Standard error estimation by an automated blocking method

Article *in* PHYSICAL REVIEW E · October 2018

DOI: 10.1103/PhysRevE.98.043304

1 author:

Marius Jonsson
University of Cambridge
6 PUBLICATIONS **36** CITATIONS

Standard error estimation by an automated blocking method

Marius Jonsson
Department of Physics, University of Oslo, N-0316 Oslo, Norway

(Received 15 May 2018; revised manuscript received 16 August 2018; published 15 October 2018)

The sample mean $\overline{X}$ is probably the most popular estimator of the expected value in all sciences and var($\overline{X}$) measures the error (standard- and mean-square-errors). Here, an alternative approach to estimation of var($\overline{X}$) for time series data is presented. The method has an accuracy similar to dependent bootstrapping, but scales in $O(n)$ time, and applies to stationary time series, including stationary Markov chains. The computational complexity is bounded by $12n$ floating point operations, but this can be reduced to $n + O(1)$ in large computations. Convergence in relative error squared is faster than $n^{-1/2}$ and the method is insensitive to the probability distribution of the observations. It is proven that a small part of the correlation structure is relevant to the convergence rate of the method. From this, proof of the Blocking method [Flyvbjerg and Petersen, J. Chem. Phys. **91**, 461 (1989)] follows as a corollary. The result is also used to propose a hypothesis test surveying the relevant part of the correlation structure. It yields a fully automatic method which is sufficiently robust to operate without supervision. An algorithm and sample code showing the implementation is available for PYTHON, C++, and R [www.github.com/computative/block]. Method validation using autoregressive AR(1) and AR(2) processes and physics applications is included. Method self-evaluation is provided by bias and mean-square-error statistics. The method is easily adapted to multithread applications and data larger than computing cluster memory, such as ultralong time series or data streams. This way, the paper provides a stringent and modern treatment of the Blocking method using rigorous linear algebra, multivariate probability theory, real analysis, and Fisherian statistical inference.

DOI: 10.1103/PhysRevE.98.043304

I. INTRODUCTION

Estimation of the variance of sample means $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ is essential in natural sciences [1,2]. This is because $\overline{X}$ is a typical estimator of the expected value of the observations $X_1, X_2, \ldots$, if the observations are identically distributed with finite variance. The variance of the mean is the expected squared error of the estimate. Already in 1867, Chebyshev [3] explained this by showing that if the observations have expected value μ, variance σ^2, and $\overline{X}$ has finite nonzero variance var($\overline{X}$), then for any real number $k > 0$

$$P(|\overline{X} - \mu| > k[\mathrm{var}(\overline{X})]^{1/2}) \leqslant \frac{1}{k^2} \quad \text{(Chebyshev's ineq.)}.$$

Here, P is the probability measure of the X_i's [4]. Chebyshev's inequality says that it is likely that the difference $|\overline{X} - \mu|$ is a small number. If the observations are independent and identically distributed, the variance of the mean is easily obtained by setting $\mathrm{var}(\overline{X}) = \sigma^2/n$ [5], but for correlated data, the computation is more complicated [6]. Here, however, I show that if there is some integer $d > 1$ such that $n = 2^d$, and $X_1, \ldots, X_n$ are observations from a stationary time series, then the computational complexity is essentially the same as that of the sample mean, and one can use an automated scheme to compute it.

The method uses so called blocking transformations [7]. This refers to forming a new sample of data by taking the mean of every pair of subsequent observations. To be precise, *blocking transformation number i* relates each element B_k of a vector $\mathbf{B} \in \mathbb{R}^{n_i}$ to the elements A_k of $\mathbf{A} \in \mathbb{R}^{n_{i-1}}$ by

$$B_k = (1/2)(A_{2k-1} + A_{2k}). \tag{1}$$

Such transformations are applied in many areas of probability theory, and Flyvbjerg and Petersen [7] popularized a method where blocking transformations reduced the correlations of the data, and proposed a way to estimate the variance of $\overline{X}$. In this study, the mathematics is developed and an automation of the method is given. The rigor is similar to modern mathematics and an automation that is robust enough to operate without supervision is provided. The philosophy of Flyvbjerg and Petersen [7] is recycled, but the mathematics is different. The method validation has physics applications and experimental results quantifying all errors involved in applications. The method works by applying blocking transformations of the type given in Eq. (1) until correlation of observations is no longer significantly different from zero. The results show that the behavior of the method is specified by the autocovariance $\gamma(1)$. Furthermore, if $\gamma(1)$ is not significantly different from zero, the method has obtained the ideal estimate. Next, I developed an automated statistical test that stops the algorithm when there is no reason to believe that $\gamma(1)$ is different from zero. Hence, speedups of a thousand or more are possible because the complexity is $O(n)$ as compared to dependent bootstrapping or other methods of complexity $O(n^2)$ or $O(n \log n)$. I give the algorithm and sample code in [8].

The preliminary background (definitions and theorems) is scattered throughout in text blocks where they are required.

First, proof of the blocking method is given and, second, an automation that is sufficiently robust to operate without supervision is derived to perform the calculations. In the discussion I compare the properties of the present blocking method with other relevant methods for computing the variance of the sample mean for correlated data.

A. Key ideas

First, the types of considered time series are defined.

Preliminaries 1. A set of random variables $\{X_i\}$ is said to be a *time series* if it is possible to think of the variables as being ordered as a function of time. The focus will be on infinite time series $X_1, X_2, \ldots$, but also the part of it that is possible to sample: that is, the first n observations: $X_1, X_2, \ldots, X_n$. The random variables $\{X_i\}$ are said to be *stationary* or *weakly stationary* if (1) there exist $\mu \in \mathbb{R}$ such that $\langle X_i \rangle = \mu$ for all i and (2) the covariances $\mathrm{cov}(X_i, X_j)$ only depend on the difference $h = |i - j|$ for all $1 \leqslant i, j \leqslant n$ [6]. A time series is *strictly stationary* if the cumulative distribution function (cdf) of all sets of the form $\{X_i, X_{i+1}, \ldots, X_{i+k}\}$ equals the cdf of the set $\{X_{i+j}, X_{i+j+1}, \ldots, X_{i+j+k}\}$ for all i, j, k [6]. A strictly stationary time series with finite variance is stationary [9]. The function $\gamma(h) = \mathrm{cov}(X_i, X_{i+h})$ is the *autocovariance* of $\{X_j\}_{j=1}^{\infty}$. ▶

$\mathrm{var}(\overline{X})$ will be estimated by a quantity σ_k^2/n_k, which is subject to an error e_k, for $k \in \{0, 1, 2, \ldots\}$. These quantities will be defined soon. The index k denotes how many sequential blocking transformations have been applied to the data. The first aim of the paper is to prove that if the autocovariance $\gamma(h) \to 0$ as $h \to \infty$, then e_k can be made as small as you may wish, by applying enough blocking transformations:

Theorem. Assume that the stationary time series $X_1, X_2, \ldots$ has autocovariance $\gamma(h) \to 0$ as $h \to \infty$. Then, for every $\varepsilon > 0$ there exists a natural number K such that $|e_k| < \varepsilon$ if $K \leqslant k \leqslant d$ for the time series $X_1, X_2, \ldots, X_{2^d}$.

This begs the question of how to find the number K that ensures that e_K is not significantly different from zero. I present a hypothesis test that automatically determines this for you using a function M_j that will be defined later:

Theorem. If $X_1, X_2, \ldots$ is a strictly stationary time series such that $\gamma(h) \to 0$ as $h \to \infty$ with $\lim_{n\to\infty} \mathrm{var}(\sum_{i=1}^n X_i) = \infty$ and $\langle |X_i|^{2+a} \rangle < \infty$, for some $a > 0$ then M_j is a test statistic that is asymptotic χ^2_{d-j} distributed under the hypothesis $\gamma_k(1) = 0$ for all $k \geqslant j$. The rejection region includes all values M_j larger than $q_{d-j}(1-\alpha)$ for all $1 \leqslant j \leqslant d-1$.

Using this theorem, calculations are automated in six steps (see Fig. 2). Users primarily interested in the algorithm can jump ahead to Sec. III B to read more. For those interested in justification of the method, it is necessary to introduce measure of dependence on time series and under blocking transformations before starting work on the first theorem:

Preliminaries 2. If the time series has finite length n, it is possible to form an n-vector or n-tuple $\mathbf{X}$ containing the elements $\{X_j\}_{j=1}^n$. For any vector $\mathbf{X}$, define $\langle \mathbf{X} \rangle$ to be the vector with elements $\langle X_i \rangle$. A pair of random variables X_i, X_j are *uncorrelated* if $\mathrm{cov}(X_i, X_j) = 0$. A time series is uncorrelated if $\gamma(h) = 0$ for all $h \neq 0$ and *asymptotic uncorrelated* if the autocovariance $\gamma(h) \to 0$ as $h \to \infty$. The matrix consisting of elements $(\Sigma)_{ij} = \gamma(|i-j|)$ is the *covariance matrix* of $\mathbf{X}$, and $\widehat{\gamma}(h)$ is the *sample covariance* and $\widehat{\sigma}^2$ is the *sample variance* according to [6] if

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-h}(X_i - \overline{X})(X_{i+h} - \overline{X}) \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\gamma}(0). \tag{2}$$

