Error estimation on averages of correlated data

Steen Bender
Copenhagen University, Department of chemistry
Nanoscience
Thur. 10/03-22

Full authors • Submitted: 08 February 1989 • Accepted: 14 March 1989 • Published Online: 31 August 1998

Error estimates on averages of correlated data

J. Chem. Phys. 91, 461 (1989); https://doi.org/10.1063/1.457480

H. Flyvbjerg

- The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen O/, Denmark

H. G. Petersen

more...

ABSTRACT

We describe how the true statistical error on a
with ease and efficiency by a renormalization g
numerical and analytical examples, having finit

The article

Hatzakis Lab

The error on the mean

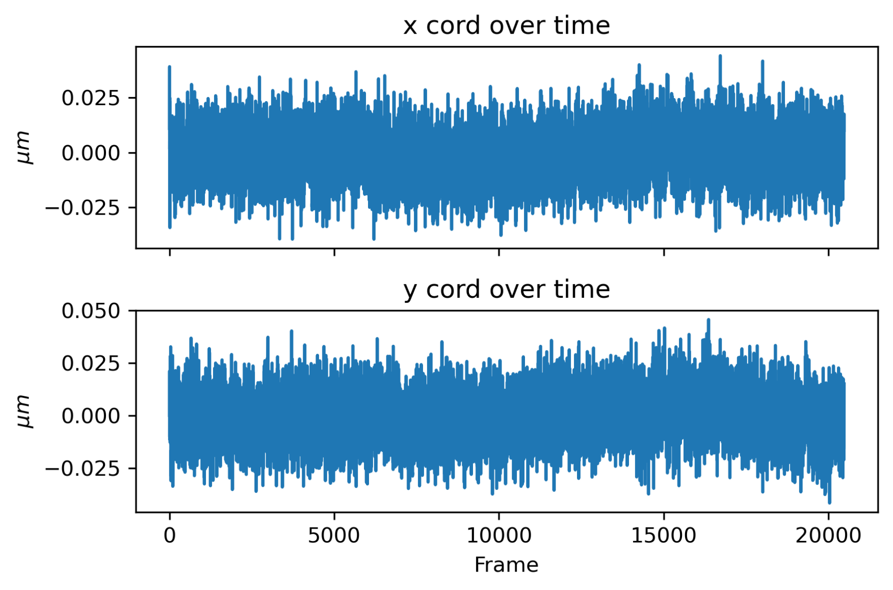
- Assuming iid
- Assuming normal distribution
- Central limit theorem
- $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $SD[\bar{x}] = \frac{SD[x]}{\sqrt{N}}$

