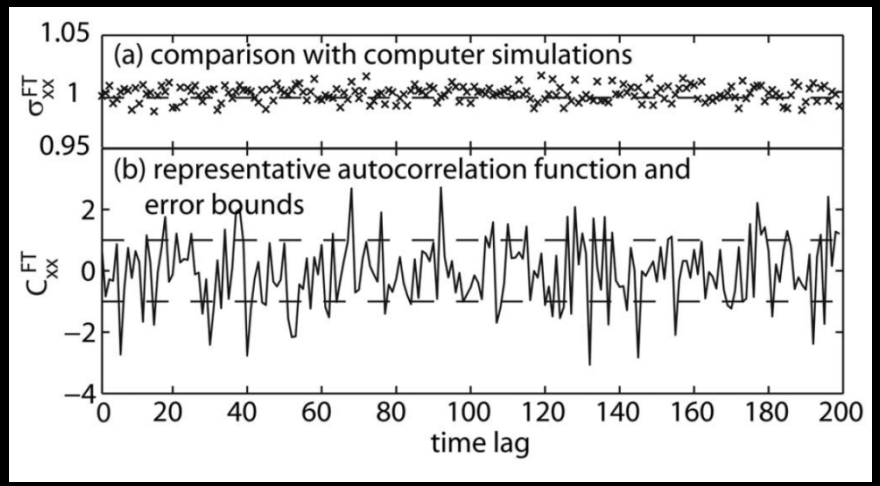
Result 2

$$\mathrm{var}\{\bar{C}_{xx}^{\mathrm{FT}}(s)\} = \frac{1}{N^3}E\{X^4\} - \frac{4}{N^3}E\{X^3\}E\{X\} + \frac{(N-3)(N+1)}{N^3}E\{X^2\}^2 - \frac{2(N^2-2N-6)}{N^3}E\{X^2\}E\{X\}^2 + \frac{N^2-2N-6}{N^3}E\{X\}^4.$$

For the Poisson series

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

Statistical test
for the existence of correlation



Result 3

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

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

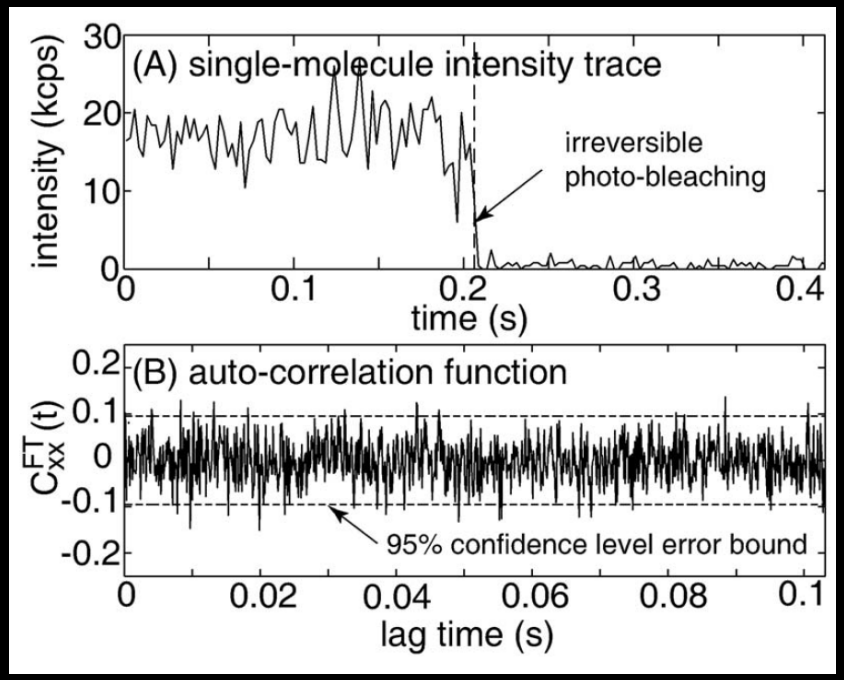
Statistical test
for the existence of correlation