Subscripts are used on the variables to indicate that $(X_k)_1, (X_k)_2, \ldots$ and $\mathbf{X}_k$ are *subject to k blocking transformations* if they are related to $X_1, X_2, \ldots$ and $\mathbf{X}$ by k repeated transformations of the type given in Eq. (1). Subscripts will also be used to denote the length of the vector $\mathbf{X}_k$ by the symbol n_k. Since $X_1, X_2, \ldots$ is subject to zero transformations, $\{(X_0)_i\}_{i=1}^{\infty} = \{X_i\}_{i=1}^{\infty}$ and $\mathbf{X} = \mathbf{X}_0$ and $n = n_0$ is used to emphasize this. It will be shown in Lemma 1 that $\{(X_k)_i\}_{i=1}^{\infty}$ is indeed stationary if $\{X_i\}_{i=1}^{\infty}$ is, so it is possible to let the mean, autocovariance, and variance of the blocking-transformed variables be given subscripts to denote which blocking iteration they belong to: $\overline{X}_k$, σ_k^2, $\widehat{\sigma}_k^2$, $\gamma_k(h)$, $\widehat{\gamma}_k(h)$. Assuming $h = |i-j|$ and using the definition of the blocking transformation, Eq. (1), and the distributive property of the covariance, it is clear that

$$\begin{aligned}
\gamma_{k+1}(h) &= \mathrm{cov}((X_{k+1})_i, (X_{k+1})_j) \\
&= \tfrac{1}{4}\,\mathrm{cov}((X_k)_{2i-1} + (X_k)_{2i}, (X_k)_{2j-1} + (X_k)_{2j}) \\
&= \begin{cases} \frac{1}{2}\gamma_k(2h) + \frac{1}{2}\gamma_k(2h+1) & \text{if } h = 0, \\ \frac{1}{4}\gamma_k(2h-1) + \frac{1}{2}\gamma_k(2h) + \frac{1}{4}\gamma_k(2h+1) & \text{else.} \end{cases}
\end{aligned} \tag{3}$$

Finally, the variance of the sample mean can be expressed in terms of the autocovariance function by

$$\begin{aligned}
\mathrm{var}(\overline{X}) &= \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\,\mathrm{cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right] \\
&= \frac{1}{n^2}[n\gamma(0) + (n-1)\gamma(1) + \cdots + \gamma(n-1) \\
&\quad + (n-1)\gamma(-1) + (n-2)\gamma(-2) \\
&\quad + \cdots + \gamma(1-n)] \\
&= \frac{\sigma^2}{n} + \frac{2}{n}\sum_{h=1}^{n-1}\left(1 - \frac{h}{n}\right)\gamma(h) \quad \text{if} \quad \gamma(0) = \sigma^2.
\end{aligned} \tag{4}$$

Using these definitions, the first theorem can be obtained. ▶

II. TIME SERIES BEHAVIOR UNDER BLOCKING TRANSFORMATIONS

Section III states the idea of the announced algorithm. However, in order to understand why the algorithm works, preliminary results are required. The first result explains which part of the correlation structure is sufficient to survey, but before showing that, consider the following lemma, which will be frequently used. One of the main reasons why it is important, is that it justifies that γ_k exists for all integers $k \geqslant 0$. It does this because it turns out that a stationary time series remains stationary after blocking transformations are applied.

Lemma 1. Let $X_1, X_2, \ldots$ be a stationary time series and $\mathbf{X}$ be the vector of the first $n = 2^d$ sequential observations from $X_1, X_2, \ldots$. Suppose $\mathbf{X}_k$ are the n_k first observations of the time series $(X_k)_1, (X_k)_2, \ldots$. Then, both $(X_k)_1, (X_k)_2, \ldots$ and $\mathbf{X}_k$ are stationary. Moreover, if $\overline{X}_k$ is the sample mean of $\mathbf{X}_k$, then $\overline{X} = \overline{X}_k$ for all $0 \leqslant k \leqslant d-1$.

Proof. We first show that the time series $(X_k)_1, (X_k)_2, \ldots$ is weakly stationary using induction. Since elements of $\{X_i\}_{i=1}^{\infty}$ are stationary, there is $\mu \in \mathbb{R}$ such that $\langle (X_0)_i \rangle = \langle X_i \rangle = \mu$ for $i \geqslant 1$ and so the base case is trivially satisfied. For the induction step, write

$$\langle (X_{k+1})_i \rangle \overset{(1)}{=} \tfrac{1}{2}\langle (X_k)_{2i-1} + (X_k)_{2i} \rangle = \tfrac{1}{2}(\mu + \mu) = \mu.$$

For the covariance, the elements of $\{X_i\}$ are stationary, and therefore $\mathrm{cov}(X_i, X_j)$ only depends on the difference $|i-j| = h$, which proves the base case. Now, if the hypothesis is true for some k, then according to Eq. (3), it is true for $k+1$ since Eq. (3) says it only depends on the difference $h = |i-j|$. This proves that the elements $\{(X_k)_i\}_{i=1}^{\infty}$ are stationary for all $k \geqslant 0$. The proof works for any smaller time series $\{(X_k)_i\}_{i=a}^{b}$ for $a \geqslant 1$. By taking $b = n_k$, this proves $\mathbf{X}_k$ is stationary.

To show that the mean satisfies $\overline{X} = \overline{X}_k$ for all $0 \leqslant k \leqslant d-1$, use induction. Here, the base case is trivially satisfied. So, write

$$\begin{aligned} n_{k+1}\overline{X}_{k+1} &= \sum_{i=1}^{n_{k+1}} (X_{k+1})_i \overset{(1)}{=} \frac{1}{2}\sum_{i=1}^{n_k/2} [(X_k)_{2i-1} + (X_k)_{2i}] \\ &= \frac{1}{2} n_k \overline{X}_k = n_{k+1}\overline{X}_k, \end{aligned}$$

which provides the induction step. ∎

Using Lemma 1 and Eq. (4) it is clear that any estimate of $\mathrm{var}(\overline{X})$ using σ_k^2/n_k has *truncation error* given by

$$e_k \equiv \frac{2}{n_k}\sum_{h=1}^{n_k-1}\left(1 - \frac{h}{n_k}\right)\gamma_k(h). \tag{5}$$

The next proposition is crucial: it says that if $\gamma_0(1), \gamma_1(1), \ldots, \gamma_{d-1}(1)$ are known, then the behavior of the truncation error e_k is known. Corollary 1 will explain the details of this, but the main idea is that if $e_{k+1} - e_k$ is known for all k, then the values of e_k are known up to a constant for all k.

Proposition 1. Suppose $2^d \geqslant 2$ is the number of observations, and σ_k^2 is finite for all $k \in \{0, 1, \ldots, d-1\}$. Then, the rate of change of the truncation error e_k is

$$e_k - e_{k+1} = \frac{\gamma_k(1)}{n_k} \quad \text{for all} \quad 0 \leqslant k < d-1. \tag{6}$$

To prove the proposition, sum each side of Eq. (3) to get

$$\sum_{h=1}^{n_{k+1}-1}\gamma_{k+1}(h) = \frac{1}{2}\sum_{h=1}^{n_k-1}\gamma_k(h) - \frac{1}{4}[\gamma_k(1) + \gamma_k(n_k-1)]. \tag{7}$$

Similarly, sum Eq. (3):

$$\sum_{h=1}^{n_{k+1}-1} h\gamma_{k+1}(h) = \frac{1}{4}\sum_{h=1}^{n_k-1} h\gamma_k(h) - \frac{n_k}{8}\gamma_k(n_k-1). \tag{8}$$

Plugging these equations into the definition of e_k given in Eq. (5) and using $n_{k+1} = n_k/2$, it is immediate that

$$\begin{aligned} e_{k+1} &= \frac{2}{n_{k+1}}\sum_{h=1}^{n_{k+1}-1}\left(1 - \frac{h}{n_{k+1}}\right)\gamma_{k+1}(h) \\ &= \frac{4}{n_k}\sum_{h=1}^{n_{k+1}-1}\gamma_{k+1}(h) - \frac{8}{n_k^2}\sum_{h=1}^{n_{k+1}-1} h\gamma_{k+1}(h) \\ &\overset{(7)(8)}{=} e_k - \frac{\gamma_k(1)}{n_k}. \end{aligned}$$

The following corollary is the most important takeaway. It shows the effect of $\gamma_k(1)$ on the error of the estimate σ_k^2/n_k. Interestingly, this provides a proof of the behavior of the Flyvbjerg and Petersen [7] blocking method.

Corollary 1. Suppose $X_1, \ldots, X_n$ and $n = 2^d > 2$ are random variables from a weakly stationary sample with σ_k^2 finite for all $k \in \{0, 1, \ldots, d-1\}$, and $i < j$:

(1) If there exists $k \in \mathbb{N}$ such that for all $i \leqslant k \leqslant j$ either $\gamma_k(1) > 0$ or $\gamma_k(1) \geqslant 0$ or $\gamma_k(1) = 0$, then the sequence of errors e_k is strictly decreasing or decreasing or constant on $i \leqslant k \leqslant j$, respectively.

(2) If there exists some $k \in \{0, 1, \ldots, d-1\}$ such that the elements of $\mathbf{X}_k$ are uncorrelated, then the sequence of errors e_j is constant, and $\sigma_{j+1}^2 = \sigma_j^2/2$ for all $j \geqslant k$.

Proof. Suppose the hypothesis is true and first let $\gamma_k(1) > 0$. That means Proposition 1 is true and there exist $k \in \mathbb{N}$ such that $\gamma_k(1) > 0$ for all $i \leqslant k < j$. If $u, v \in \{i, i+1, \ldots, j+1\}$ are distinct natural numbers, assume without loss of generality that $u < v$. By hypothesis, $n_k > 0$ and $\gamma_k(1) > 0$, and a sum of such terms must be positive. That means

$$\begin{aligned} 0 < \sum_{k=u}^{v-1}\frac{\gamma_k(1)}{n_k} &= \frac{\gamma_u(1)}{n_u} + \frac{\gamma_{u+1}(1)}{n_{u+1}} + \cdots + \frac{\gamma_{v-1}(1)}{n_{v-1}} \\ &\overset{(6)}{=} (e_u - e_{u+1}) + (e_{u+1} - e_{u+2}) + \cdots + (e_{v-1} - e_v) \\ &= e_u - e_v. \end{aligned}$$

Now, by adding e_v to each side of the inequality, the first part is proven. To obtain the result in the case $0 \leqslant \gamma_k(1)$, replace $<$ with $\leqslant$ in the argument above. The case $\gamma_k(1) = 0$ is obtained by replacing $<$ with $=$.