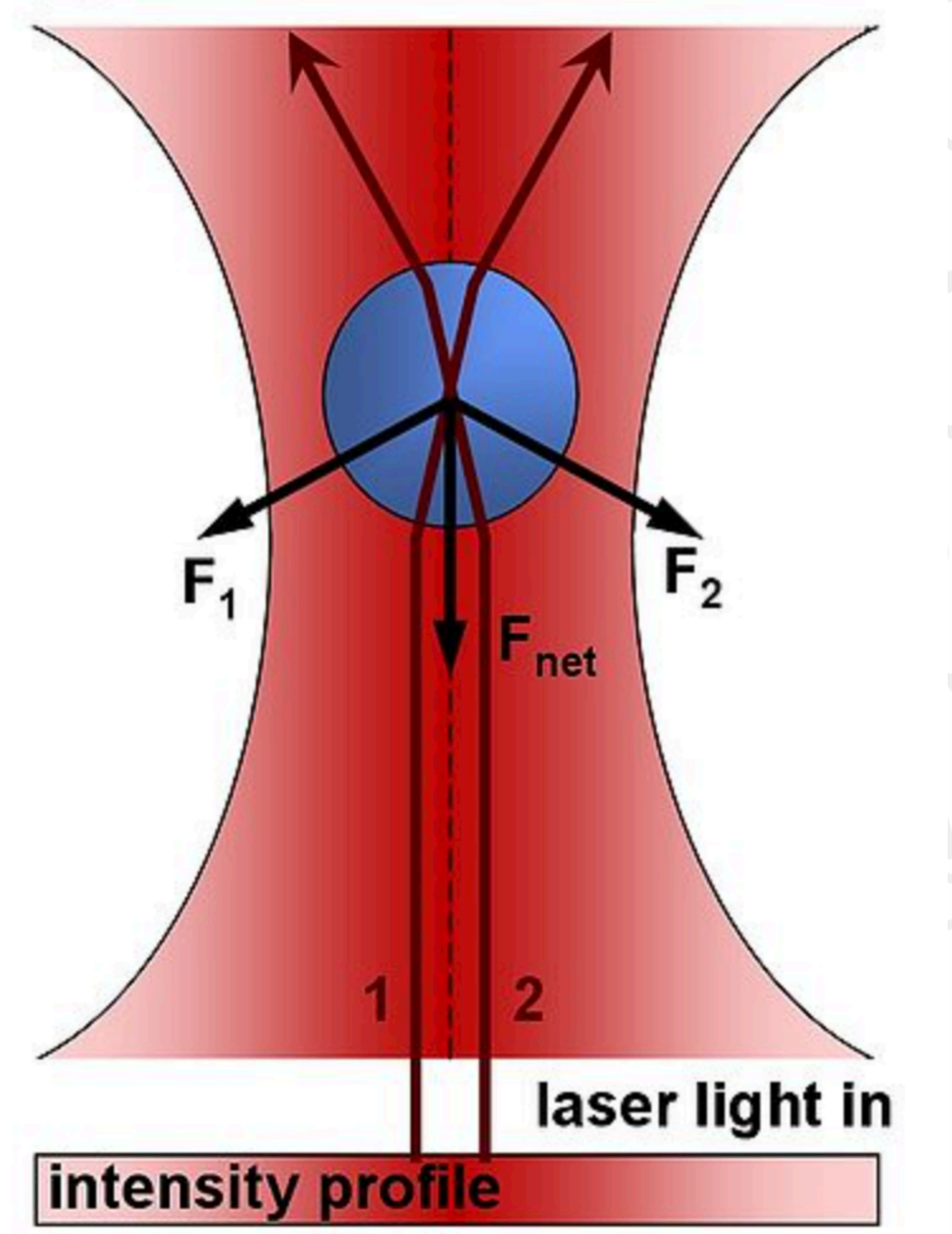
The curse of correlation

- Many of statistical method depends on assuming iid or just dependence
- Cannot determine parameters or statistics as it depends on other parameters
- Simulations and experiments of physical system normally generate data in a finite time-series
  - Correlated through last position
  - Optical tweezer as an example