Result 3

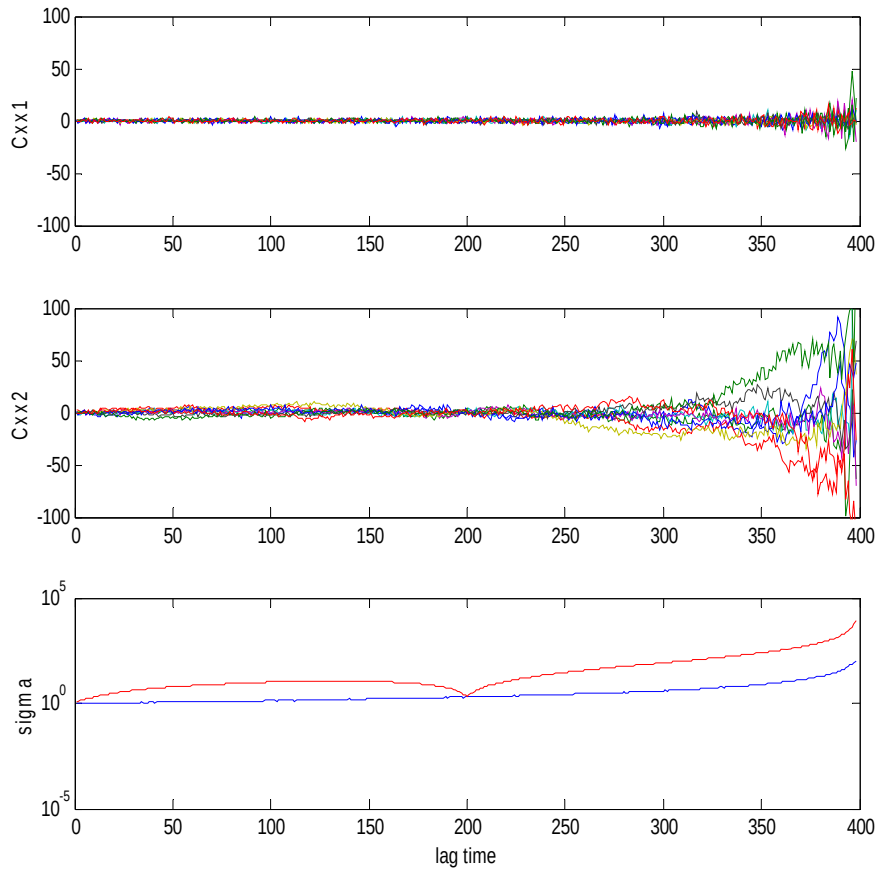
$$Z_N = \left| \frac{1}{n_t} \sum_{j_s=1}^{n_t} \bar{C}_{xx}^{\mathrm{FT}}(j_s) \right| > c_{1-\alpha},$$

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

E. coli adenylate kinase labeled w. Alexa fluor 555



The dip is counterintuitive

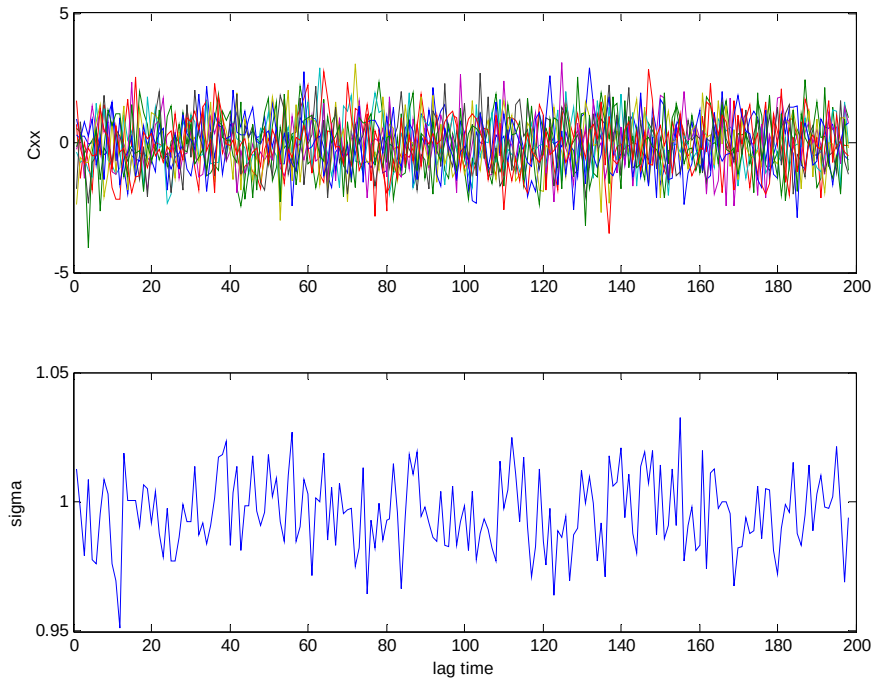


$$\bar{C}_{xx}(s) = \frac{1}{N - j_s} \sum_{i=1}^{N - j_s} x_i x_{i + j_s} - \bar{x}^2$$

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

$$\bar{C}_{xx}(s) = \frac{1}{N - j_s} \sum_{i=1}^{N - j_s} x_i x_{i + j_s} - \left(\frac{1}{N - j_s} \sum_{i=1}^{N - j_s} x_i \right) \left(\frac{1}{N - j_s} \sum_{i = j_s + 1}^{N} x_i \right)$$

$C^{FT}(s)$



Fourier transform of the correation function

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