Suppose δ_{ij} is the Kronecker delta. To obtain part 2, use induction: assume that the elements of $\mathbf{X}_k$ are uncorrelated. Then, the base case is trivially satisfied since all uncorrelated variables have zero covariance [3]. The induction step follows for $k+1$ since Eq. (3) says that $\gamma_{k+1}(i) = \delta_{i0}\sigma_k^2/2$. This proves $\gamma_j(1)$ is zero for all $j \geqslant k$, so the error is constant by what was proved above. ∎

These results will be useful in the automation of the blocking method below. And, as stated, the corollary proves the behavior of the blocking method. But, experts will spot a problem: The sequence e_k may be decreasing and eventually constant if the elements of $\mathbf{X}_k$ become uncorrelated. But, there is no guarantee that the variables become uncorrelated. However, existing users of the blocking method [7] know the variables do indeed become uncorrelated (and the constant from part 2 of the corollary is zero). But, so far this is not guaranteed. So, although our present results are promising and hint at the conclusions to come, a bit more work is

required. The next part is technical, but the purpose is to obtain a decomposition of $\gamma_k(h)$ in Lemma 3. Start by fixing some $k \in \{0, 1, \cdots\}$ and consider this interesting sequence of functions and its properties:

$$f_k(i) = \begin{cases} i & \text{if } 0 \leqslant i \leqslant 2^k, \\ 2^{k+1} - i & \text{if } 2^k \leqslant i \leqslant 2^{k+1}, \\ 0 & \text{else.} \end{cases} \tag{9}$$

Lemma 2. The sequence $\{f_k\}$ has the following nice properties:

(1) $f_k(i) \leqslant i$ for all $i \in \mathbb{N}$.
(2) $\sum_{i=1}^{2^{k+1}-1} f_k(i) = 2^{2k}$.
(3) $f_{k+1}(i) = f_k(i) + 2f_k(i - 2^k) + f_k(i - 2^{k+1})$.

Proof. See the Appendix. ∎

Lemma 3. Suppose $X_1, X_2, \ldots$ is a stationary time series and h and k are positive natural numbers, then,

$$\gamma_k(h) = 2^{-2k} \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma[2^k(h-1) + i]. \tag{10}$$

Proof. We prove the lemma by induction. Assume $k = 1$ and write

$$\begin{aligned} \gamma_1(h) &\overset{(3)}{=} 2^{-2}(\gamma_0(2h-1) + 2\gamma_0(2h) + \gamma_0(2h+1)) \\ &= 2^{-2k} \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma[2^k(h-1) + i]. \end{aligned}$$

Assume now that Eq. (10) is true for some $k \geqslant 1$ and write

$$\begin{aligned} 2^{2(k+1)}\gamma_{k+1}(h) &\overset{(3)}{=} 2^{2k}\big[\gamma_k(2h-1) + 2\gamma_k(2h) + \gamma_k(2h+1)\big] \\ &\overset{(10)}{=} \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma[2^k(2h-2) + i] \\ &\quad + 2\sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma[2^k(2h-1) + i] \\ &\quad + \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma[2^k(2h) + i] \\ &\overset{(9)}{=} \sum_{i=1}^{2^{k+1+1}-1} \gamma[2^{k+1}(h-1) + i] \\ &\quad \times [f_k(i) + 2f_k(i-2^k) + f_k(i-2^{k+1})] \\ &\overset{\text{Lemma 2}}{=} \sum_{i=1}^{2^{k+1+1}-1} f_{k+1}(i)\gamma[2^{k+1}(h-1) + i]. \end{aligned}$$

In the third equality, the summation limits were shifted by 0, 2^k, and 2^{k+1}, respectively; in addition, I used that $f_n(i - 2^j) = 0$ whenever $i \leqslant 2^j$ or $2^{k+1} + 2^j \leqslant i$ from Eq. (9). This allowed to factor out the term $\gamma[2^{k+1}(h-1) + i]$. ∎

Proposition 1 shows that $\gamma_k(1)$ is of special interest to us and therefore the following corollary is interesting.

Corollary 2. Suppose $X_1, X_2, \ldots$ is a stationary time series and k is a positive natural number, then,

$$\begin{aligned} 2^{2k}\gamma_k(1) &= \gamma(1) + 2\gamma(2) + \cdots + 2^k\gamma(2^k) \\ &\quad + (2^k - 1)\gamma(2^k + 1) + \cdots + \gamma(2^{k+1} - 1). \end{aligned} \tag{11}$$

Proof. Use the previous lemma with $h = 1$. ∎

Using these results, everything is now set to finalize the investigation of γ under blocking transformations. The following proposition may sound technical at first, but it will carry us all the way to the final proof of the blocking method.

Proposition 2. Assume that the stationary time series $X_1, X_2, \ldots$ has autocovariance $\gamma(h) \to 0$ as $h \to \infty$. Then, $\{\gamma_k\}_{k=1}^{\infty}$ converges uniformly to the zero function on $\mathbb{N}$.

Proof. Pick $\varepsilon > 0$. By assumption $\gamma(i) \to 0$ as $i \to \infty$. So, there exists $I \in \mathbb{N}$ such that $\gamma(i) < \varepsilon/2$ when $i \geqslant I$. Define $S = |\sum_{i=1}^{I} i\gamma(i)|$. Set $K = \max\{\log_2(I), (1/2)\log_2(2S/\varepsilon)\}$. Assume first that $h \geqslant 2$ and let $j \in \mathbb{N}$ be any natural number, then, by construction, if $k \geqslant K$ then we have

$$\begin{aligned} k \geqslant K \geqslant \log_2 I \geqslant \log_2 \frac{I}{h-1} \quad &\text{only if} \quad 2^k(h-1) + j \geqslant I \\ &\text{only if} \quad \gamma[2^k(h-1) + j] < \frac{\varepsilon}{2} \end{aligned}$$

since $\log_2$ is a monotonous function. Thus, by Lemma 3 and the triangular inequality

$$\begin{aligned} |\gamma_k(h) - 0| &\leqslant 2^{-2k} \sum_{j=1}^{2^{k+1}-1} f_k(j)|\gamma[2^k(h-1) + j]| \\ &\leqslant \frac{\varepsilon}{2}2^{-2k} \sum_{j=1}^{2^{k+1}-1} f_k(j) = \frac{\varepsilon}{2}2^{-2k}2^{2k} = \frac{\varepsilon}{2} < \varepsilon, \end{aligned}$$

where Lemma 2 was used in the third step. By construction, it is possible to assume $k \geqslant K \geqslant (1/2)\log_2(2S/\varepsilon)$, so $\varepsilon/2 \geqslant 2^{-2k}S$. Assume now that $h = 1$. Then, by Lemmas 2 and 3 and the triangular inequality

$$\begin{aligned} |\gamma_k(h) - 0| &\leqslant 2^{-2k}\left|\sum_{i=1}^{I} \underbrace{f_i(h)}_{\leqslant i}\gamma(i)\right| + 2^{-2k}\left|\sum_{i=I+1}^{2^{k+1}-1} f_i(h)\gamma(i)\right| \\ &< 2^{-2k}S + 2^{-2k}\frac{\varepsilon}{2}\left|\sum_{i=I+1}^{2^{k+1}-1} f_i(h)\right| \\ &< \frac{\varepsilon}{2} + 2^{-2k}\frac{\varepsilon}{2}\left|\underbrace{\sum_{i=1}^{2^{k+1}-1} f_i(h)}_{=2^{2k}}\right| \leqslant \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \end{aligned}$$

which was required. ∎

The blocking method follows immediately. The theorem says that end users can apply blocking transformations to get the truncation error smaller than every $\varepsilon > 0$ if the time series is large enough.

Theorem 1 (The blocking method). Assume that the stationary time series $X_1, X_2, \ldots$ has autocovariance $\gamma(h) \to 0$ as $h \to \infty$. Then, for every $\varepsilon > 0$ there exists a natural number K such that the magnitude of $e_k < \varepsilon$ if $K \leqslant k \leqslant d$ for the time series $X_1, X_2, \ldots, X_{2^d}$.

Proof. Suppose $\varepsilon > 0$ is given. Since $\{\gamma_k\}_{k=1}^{\infty}$ converges uniformly and identically to zero on $\mathbb{N}$ by Proposition 2, there exists $K \in \mathbb{N}$ such that if $k \geqslant K$ then $\gamma_k < \varepsilon/2$ on $\mathbb{N}$.

Moreover, if $d \geqslant k$, then by the triangular inequality

$$|e_k| \leqslant \frac{2}{n_k}\sum_{h=1}^{n_k-1}\left(1-\frac{h}{n_k}\right)|\gamma_k(h)| \leqslant \frac{2}{n_k}\sum_{h=1}^{n_k-1}|\gamma_k(h)|$$
$$\leqslant \frac{\varepsilon}{n_k}\sum_{h=1}^{n_k-1}1 = \varepsilon\frac{n_k-1}{n_k} < \varepsilon,$$

which is the theorem. ■

III. AUTOMATING CALCULATIONS

The previous section provided proof that if $X_1, X_2, \ldots$ is a stationary time series, then the error $\{|e_k|\}_{k=0}^{\infty}$ is a decreasing sequence that converges to zero whenever $\gamma(h) \to 0$ as $h \to \infty$. The next objective is to provide an algorithm that automates calculations. Users that just want to know the algorithm can skip to Sec. III B. Else, I introduce hypothesis testing and maximum likelihood estimation, which will be required to understand the results from the the present section.