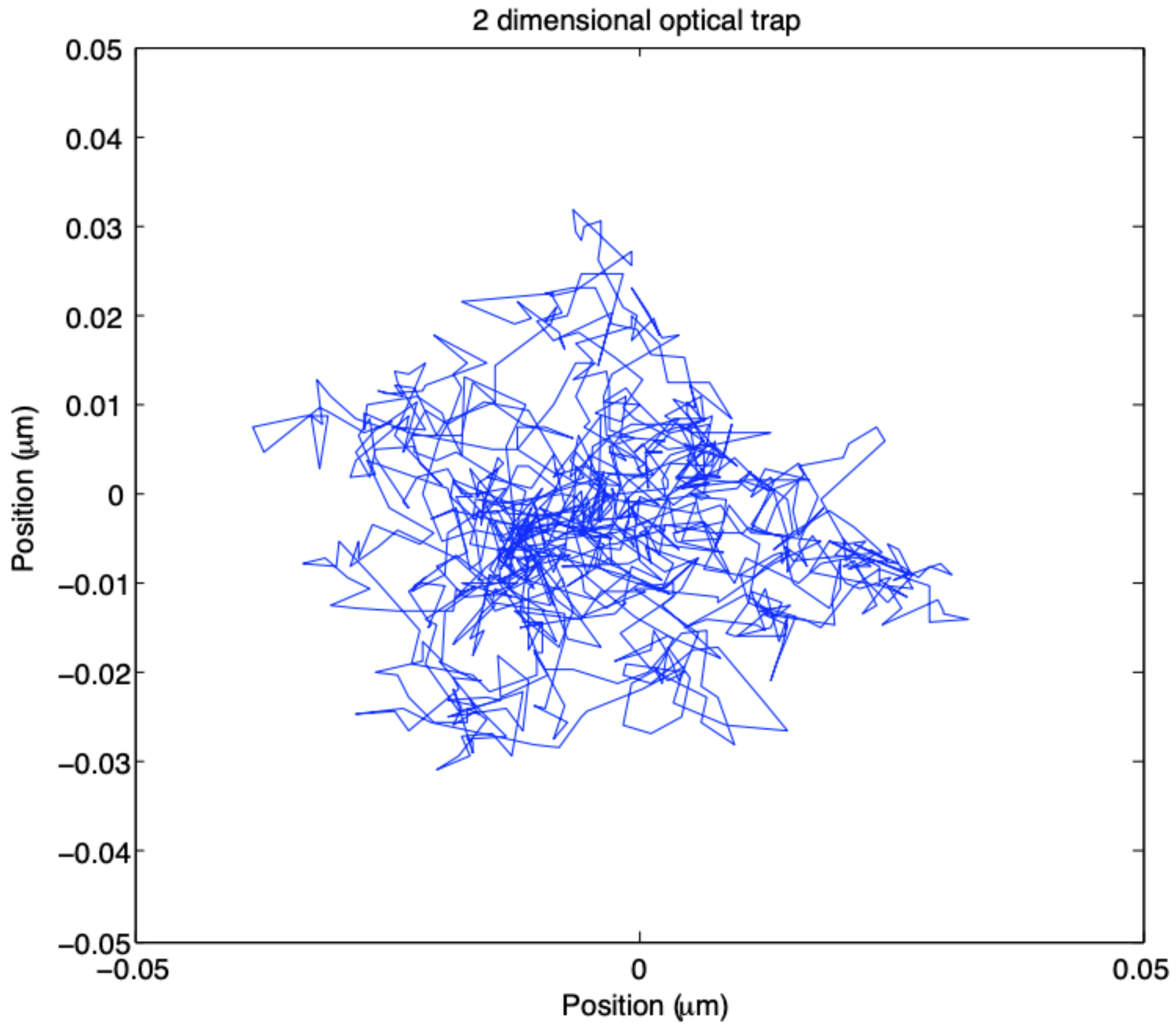
Optical tweezer

- Browian motion in x and y direction
- Hookian spring
  - $x_{j+1} = C \cdot x_j + \Delta x_j$
  - $C = e^{2\pi f_c \Delta t}$
- Stationary state





http://soft-matter.seas.harvard.edu/index.php/File:Optical-tweezer-fig2.jpg


2 dimensional optical trap
0.05
0.04
0.03
0.02
0.01
0
−0.01
−0.02
−0.03
−0.04
−0.05
Position (μm)
−0.05
0
0.05
Position (μm)

The problem and a solution

- To estimate the error on the average of correlated data
- Correlate function based estimators

  - $\gamma_{i,j} \equiv \langle x_i, x_j \rangle - \langle x_i \rangle \langle x_j \rangle \equiv \gamma_t$
  - $t = |i - j|$

- $$C_t \equiv \frac{1}{N-t} \sum_{k=1}^{n-t} (x_k - \bar{x})(x_{k+1} - \bar{x})$$
- $$\sigma^2(m) \approx \langle \frac{C_0 + 2 \cdot \sum_{t=1}^{T} (1 - \frac{t}{n}) \cdot C_t}{N - 2T - 1 + \frac{T(T+1)}{N}} \rangle$$

$$c_t \equiv \frac{1}{n-t} \sum_{k=1}^{n-t} (x_k - \bar{x})(x_{k+t} - \bar{x}), \tag{8}$$

is a *biased* estimator; its expectation value is *not* γ_t, but

$$\langle c_t \rangle = \gamma_t - \sigma^2(m) + \Delta_t, \tag{9}$$

where

$$\Delta_t = 2\left(\frac{1}{n}\sum_{i=1}^{n} - \frac{1}{n-t}\sum_{i=1}^{n-t}\right)\frac{1}{n}\sum_{j=1}^{n}\gamma_{i,j}\,. \tag{10}$$