Preliminaries 3. It is necessary to discuss which distribution the random variables and vectors represent: $\mathbf{A} \sim \alpha(\theta)$ indicates that the vector $\mathbf{A}$ is *αdistributed with parameter* θ. In general, θ can have any dimension, and a 1×1 matrix or one-dimensional vector is a scalar. For example, let $\mathrm{N}(\mu, \Sigma)$ denote the *multivariate normal distribution* with expected value μ and covariance matrix Σ. If $\mathbf{Y} \sim \mathrm{N}(\mu, \Sigma)$ and Σ is positive definite, then the probability density function (pdf) of $\mathbf{Y}$ is an $\mathbb{R}^n \to \mathbb{R}$ function

$$f(\mathbf{y}) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{y}-\mu)^{\mathsf{T}}\Sigma^{-1}(\mathbf{y}-\mu)\right),$$

where $|\cdot|$ denotes the determinant [10]. It turns out that $\mathbf{Y}$ is multivariate normal if and only if every linear combination of the elements of $\mathbf{Y}$ is normally distributed [11]. The components of $\mathbf{Y}$ are independent if and only if there exists $\sigma_1, \ldots, \sigma_n > 0$ such that $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ [10]. A random variable $X_i \sim \chi_1^2$ is *χ^2 distributed* with 1 degree of freedom (later: dof) if it is the square of a *standard normal* random variable [3]. That means there exists a random variable $Z \sim \mathrm{N}(0, 1)$ such that $X_i = Z^2$. A sum of n independent χ^2 random variables with 1 dof, $\sum_{i=1}^{n} X_i$, is *χ^2 distributed* with n dof and its pdf is

$$g(x) = \frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}.$$

Conveniently, if Σ is invertible and $(Y_1, \ldots, Y_n)^{\mathsf{T}} = \mathbf{Y} \sim \mathrm{N}(\mu, \Sigma)$, then $(\mathbf{Y}-\mu)^{\mathsf{T}}\Sigma^{-1}(\mathbf{Y}-\mu) \sim \chi_n^2$ [*Proof.* Write $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X}-\mu)$. What is the distribution of $\mathbf{Z}$ and $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$?] If $X \sim \chi_\nu^2$, the $100(1-\alpha)$ *percentile* is the value $q_\nu(1-\alpha) \in \mathbb{R}$ such that $P[X > q_\nu(1-\alpha)] = \alpha$. χ^2 percentiles are tabulated in Table I. If $X_1, X_2, \ldots$ is a strictly stationary time series that is asymptotic uncorrelated such that $\mathrm{var}(X_i) < \infty$ and $\lim_{n\to\infty}\mathrm{var}(\sum_{i=1}^{n} X_i) = \infty$, and $\langle|X_i|^{2+k}\rangle < \infty$ for some $k > 0$, then [12] has proved that the central limit theorem holds [13].

The Fisherian approach to inference is the most common type of inference in natural sciences [14] and will be used in the following. Suppose there is a pdf $f(\mathbf{y};\theta)$ for $\mathbf{Y}$ that depends on a parameter θ and it is necessary to test whether there exists evidence that $\theta = \theta_0$ or if $\theta \neq \theta_0$ based on the observations $\mathbf{Y}$. Let $\widehat{\theta} = \widehat{\theta}(\mathbf{Y})$ be an estimator for θ. The hypothesis $\theta = \theta_0$ is called the *null hypothesis*, denoted H_0. It is common to pick H_0 such that the consequences of an incorrect test conclusion are minimized. The *alternative hypothesis* denoted by H_a is typically the negation of H_0.

TABLE I. χ^2 90, 95, and 99 percentiles: percentiles in the χ^2 distribution for $1 \leqslant d-k \leqslant 48$, which suffices for any error estimation with $\leqslant 10^{14}$ observations, at the three significance levels that performed best, $1-\alpha = 0.99,\ 0.95$, and 0.90. See for example [3] for additional values.

$d-k$	$q_{d-k}(0.99)$	$q_{d-k}(0.95)$	$q_{d-k}(0.9)$
1	6.634897	3.841459	2.705543
2	9.210340	5.991465	4.605170
3	11.344867	7.814728	6.251389
4	13.276704	9.487729	7.779440
5	15.086272	11.070498	9.236357
6	16.811894	12.591587	10.644641
7	18.475307	14.067140	12.017037
8	20.090235	15.507313	13.361566
9	21.665994	16.918978	14.683657
10	23.209251	18.307038	15.987179
11	24.724970	19.675138	17.275009
12	26.216967	21.026070	18.549348
13	27.688250	22.362032	19.811929
14	29.141238	23.684791	21.064144
15	30.577914	24.995790	22.307130
16	31.999927	26.296228	23.541829
17	33.408664	27.587112	24.769035
18	34.805306	28.869299	25.989423
19	36.190869	30.143527	27.203571
20	37.566235	31.410433	28.411981
21	38.932173	32.670573	29.615089
22	40.289360	33.924438	30.813282
23	41.638398	35.172462	32.006900
24	42.979820	36.415029	33.196244
25	44.31410	37.65248	34.38159
26	45.64168	38.88514	35.56317
27	46.96294	40.11327	36.74122
28	48.27824	41.33714	37.91592
29	49.58788	42.55697	39.08747
30	50.89218	43.77297	40.25602
31	52.19139	44.98534	41.42174
32	53.48577	46.19426	42.58475
33	54.77554	47.39988	43.74518
34	56.06091	48.60237	44.90316
35	57.34207	49.80185	46.05879
36	58.61921	50.99846	47.21217
37	59.89250	52.19232	48.36341
38	61.16209	53.38354	49.51258
39	62.42812	54.57223	50.65977
40	63.69074	55.75848	51.80506
41	64.95007	56.94239	52.94851
42	66.20624	58.12404	54.09020
43	67.45935	59.30351	55.23019
44	68.70951	60.48089	56.36854
45	69.95683	61.65623	57.50530
46	71.20140	62.82962	58.64054
47	72.44331	64.00111	59.77429
48	73.68264	65.17077	60.90661

Suppose there exists a $\mathbb{R}^k \to \mathbb{R}$ function $G(\widehat{\theta})$ such that G has known pdf g whenever H_0 is true, then $G(\widehat{\theta})$ is called a *test statistic*. The values of $G(\widehat{\theta})$ that are sufficiently unlikely according to g whenever H_0 is true are called the *rejection region*. And, if the estimated value $G(\widehat{\theta})$ is in the rejection region, H_0 is rejected in favor of H_a. Prior investigation $\alpha \in (0, 1)$ is chosen such that

$$P(\text{Rejecting } H_0 | H_0 \text{ is true}) \leqslant \alpha.$$

Here, $P(A|B)$ denotes conditional probability. Since g is known, this determines the size of the rejection region. It is convention to let $\alpha = 0.05$ and say that the test result is *significant* if the estimated value $G(\widehat{\theta})$ is in the rejection region. It is possible to determine the largest value of α such that the test concludes that H_0 is false. This value is called the *p value*, denoted p. The p value is a measure of the probability that it is a mistake to reject H_0 in favor of H_a. The *likelihood* L of θ is the function $L(\theta) = f(\mathbf{Y}; \theta)$. The estimator $\widehat{\theta}$ maximizing L is called the *maximum likelihood estimator*, and is asymptotically multivariate normal distributed [11]. The estimator $\widehat{\gamma}_k(1)$ is a maximum likelihood estimator if $\mathbf{X}$ is multivariate normal. The bias of an estimator $\widehat{\theta}$ given a parameter θ is defined as $\text{Bias}(\widehat{\theta}; \theta) = \langle \widehat{\theta} - \theta \rangle$ and measures how far from θ one can expect $\widehat{\theta}$. ▶

According to the variant of the central limit theorem introduced in Preliminaries 3, the elements of $\mathbf{X}_j$ are asymptotic multivariate normal if $\gamma(h) \to 0$ as $h \to \infty$ (in addition to some technical assumptions). This is because the elements of $\mathbf{X}_k$ are a mean of the elements of $\mathbf{X}$, which you can check. In that case, Preliminaries 3 say that $\widehat{\gamma}_j(1)$ is the maximum likelihood estimator of $\gamma_j(1)$. Hence, if $\widehat{\boldsymbol{\gamma}}_j \equiv (\widehat{\gamma}_j(1), \ldots, \widehat{\gamma}_{d-1}(1))$, then $\widehat{\boldsymbol{\gamma}}_j \sim \text{N}(\boldsymbol{\mu}_j, \Sigma_j)$ is asymptotic multivariate normal according to Preliminaries 3. The idea is to find the first index j such that $\gamma_k(1) = 0$ for all $k \geqslant j$ because, by Corollary 1, the error e_j becomes constant and there is no reason to expect that σ_k^2/n_k is a better estimate than σ_j^2/n_j for any $k > j$. To test this, define

$$M_j = (\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\mu}_j)^{\mathsf{T}} \Sigma_j^{-1} (\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\mu}_j) \sim \chi^2_{d-j}. \tag{12}$$

Hence, M_j has a known distribution (according to Preliminaries 3). This means that if $\boldsymbol{\mu}_j$ is evaluated in the case that $\gamma_k(1) = 0$ for all $k \geqslant j$, it is a test statistic for the hypothesis test:

$$\begin{aligned} &H_0 : \gamma_k(1) = 0 \text{ for all } k \geqslant j, \\ &H_a : \text{there exists } k \geqslant j \text{ such that } \gamma_k(1) \neq 0. \end{aligned} \tag{13}$$

The idea is to pick the smallest j such that the hypothesis test finds no evidence for H_a and take $\text{var}(\overline{X}) = \sigma_j^2/n_j$. Thus, an appropriate estimator is

$$\widehat{\text{var}}(\overline{X}) = \widehat{\sigma}_j^2/n_j. \tag{14}$$

This works according to Preliminaries 3: Whenever H_0 is true, the distribution of M_j is known [χ^2 with $d - j$ dof by Eq. (12)], so for all j such that a sufficiently improbable value of M_j is observed, the hypothesis test concludes that H_0 is false. However, once there is a j such that M_j is smaller than the $100(1 - \alpha)$ percentile, there is no longer evidence for H_a and the method concludes that H_0 is true,

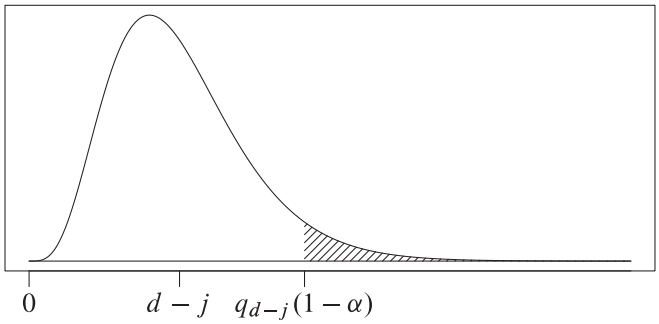

FIG. 1. Whenever H_0 is true, the pdf of M_j is known and plotted above. The test concludes that H_0 is false if the observed value of M_j is sufficiently unlikely. That is, if the observed value of M_j is larger than $100(1 - \alpha)$ percentile for a suitable α; the shaded area represents a probability of α. The value $d - j = \langle M_j \rangle$ is the expected value of M_j whenever H_0 is true since then $M_j \sim \chi^2_{d-j}$ is χ^2 distributed.

i.e., that the error becomes constant, and iterating further does not improve the estimate. See Fig. 1 for illustration. However, an expression of the covariance matrix Σ_j has to be determined. In this paper, the following approximation will be used:

$$\Sigma_j \simeq \text{diag}\big(\sigma_j^4/n_j, \cdots, \sigma_{d-1}^4/n_{d-1}\big). \tag{15}$$

In the next section, we will see that the approximation is obtained by (a) considering the elements only up to leading order in $1/n_k$ and (b) by setting all off-diagonal elements equal to zero. The benefit of the approximation is that inversion of Σ_j is easy. It is possible to question the expected error of setting the off-diagonal elements equal to zero. But, the expected error is zero, as Proposition 6 explains. However, before proving that, more work is required. We end this section with a proposition which says that the estimators of any blocking method, $\overline{X}$ and $\widehat{\sigma}_j^2/n_j$, are necessarily asymptotically unbiased in the following sense.

Proposition 3. Assume that $X_1, X_2, \ldots$ is a weakly stationary time series. Then, $\overline{X}$ is an unbiased estimator of $\langle X_i \rangle$ for all $i \in \mathbb{N}$ and

$$\text{Bias}(\widehat{\sigma}_j^2/n_j; \overline{X}) = -V(\overline{X})/n_j - e_j \quad \text{for all} \quad j \geqslant 0. \tag{16}$$

If in addition $\gamma(h) \to 0$ as $h \to \infty$, then for every $\varepsilon > 0$, there exists $K \in \mathbb{N}$ such that $|\text{Bias}(\widehat{\sigma}_K^2/n_K; \overline{X})| < \varepsilon$.

Proof. Whenever $X_1, X_2, \ldots$ is a weakly stationary time series, there exists $\mu \in \mathbb{R}$ such that $\langle X_i \rangle = \mu$ for all $i \in \mathbb{N}$, so

$$\langle \overline{X} \rangle = \left\langle \frac{1}{n} \sum_{i=1}^{n} X_i \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \langle X_i \rangle = \frac{1}{n} n\mu = \langle X_i \rangle.$$

To obtain the bias formula, use both the variance formula $\text{var}(Y) = \langle Y^2 \rangle - \langle Y \rangle^2$ $(*)$ and $\sum_{i=1}^{n_k} (X_k)_i = n\overline{X}$ $(\ddagger)$. Then, the following expression is obtained:

$$\begin{aligned} \big\langle \widehat{\sigma}_K^2 \big\rangle &= \frac{1}{n_K} \sum_{i=1}^{n_K} \langle [(X_K)_i - \overline{X}]^2 \rangle \\ &\overset{(\ddagger)}{=} \frac{1}{n_K} \sum_{i=1}^{n_K} \big[\big\langle (X_K)_i^2 \big\rangle \big] + \langle \overline{X}^2 \rangle - 2\langle \overline{X}^2 \rangle \overset{(*)}{=} \sigma_K^2 - \text{var}(\overline{X}). \end{aligned} \tag{17}$$

Use Eq. (4) and write

$$\langle \widehat{\mathrm{var}}(\overline{X})\rangle = \frac{\langle \widehat{\sigma}_j^2\rangle}{n_j} = \frac{\sigma_j^2}{n_j} - \frac{1}{n_j}\mathrm{var}(\overline{X}) = \mathrm{var}(\overline{X}) - e_j - \frac{1}{n_j}\mathrm{var}(\overline{X}). \tag{18}$$

Subtract $\mathrm{var}(\overline{X})$ from each side of the equation, and the bias formula (16) is obtained.

For the third part, suppose that $\varepsilon > 0$ is given. Since the hypothesis of Theorem 1 is true, there is a $K \in \mathbb{N}$ such that $|e_k| < \varepsilon/2$ if $K \leqslant k \leqslant d$ for the time series $X_1, X_2, \ldots, X_{2^d}$. Pick $k = K$ and let $d > 1 + K + \log_2 \mathrm{var}(\overline{X}) - \log_2 \varepsilon$. Because $d - K > 1 + \log_2 \mathrm{var}\,\overline{X} - \log_2 \varepsilon$,

$$\frac{\mathrm{var}(\overline{X})}{n_K} = \mathrm{var}(\overline{X})2^{-(d-K)} < \mathrm{var}(\overline{X})2^{\log_2(\varepsilon)-\log_2(\overline{X})-1} = \frac{\varepsilon}{2}.$$

Apply the triangular inequality and see

$$|\mathrm{Bias}(\widehat{\sigma}_K^2/n_K; \mathrm{var}(\overline{X}))| = \left|-\frac{\mathrm{var}(\overline{X})}{n_K} - e_K\right| \leqslant \left|\frac{\mathrm{var}(\overline{X})}{n_K}\right| + |e_K| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which was required. ■

A. Covariance matrix of $\widehat{\gamma}_i(h)$ and the matrix Σ_j

It is necessary to compute the covariance matrix of the estimators $\widehat{\gamma}(1)$ because the estimator of the hypothesis test from the previous section depends on it. It is impractical to compute the covariance matrix directly. However, developing linear algebra for the task is a fruitful alternative. Two lemmas and two propositions are required. The idea is to lay the foundation to apply the following theorem from the theory of quadratic forms of random variables.

Preliminaries 4. Assume $\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$ with Σ singular. If A, B are symmetric $n \times n$ matrices and there exists some $n \times r$ matrix Q of rank r such that $\Sigma = QQ^{\mathsf{T}}$, then

$$\mathrm{cov}(\mathbf{Y}^{\mathsf{T}}A\mathbf{Y}, \mathbf{Y}^{\mathsf{T}}B\mathbf{Y}) = 2\,\mathrm{Tr}(\Sigma A \Sigma B) + 4\boldsymbol{\mu}^{\mathsf{T}}A\Sigma B\boldsymbol{\mu}. \tag{19}$$

For proof, see [15]. ▶

Consider now a lemma that contains all the information required about probability distributions:

Lemma 4. Assume $\mathbf{1}$ denotes the vector of ones, $\mathbf{X} \sim \mathrm{N}(m\mathbf{1}, \sigma^2 I_n)$ and $\mathbf{Y} = \mathbf{X} - \overline{X}\mathbf{1}$. Then, $\mathbf{Y}$ is multivariate normal with expected value $\boldsymbol{\mu} = \mathbf{0}$, and there exists some $n \times (n-1)$ matrix Q of rank $n-1$ such that the covariance matrix $\Sigma_{\mathbf{Y}} = QQ^{\mathsf{T}}$ and

$$\Sigma_{\mathbf{Y}} = \frac{\sigma^2}{n}(nI_n - \mathbf{1}\mathbf{1}^{\mathsf{T}}). \tag{20}$$

Proof. First note that Y_i is a linear combination of elements of $\mathbf{X}$ because $\mathbf{X}$ is multivariate normal, which means that Y_i is univariate normal. This holds also for every linear combination of the elements of $\mathbf{Y}$, so $\mathbf{Y}$ is multivariate normal by Preliminaries 3. The expected value of $\mathbf{Y}$ is $\mathbf{0}$ since $\langle Y_i\rangle = \langle X_i - \overline{X}\rangle = m - m = 0$. To get Eq. (20), notice that the covariance matrix of $\mathbf{X}$ is diagonal, which means that the elements of $\mathbf{X}$ are independent since $\mathbf{X}$ is multivariate normal, and if δ_{ij} denotes the Kronecker delta, then the elements of Σ_Y are

$$\begin{aligned}(\Sigma_{\mathbf{Y}})_{ij} &= \mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(X_i - \overline{X}, X_j - \overline{X}) \\ &= \sigma^2\delta_{ij} - \frac{1}{n}\sum_{k=1}^{n}\mathrm{cov}(X_i, X_k) \\ &\quad - \frac{1}{n}\sum_{k=1}^{n}\mathrm{cov}(X_j, X_k) + \mathrm{var}\,\overline{X} \\ &= \sigma^2\delta_{ij} - \frac{1}{n}\sum_{k=1}^{n}\sigma^2\delta_{ik} - \frac{1}{n}\sum_{k=1}^{n}\sigma^2\delta_{jk} \\ &\quad + \frac{\sigma^2}{n} = \sigma^2\delta_{ij} - \frac{\sigma^2}{n}\end{aligned} \tag{21}$$

only if $\Sigma_{\mathbf{Y}} = (\sigma^2/n)(nI_n - \mathbf{1}\mathbf{1}^{\mathsf{T}})$. This proves that $\Sigma_{\mathbf{Y}}$ is symmetric. Note that $\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{1} = n\mathbf{1}$, so $\mathbf{1}$ is an eigenvector of $\Sigma_{\mathbf{Y}}$ with eigenvalue 0. Furthermore, if $k \in \{1, 2, \ldots, n-1\}$ and $\mathbf{q}_k = \mathbf{e}_k - \mathbf{e}_n$, then

$$\Sigma_{\mathbf{Y}}\mathbf{q}_k \overset{(21)}{=} \frac{\sigma^2}{n}\left[n\mathbf{q}_k - \mathbf{1}(\underbrace{\mathbf{1}^{\mathsf{T}}\mathbf{e}_k - \mathbf{1}^{\mathsf{T}}\mathbf{e}_n}_{=1-1=0})\right] = \sigma^2\mathbf{q}_k,$$

which proves that σ^2 is an eigenvalue of $\Sigma_{\mathbf{Y}}$ with multiplicity $n-1$. And since $\Sigma_{\mathbf{Y}}$ is symmetric, it has a spectral decomposition [16]:

$$\Sigma_{\mathbf{Y}} = \sigma^2\sum_{k=1}^{n-1}\mathbf{q}_k\mathbf{q}_k^{\mathsf{T}} = \sigma^2[\mathbf{q}_1\ \mathbf{q}_2\ \ldots\ \mathbf{q}_{n-1}][\mathbf{q}_1^{\mathsf{T}}\ \mathbf{q}_2^{\mathsf{T}}\ \ldots\ \mathbf{q}_{n-1}^{\mathsf{T}}]^{\mathsf{T}}.$$