However, if the largest correlation time in γ_t is finite, call it τ, then Eq. (5) reads

$$\sigma^2(m) = \frac{1}{n}\left[\gamma_0 + 2\sum_{t=1}^{T}\left(1-\frac{t}{n}\right)\gamma_t\right] + \mathcal{O}\left[\frac{\tau}{n}\exp(-T/\tau)\right], \tag{11}$$

where T is a cutoff parameter in the sum. For $\exp(-T/\tau) \ll 1$ the explicitly written terms in Eq. (8) clearly give a very good approximation to $\sigma^2(m) \sim \mathcal{O}(\tau/n)$. Furthermore, assuming $n \gg \tau$,

$$\Delta_t = \mathcal{O}\left(\frac{t\tau}{n^2}\right) \quad \text{for } t \ll \tau, \tag{12a}$$

growing to

$$\Delta_t = \mathcal{O}\left(\frac{\tau^2}{n^2}\right) \quad \text{for } t \gg \tau. \tag{12b}$$

So we may neglect Δ_t in Eq. (9), since it is at least a factor τ/n smaller than the term $\sigma^2(m) = \mathcal{O}(\tau/n)$. Doing that, and using Eq. (9) to eliminate γ_t from Eq. (5), we find

$$\sigma^2(m) = \frac{1}{n}\left[\langle c_0 \rangle + 2\sum_{t=1}^{T}\left(1-\frac{t}{n}\right)\langle c_t \rangle\right] + \sigma^2(m)\left(\frac{1+2T}{n} - \frac{T(T+1)}{n^2}\right). \tag{13}$$

Solving for $\sigma^2(m)$ we find

$$\sigma^2(m) \approx \left\langle \frac{c_0 + 2\Sigma_{t=1}^{T}(1-\frac{t}{n})c_t}{n - 2T - 1 + \frac{T(T+1)}{n}} \right\rangle, \tag{14}$$

$$\sigma^2(m) \approx \left\langle \frac{c_0 + 2\Sigma_{t=1}^{T} c_t}{n - 2T - 1} \right\rangle. \tag{15}$$

One also sees the approximation

$$\sigma^2(\mathrm{m}) \approx \left\langle \frac{c_0 + 2\Sigma_{t=1}^{T} c_t}{n} \right\rangle. \tag{16}$$

All these variants of Eq. (14) are equally good when T/n is sufficiently small. There is no reason *not* to use Eq. (14) itself, though, when any of the formulas are appropriate. It is as easy to compute as any of its approximations.

A variant of Eq. (8) in use is

$$c_t \equiv \frac{1}{n-t}\sum_{k=1}^{n-t}\left(x_k - \frac{1}{n-t}\sum_{k=1}^{n-t} x_k\right) \times \left(x_{k+t} - \frac{1}{n-t}\sum_{k=1}^{n-t} x_{k+t}\right). \tag{17}$$