So if $Q = \sigma[\mathbf{q}_1\ \mathbf{q}_2\ \ldots\ \mathbf{q}_{n-1}]$, then Q is an $n \times (n-1)$ matrix and $\Sigma_{\mathbf{Y}} = QQ^{\mathsf{T}}$. Moreover, according to the spectral theorem [16], the dimension of $\mathrm{span}\{\mathbf{q}_1, \ldots, \mathbf{q}_{n-1}\}$ equals the multiplicity of σ^2. That means the columns of Q are $n-1$ linearly independent vectors, which also equals its rank. ■

Define transformations $S_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ and $T_i : \mathbb{R}^{n_{i-1}} \to \mathbb{R}^{n_i}$ with standard matrices

$$S_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix},$$

$$T_i = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{22}$$

According to Eq. (1), the matrices T_i generate the observations $\mathbf{X}_k$ subject to i blocking transformations by

$$\mathbf{X}_i = T_i T_{i-1} \ldots T_1 \mathbf{X}. \tag{23}$$

Using the matrices $\{S_i\}_{i=0}^{d-1}$ and $\{T_i\}_{i=1}^{d-1}$, define the $n \times n$ matrices $\{\Gamma_i\}_{i=0}^{d-1}$ by

$$\Gamma_i = \frac{1}{2}\frac{1}{n_i} T_1^\mathsf{T} T_2^\mathsf{T} \dots T_i^\mathsf{T} \big(S_i + S_i^\mathsf{T}\big) T_i \dots T_1. \tag{24}$$

According to the following proposition, these matrices are interesting because they generate the estimator $\widehat{\gamma}_i(1)$ from the vector $\mathbf{Y}$ whose probability distribution is multivariate normal.

Proposition 4. The matrices $\{\Gamma_i\}$ are symmetric. Hence, if $\mathbf{Y} = \mathbf{X} - \overline{X}\mathbf{1}$, then $\mathbf{Y}^\mathsf{T}\Gamma_i\mathbf{Y}$ is a quadratic form and $\mathbf{Y}^\mathsf{T}\Gamma_i\mathbf{Y} = \widehat{\gamma}_i(1)$.

Proof. Fix $0 \leqslant i \leqslant d-1$. It is clear that Γ_i is symmetric by construction:

$$\begin{aligned}\Gamma_i^\mathsf{T} &= \frac{1}{2}\frac{1}{n_i}\big(T_1^\mathsf{T} T_2^\mathsf{T} \dots T_i^\mathsf{T}\big(S_i + S_i^\mathsf{T}\big) T_i \dots T_1\big)^\mathsf{T} \\ &= \frac{1}{2}\frac{1}{n_i}(T_i \dots T_1)^\mathsf{T}\big(S_i + S_i^\mathsf{T}\big)^\mathsf{T}\big(T_1^\mathsf{T} T_2^\mathsf{T} \dots T_i^\mathsf{T}\big)^\mathsf{T} = \Gamma_i.\end{aligned}$$

That means $\mathbf{Y}^\mathsf{T}\Gamma_i\mathbf{Y}$ is a quadratic form. It remains to prove $\mathbf{Y}^\mathsf{T}\Gamma_i\mathbf{Y} = \widehat{\gamma}_i(1)$. First, use the definition of blocking transformation and that any real number equals it own transpose to obtain

$$\begin{aligned} n_i \mathbf{Y}^\mathsf{T}\Gamma_i\mathbf{Y} &\overset{(24)}{=} \tfrac{1}{2}\mathbf{Y}^\mathsf{T} T_1^\mathsf{T} \dots T_i^\mathsf{T} S_i T_i \dots T_1 \mathbf{Y} \\ &\quad + \tfrac{1}{2}\big[\mathbf{Y}^\mathsf{T} T_1^\mathsf{T} \dots T_i^\mathsf{T} S_i T_i \dots T_1 \mathbf{Y}\big]^\mathsf{T} \\ &= [T_i \dots T_1 \mathbf{Y}]^\mathsf{T} S_i T_i \dots T_1 \mathbf{Y} \overset{(23)}{=} \mathbf{Y}_i^\mathsf{T} S_i \mathbf{Y}_i. \end{aligned} \tag{25}$$

Second, fix $1 \leqslant k \leqslant n_i$ and use induction to see $(\mathbf{Y}_i)_k = (\mathbf{X}_i)_k - \overline{X}_i$. The base case is satisfied by hypothesis and the induction step follows by

$$\begin{aligned} (\mathbf{Y}_{i+1})_k &\overset{(1)}{=} \tfrac{1}{2}[(\mathbf{Y}_i)_{2k-1} + (\mathbf{Y}_i)_{2k}] \\ &= \tfrac{1}{2}[(\mathbf{X}_i)_{2k-1} - \overline{X}_i + (\mathbf{X}_i)_{2k} - \overline{X}_i] \\ &\overset{(1)}{=} (\mathbf{X}_{i+1})_k - 2\tfrac{1}{2}\overline{X}_i \overset{\text{Lemma (1)}}{=} (\mathbf{X}_{i+1})_k - \overline{X}_{i+1}, \end{aligned} \tag{26}$$

where Lemma 1 was used twice to get $\overline{X}_i = \overline{X} = \overline{X}_{i+1}$. Third, using the definition of the matrices S_i from Eq. (22) notice that S_i shifts the indices of vectors by one:

$$\begin{aligned} n_i \mathbf{Y}^\mathsf{T}\Gamma_i\mathbf{Y} &\overset{(25)}{=} \mathbf{Y}_i^\mathsf{T} S_i \mathbf{Y}_i = \mathbf{Y}_i^\mathsf{T}(S_i \mathbf{Y}_i) \\ &= ((\mathbf{Y}_i)_1, \cdots, (\mathbf{Y}_i)_{n_i})^\mathsf{T}((\mathbf{Y}_i)_2, \cdots, (\mathbf{Y}_i)_{n_i}, 0) \\ &= \sum_{h=1}^{n_i - 1} (\mathbf{Y}_i)_{h+1}(\mathbf{Y}_i)_h \overset{(2)(26)}{=} n_i \widehat{\gamma}_i(1), \end{aligned} \tag{27}$$

using $(\mathbf{Y}_i)_k = (\mathbf{X}_i)_k - \overline{X}_i$ and the definition of $\widehat{\gamma}_i(1)$ in the final step. ∎

Experts will immediately see that the above result is easily generalized to $\widehat{\gamma}_k(h)$ for any $h \geqslant 0$ by considering the operators S_k^h by raising to a power $h \in \mathbb{Z}$. But, according to Proposition 1, it suffices to consider $h = 1$. Some readers may later aim to compute the mean-squared error (MSE) of the blocking method estimator $\widehat{\mathrm{var}}(\overline{X})$. In doing so, they may want to repeat calculations for $h = 0$ after reading Sec. III B. In the case of $h = 1$, we have a technical lemma; the following three quantities determine the expression for the covariance matrix of $\widehat{\gamma}_k(1)$.

Lemma 5. If $\mathbf{1}$ denotes the vector of ones and $i \geqslant j$, then Γ_i and Γ_j constitute the following:

$$\begin{aligned} n_i \mathbf{1}^\mathsf{T}\Gamma_i\mathbf{1} &= n_i - 1, \\ n^2 n_i n_j \operatorname{Tr}[\Gamma_i\Gamma_j] &= \tfrac{1}{2}n_i^2(n_i - 1), \\ 2nn_i n_j \mathbf{1}^\mathsf{T}\Gamma_i\Gamma_j\mathbf{1} &= 2n_j(n_i - 1) - n_i. \end{aligned}$$

Proof. The following is used throughout: If $\{\mathbf{e}_k\}_{k=1}^{n_i}$ denotes the standard basis of $\mathbb{R}^{n_i}$, then according to Eq. (22),

$$S_i\mathbf{e}_k = \begin{cases} \mathbf{0} & \text{if } k = 1, \\ \mathbf{e}_{k-1} & \text{else} \end{cases} \quad \text{and} \quad S_i^\mathsf{T}\mathbf{e}_k = \begin{cases} \mathbf{0} & \text{if } k = n_i, \\ \mathbf{e}_{k+1} & \text{else.} \end{cases}$$

By multiplying T_i by each vector from $\{\mathbf{e}_k\}_{k=1}^{n_i}$ and summing over k, it is clear that

$$\begin{aligned} T_i \sum_{k=1}^{n_{i-1}} \mathbf{e}_k &= T_i\mathbf{e}_1 + T_i\mathbf{e}_2 + T_i\mathbf{e}_3 + \cdots + T_i\mathbf{e}_{n_{i-1}} \\ &\overset{(22)}{=} \frac{1}{2}\mathbf{e}_1 + \frac{1}{2}\mathbf{e}_1 + \frac{1}{2}\mathbf{e}_2 + \cdots \frac{1}{2}\mathbf{e}_{n_i} = \sum_{k=1}^{n_i} \mathbf{e}_k. \end{aligned} \tag{28}$$