Like Eq. (8), Eq. (17) is a biased estimator for γ_t since

$$\langle c_t \rangle = \gamma_t - \sigma^2(m) + \tilde{\Delta}_t, \tag{18}$$

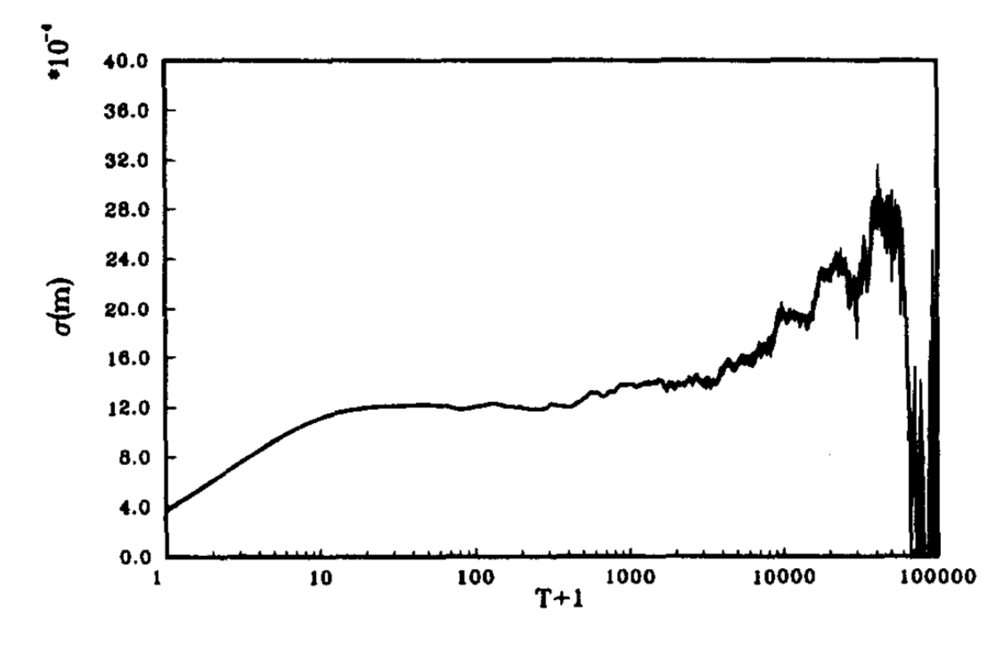
where

$$\tilde{\Delta}_t = \left(\frac{1}{n^2}\sum_{i,j=1}^{n} - \frac{1}{(n-t)^2}\sum_{i=1}^{n-t}\sum_{j=t+1}^{n}\right), \quad \gamma_{i,j} = \mathcal{O}\left(\frac{\tau t^2}{n^3}\right). \tag{19}$$

Neglecting $\tilde{\Delta}_t$ relatively to $\sigma^2(m)$ in Eq. (19) leads again to Eqs. (13) and (14). Using Eq. (17) instead of Eq. (8) as estimator for $\langle c_t \rangle$ in Eq. (14) is a better approximation, when $|\Sigma_{t=1}^{T}(n-t)\tilde{\Delta}_t| < |d\,\Sigma_{t=1}^{T}(n-t)\Delta_t|$, i.e., roughly when $T^2 < \tau n$.

The problem and a solution

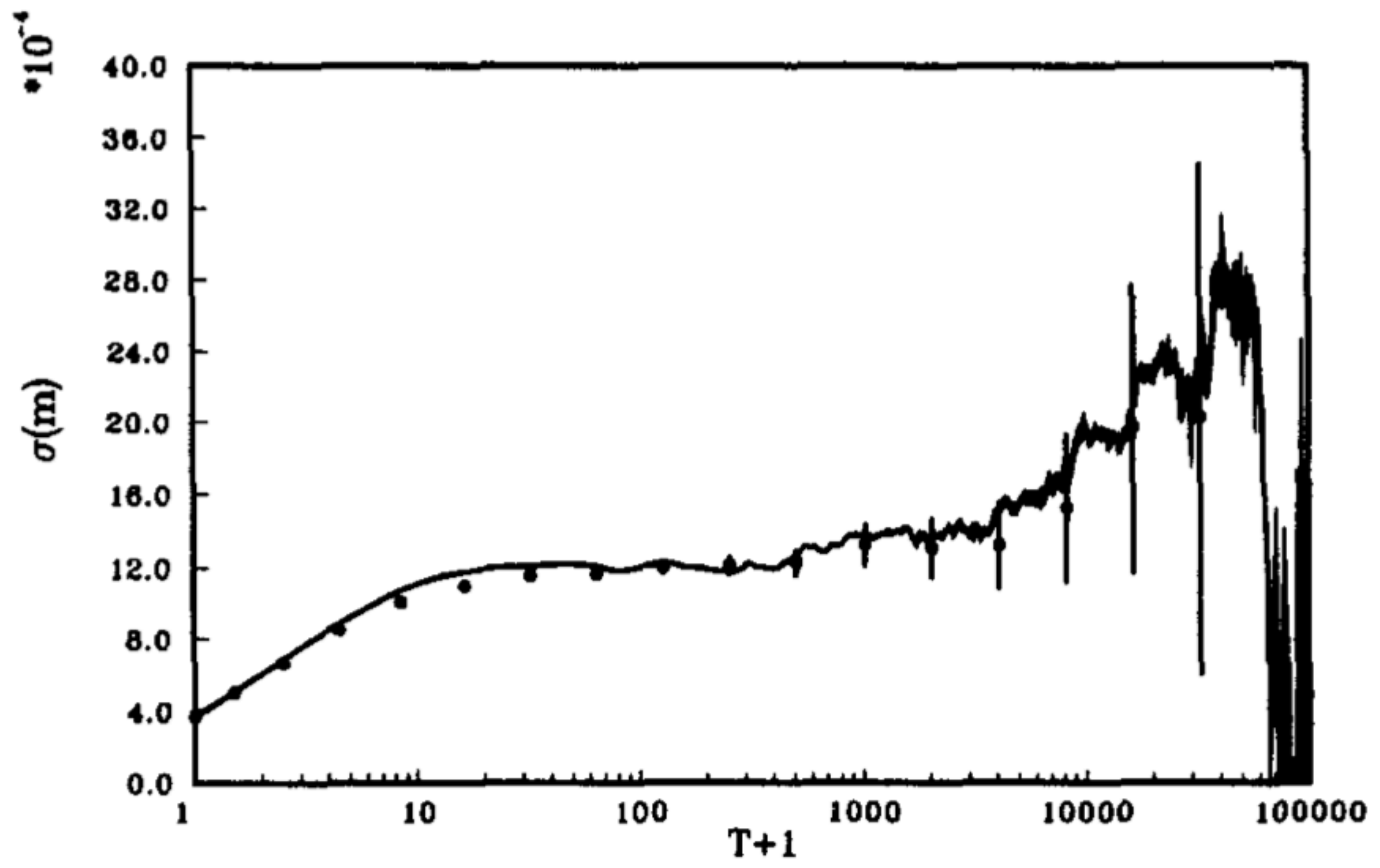
- Correlate function based estimators
  - Most commonly used method
    - Can be used for most correlation problems
    - Many different versions for different problems
  - General drawbacks
    - Manually parameter determination
    - Computational inefficient
      - $\mathcal{O}(nT_{max})$



The “blocking” method

- Let X be a finite time series with N entries containing position $\{x_1, x_2, \ldots, x_N\}$
- Calculate the standard deviation on the mean.
  - $$s = \frac{SD[\bar{x}]}{\sqrt{N}}$$
- Half the dataset
  - $$x'_i = \frac{1}{2}(x_{2i-1} + x_{2i}), \;\; N' = \frac{1}{2}N, \;\; s' = \frac{SD[\bar{x}']}{\sqrt{N'}}$$
- Repeat with s = s’, N = N’ until N’ = 2

The “blocking” method



Conclusion

- “Blocking” method more user friendly and easier to interpret
- Evaluates with as uncertainty on the error estimate.
- Can only be used for large N as it has to converge before N’ = 2
- For function has to be defined for whole range
- Function who has to be treated in log-space

$$x'_i = \frac{1}{2}(x_{2i-1} + x_{2i}), \tag{20}$$

$$n' = \tfrac{1}{2}n. \tag{21}$$

We define m' as $\bar{x}'$, the average of the n' "new" data, and have

$$m' = m. \tag{22}$$

We also define $\gamma'_{i,j}$ and γ'_t as in Eqs. (6) and (7) but from primed variables x'_i. One easily shows that

$$\gamma'_t = \begin{cases} \tfrac{1}{2}\gamma_0 + \tfrac{1}{2}\gamma_1 & \text{for } t = 0 \\ \tfrac{1}{4}\gamma_{2t-1} + \tfrac{1}{2}\gamma_{2t} + \tfrac{1}{4}\gamma_{2t+1} & \text{for } t > 0 \end{cases} \tag{23}$$

and that

$$\sigma^2(m') = \frac{1}{n'^2}\sum_{i,j=1}^{n'} \gamma'_{i,j} = \sigma^2(m). \tag{24}$$

$$\sigma^2(m) \geqslant \frac{\gamma_0}{n}$$

$$\sigma^2(m) \geqslant \left\langle \frac{c_0}{n-1} \right\rangle$$

At the fixed point the "blocked" variables $(x'_i)_{i=1,\ldots n'}$ are independent Gaussian variables—Gaussian by the central limit theorem, and independent by virtue of the fixed point value of γ'_t. Consequently, we can easily estimate the standard deviation on our estimate $c'_0/(n'-1)$ for $\sigma^2(m)$. It is $(\sqrt{2/(n-1)})\, c'_0/(n'-1)$:

$$\sigma^2(m) \approx \frac{c'_0}{n'-1} \pm \sqrt{\frac{2}{n'-1}}\frac{c'_0}{n'-1}, \tag{27}$$

$$\sigma(m) \approx \sqrt{\frac{c'_0}{n'-1}}\left(1 \pm \frac{1}{\sqrt{2(n'-1)}}\right). \tag{28}$$