Write $\mathbf{1}$ as $\sum_{u=1}^{n} \mathbf{e}_u = \mathbf{1}$, and get the first equation:

$$\begin{aligned} n_i \mathbf{1}^\mathsf{T}\Gamma_i\mathbf{1} &= \frac{1}{2}\sum_{u=1}^{n} \mathbf{e}_u^\mathsf{T} T_1^\mathsf{T} T_2^\mathsf{T} \dots T_i^\mathsf{T}\big(S_i + S_i^\mathsf{T}\big) T_i \dots T_1 \sum_{v=1}^{n} \mathbf{e}_v \\ &\overset{(28)}{=} \frac{1}{2}\sum_{u=1}^{n_i}\sum_{v=1}^{n_i} \mathbf{e}_u^\mathsf{T}\big(S_i + S_i^\mathsf{T}\big)\mathbf{e}_v \\ &= \frac{1}{2}\sum_{u=1}^{n_i}\sum_{v=2}^{n_i} \mathbf{e}_u^\mathsf{T}\mathbf{e}_{v-1} + \frac{1}{2}\sum_{u=1}^{n_i}\sum_{v=1}^{n_i - 1} \mathbf{e}_u^\mathsf{T}\mathbf{e}_{v+1} = n_i - 1, \end{aligned}$$

where orthonormality of $\{\mathbf{e}_k\}$ was used in the final step. Next, it is necessary to show $T_i T_i^\mathsf{T} = (1/2)I_{n_i}$ for all $1 \leqslant i \leqslant d-1$. To see this is true, write T_i as a Kronecker product $T_i = (1/2)I_{n_i} \otimes (1, 1)$ and use the mixed product rule [17]:

$$T_i T_i^\mathsf{T} = \frac{1}{4}(I_{n_i} \otimes (1,1))(I_{n_i}^\mathsf{T} \otimes (1,1)^\mathsf{T}) = \frac{1}{4}\underbrace{I_{n_i}^2}_{I_{n_i}}\underbrace{(1,1)(1,1)^\mathsf{T}}_{=2}.$$

Using this and working in a similar way as before, the following is obtained:

$$2nn_i n_j \mathbf{1}^\mathsf{T}\Gamma_i\Gamma_j\mathbf{1} = 2n_i n_j - n_i - 2n_j. \tag{29}$$

Prove now two more properties of $\{\mathbf{e}_k\}$: First, see that if M is any $n \times n$, then a diagonal element $m_{kk} = \mathbf{e}_k^\mathsf{T} M \mathbf{e}_k$, so $\operatorname{Tr}(M) = \sum_{k=1}^{n} \mathbf{e}_k^\mathsf{T} M \mathbf{e}_k$, as you can check. Second, if M is a $n_{j+h} \times n_{j+h}$ matrix, then there is a real number $K \in \mathbb{R}$ such that

$$\begin{aligned} &\sum_{k=1}^{n_j - 1} \mathbf{e}_k^\mathsf{T} T_{j+1}^\mathsf{T} \dots T_{j+h}^\mathsf{T} M T_{j+h} \dots T_{j+1} \mathbf{e}_{k+1} \\ &\qquad = 2^{-2h} \sum_{k=1}^{n_{j+h} - 1} \mathbf{e}_k^\mathsf{T} M \mathbf{e}_{k+1} + K \sum_{k=1}^{n_{j+h}} \mathbf{e}_k^\mathsf{T} M \mathbf{e}_k. \end{aligned} \tag{30}$$

Prove this by induction. If $h = 1$, then

$$\begin{aligned}\sum_{k=1}^{n_j-1} \mathbf{e}_k^\mathsf{T} T_{j+1}^\mathsf{T} M T_{j+1} \mathbf{e}_{k+1} &= \frac{1}{4}\mathbf{e}_1^\mathsf{T} M \mathbf{e}_1 + \frac{1}{4}\mathbf{e}_1^\mathsf{T} M \mathbf{e}_2 \\ &\quad + \cdots + \frac{1}{4}\mathbf{e}_{n_{j+1}-1}^\mathsf{T} M \mathbf{e}_{n_{j+1}} + \frac{1}{4}\mathbf{e}_{n_{j+1}}^\mathsf{T} M \mathbf{e}_{n_{j+1}} \\ &= 2^{-2}\sum_{k=1}^{n_{j+1}-1} \mathbf{e}_k^\mathsf{T} M \mathbf{e}_{k+1} + \frac{1}{4}\sum_{k=1}^{n_{j+1}} \mathbf{e}_k^\mathsf{T} M \mathbf{e}_k,\end{aligned}$$

which proves the base case. To get the induction step, assume the hypothesis is true for h then define the matrix $N = T_{j+h+1}^\mathsf{T} M T_{j+h+1}$. This matrix is $n_{j+h} \times n_{j+h}$, so it is possible to use it in the place of the matrix M. Then, use the same procedure as before to prove the result for $h + 1$.

To get the final equation from the lemma, use again that $T_j T_j^\mathsf{T} = 2^{-1} I_{n_j}$, as well as cyclic permutation of the factors and write $\mathrm{Tr}[\Gamma_i \Gamma_j]$ in the following way:

$$\begin{aligned}4 n_i n_j \,\mathrm{Tr}[\Gamma_i \Gamma_j] &\overset{(24)}{=} 2^{-2j}\,\mathrm{Tr}\left[T_{j+1}^\mathsf{T} \ldots T_i^\mathsf{T}\left(S_i + S_i^\mathsf{T}\right)\right. \\ &\quad \left. \times\, T_i \ldots T_{j+1}\left(S_j + S_j^\mathsf{T}\right)\right]. \qquad (31)\end{aligned}$$

Distribute the terms in the trace. One of the terms is $\mathrm{Tr}[T_{j+1}^\mathsf{T} \ldots T_i^\mathsf{T} S_i T_i \ldots T_{j+1} S_j^\mathsf{T}]$. To evaluate it, use what was just proven and write

$$\begin{aligned}&\sum_{k=1}^{n_j} \mathbf{e}_k^\mathsf{T} T_{j+1}^\mathsf{T} \ldots T_i^\mathsf{T} S_i T_i \ldots T_{j+1} S_j^\mathsf{T} \mathbf{e}_k \\ &\quad = \sum_{k=1}^{n_j-1} \mathbf{e}_k^\mathsf{T} T_{j+1}^\mathsf{T} \ldots T_i^\mathsf{T} S_i T_i \ldots T_{j+1} \mathbf{e}_{k+1} \\ &\quad \overset{(30)}{=} 4^{-(i-j)} \sum_{k=1}^{n_i-1} \mathbf{e}_k^\mathsf{T} S_i \mathbf{e}_{k+1} + K \sum_{k=1}^{n_i} \underbrace{\mathbf{e}_k^\mathsf{T} S_i \mathbf{e}_k}_{\mathbf{e}_k^\mathsf{T}\mathbf{e}_{k-1}=0}.\end{aligned}$$

This term equals $4^{-(i-j)}(n_i - 1)$ since $\mathbf{e}_k^\mathsf{T} S_i \mathbf{e}_{k+1} = \mathbf{e}_k^\mathsf{T}\mathbf{e}_k = 1$. Make the replacement $S_i \mapsto S_i^\mathsf{T}$ throughout the above equation, in which case the term will evaluate to zero since $\mathbf{e}_k^\mathsf{T} S_i^\mathsf{T} \mathbf{e}_{k+1} = \mathbf{e}_k^\mathsf{T}\mathbf{e}_{k+2} = 0$. The third and fourth terms from Eq. (31) are evaluated in a similar way. The sum of all four terms is $2 \times 4^{-(i-j)}(n_i - 1)$, hence,

$$\begin{aligned}n^2 n_i n_j \,\mathrm{Tr}[\Gamma_i \Gamma_j] &\overset{(31)}{=} 2\tfrac{1}{4} 2^{-2j} n^2 4^{-(i-j)}(n_i - 1) \\ &= \tfrac{1}{2} n_i^2 (n_i - 1),\end{aligned}$$

which is the final part of the lemma. ■

The following proposition gives an expression for the elements of the covariance matrix of $\widehat{\gamma}_i(1)$.

Proposition 5. If there is some $m \in \mathbb{R}$ such that the vector $\mathbf{X} \sim \mathrm{N}(m\mathbf{1}, \sigma^2 I_n)$, then the expected value of $\widehat{\gamma}_i(1)$ is $-\sigma_i^2(n_i - 1)/n_i^2$. Furthermore, the covariance matrix of $(\widehat{\gamma}_0(1), \ldots, \widehat{\gamma}_{d-1}(1))^\mathsf{T}$ has elements

$$\mathrm{cov}(\widehat{\gamma}_i(1), \widehat{\gamma}_j(1)) = 2\left(\frac{\sigma_i \sigma_j}{n_i n_j}\right)^2 \left[1 + (n_i - 1)\left(\frac{1}{2} n_i^2 - n_j\right)\right]$$

whenever $0 \leqslant j \leqslant i \leqslant d - 1$.

Proof. Assume $0 \leqslant j \leqslant i \leqslant d - 1$. To obtain the expectation, use the defining equation (2) and notice that the elements of $\mathbf{X}_i$ are independent by hypothesis, so

$$\begin{aligned}n_i \langle \widehat{\gamma}_i(1) \rangle &= \sum_{j=1}^{n_i-1} \underbrace{\langle (\mathbf{X}_i)_j (\mathbf{X}_i)_{j+1} \rangle}_{\gamma_i(1) + m^2 = 0 + m^2} + \underbrace{\langle \overline{X}^2 \rangle}_{\frac{\sigma_i^2}{n_i} + m^2} - \underbrace{\langle [(\mathbf{X}_i)_j + (\mathbf{X}_i)_{j+1}] \overline{X} \rangle}_{2(m^2 + \sigma_i^2/n_i)},\end{aligned}$$

and the first part is proven. Assume now that $\mathbf{Y} = \mathbf{X} - \overline{X}\mathbf{1}$. To get the covariance matrix, note that by Lemma 4, $\mathbf{Y}$ is multivariate normal with expected value $\mathbf{0}$ and there exists a $n \times (n - 1)$ matrix Q of rank $n - 1$ such that the covariance matrix $\Sigma_\mathbf{Y} = QQ^\mathsf{T}$. According to Proposition 4, Γ_i and Γ_j are symmetric, so according to Preliminaries 4,

$$\begin{aligned}\mathrm{cov}(\widehat{\gamma}_i(1), \widehat{\gamma}_j(1)) &= \mathrm{cov}(\mathbf{Y}^\mathsf{T}\Gamma_i \mathbf{Y}, \mathbf{Y}^\mathsf{T}\Gamma_j \mathbf{Y}) \\ &\overset{(19)}{=} 2\,\mathrm{Tr}(\Sigma_\mathbf{Y}\Gamma_i \Sigma_\mathbf{Y}\Gamma_j). \qquad (32)\end{aligned}$$

Recall that it is possible to cyclic permute the elements of a trace and that $\Gamma_i = \Gamma_j^\mathsf{T}$, and $\mathrm{Tr}(M) = \mathrm{Tr}(M^\mathsf{T})$ and $\mathrm{Tr}(\mathbf{1}^\mathsf{T} M \mathbf{1}) = \mathbf{1}^\mathsf{T} M \mathbf{1}$ (since it is a real number) for all square matrices M. Using this and Lemma 4, write $(n^2/\sigma^4)\mathrm{Tr}(\Sigma_\mathbf{Y}\Gamma_i \Sigma_\mathbf{Y}\Gamma_j)$ in the following way:

$$\begin{aligned}\frac{n^2}{\sigma^4}\mathrm{Tr}(\Sigma_\mathbf{Y}\Gamma_i \Sigma_\mathbf{Y}\Gamma_j) &= n^2\,\mathrm{Tr}(\Gamma_i \Gamma_j) + \mathbf{1}^\mathsf{T}\Gamma_i \mathbf{1}\mathbf{1}^\mathsf{T}\Gamma_j \mathbf{1} \\ &\quad - 2n\mathbf{1}^\mathsf{T}\Gamma_i \Gamma_j \mathbf{1}.\end{aligned}$$

To complete the proof, use now Lemma 5 and that $n_i = n/2^i$. Furthermore, since the elements of $\mathbf{X}$ are independent, $\sigma_j^2 = \sigma^2/2^j$ by Corollary 1. Thus,

$$\begin{aligned}\mathrm{Tr}(\Sigma_\mathbf{Y}\Gamma_i \Sigma_\mathbf{Y}\Gamma_j) &= \frac{\sigma^4}{n^2}[n_i\ n_j]^{-1}\left[\frac{1}{2} n_i^2 (n_i - 1) + (n_i - 1)(n_j - 1)\right. \\ &\quad \left. + n_i - 2n_j(n_i - 1)\right] \\ &= \left(\frac{\sigma_i \sigma_j}{n_i n_j}\right)^2 \left[1 + (n_i - 1)\left(\frac{1}{2} n_i^2 - n_j\right)\right].\end{aligned}$$

Multiply each side of the equation by 2 and recall Eq. (32) above, which proves the proposition true. ■

As we discussed, the diagonal elements of the covariance matrix of $\widehat{\gamma}_i(1)$ are of special interest to us, and they are given by the following corollary.

Corollary 3. Assume there is some $m \in \mathbb{R}$ such that the vector $\mathbf{X} \sim \mathrm{N}(m\mathbf{1}, \sigma^2 I_n)$. Then, the variance of $\widehat{\gamma}_j(k)$ is exactly

$$\mathrm{var}(\widehat{\gamma}_j(1)) = \left(\frac{\sigma_j}{n_j}\right)^4 [2 + n_j(n_j - 1)(n_j - 2)],$$

whenever $0 \leqslant j \leqslant d - 1$.

It remains to determine the error of discarding the off-diagonal elements of Σ_j in Eq. (15). In order to understand how the error may influence the calculation, recall that the purpose of finding Σ_j is to evaluate M_j. If we form a diagonal matrix Σ'_j that contains the diagonal elements of Σ_j and

make the approximation

$$M'_j \equiv (\widehat{\gamma}_j - \boldsymbol{\mu}_j)^{\mathsf{T}} {\Sigma'_j}^{-1} (\widehat{\gamma}_j - \boldsymbol{\mu}_j),$$

then the expected error $\langle M_j - M'_j \rangle$ is zero according to the following proposition:

Proposition 6. If $E_j = M_j - M'_j$, then $\langle E_j \rangle = 0$.

Proof. Set $G_j = \widehat{\gamma}_j - \boldsymbol{\mu}_j$ and use the definition of Σ_j to write Σ_j in a convenient way:

$$\Sigma_j = \langle (\widehat{\gamma}_j - \boldsymbol{\mu}_j)(\widehat{\gamma}_j - \boldsymbol{\mu}_j)^{\mathsf{T}} \rangle = \langle G_j G_j^{\mathsf{T}} \rangle. \tag{33}$$

In this notation we can write $M_j = G_j^{\mathsf{T}} {\Sigma_j}^{-1} G_j$ and similarly for M'_j. Since E_j is a 1×1 matrix and using that the trace is invariant under cyclic permutation of its elements (†),

$$\begin{aligned} E_j = \operatorname{Tr} E_j = \operatorname{Tr}[M_j - M'_j] &= \operatorname{Tr}\left[G_j^{\mathsf{T}} \left(\Sigma_j^{-1} - {\Sigma'_j}^{-1}\right) G_j\right] \\ &\overset{(\dagger)}{=} \operatorname{Tr}\left[\left(\Sigma_j^{-1} - {\Sigma'_j}^{-1}\right) G_j G_j^{\mathsf{T}}\right]. \end{aligned}$$

It follows directly from the definition of matrix trace and linearity of expected value that if A is a (nonrandom) matrix then $\langle \operatorname{Tr} A G_j G_j^{\mathsf{T}} \rangle = \operatorname{Tr} \langle A G_j G_j^{\mathsf{T}} \rangle = \operatorname{Tr} A \langle G_j G_j^{\mathsf{T}} \rangle$ [10]. That means

$$\langle E_j \rangle = \operatorname{Tr}\left[\left(\Sigma_j^{-1} - {\Sigma'_j}^{-1}\right)\langle G_j G_j^{\mathsf{T}} \rangle\right] \overset{(33)}{=} \operatorname{Tr}\left[\left(I - {\Sigma'_j}^{-1} \Sigma_j\right)\right]. \tag{34}$$

Define $m = d - j$, then Σ_j and Σ'_j are $m \times m$ matrices. Also, note that by assumption, Σ'_j is the diagonal matrix consisting of diagonal elements of Σ_j. That means (1) the inverse of Σ'_j is a diagonal matrix, and (2) it has diagonal elements that are the multiplicative inverse of the diagonal elements of Σ_j. So,

$$\left({\Sigma'_j}^{-1}\right)_{ik} = (\Sigma_j)^{-1}_{ik} \delta_{ik}.$$

Consequently, each diagonal element of ${\Sigma'_j}^{-1} \Sigma_j$ equals 1 because

$$\begin{aligned} \left({\Sigma'_j}^{-1} \Sigma_j\right)_{ii} &= \sum_{k=1}^{m} \left({\Sigma'_j}^{-1}\right)_{ik} (\Sigma_j)_{ki} = \sum_{k=1}^{m} (\Sigma_j)^{-1}_{ik} \delta_{ik} (\Sigma_j)_{ki} \\ &= (\Sigma_j)^{-1}_{ii} (\Sigma_j)_{ii} = 1. \end{aligned} \tag{35}$$

So, since the matrix trace is the sum of all the diagonal elements,

$$\operatorname{Tr} {\Sigma'_j}^{-1} \Sigma_j = \sum_{i=1}^{m} \left({\Sigma'_j}^{-1} \Sigma_j\right)_{ii} \overset{(35)}{=} \sum_{i=1}^{m} 1 = m \tag{36}$$

and has the consequence that

$$\langle E_j \rangle \overset{(34)(33)}{=} \underbrace{\operatorname{Tr} I}_{=m} - \operatorname{Tr} {\Sigma'_j}^{-1} \Sigma_j \overset{(36)}{=} m - m = 0,$$

which is the proposition. ∎

Applying this proposition in combination with Corollary 3 to leading order in $1/n_k$ provides justification for the use of $\operatorname{diag}(\sigma_j^4/n_j, \cdots, \sigma_{d-1}^4/n_{d-1})$ in the place of Σ_j in computation of M_j.

B. Algorithm

By summarizing the results from the two previous sections, the following is clear:

Theorem 2. If $X_1, X_2, \ldots$ is a strictly stationary time series such that $\gamma(h) \to 0$ as $h \to \infty$ with $\lim_{n\to\infty} \operatorname{var}(\sum_{i=1}^{n} X_i) = \infty$ and $\langle |X_i|^{2+a} \rangle < \infty$, for some $a > 0$, then M_j is a test statistic that is asymptotic χ^2_{d-j} distributed under the hypothesis $\gamma_k(1) = 0$ for all $k \geqslant j$. The rejection region includes all values M_j larger than $q_{d-j}(1-\alpha)$ for all $1 \leqslant j \leqslant d-1$.

The above theorem outlines the algorithm. Form a vector $\mathbf{X}$ consisting of 2^d observations. The algorithm proceeds as follows: Compute $\widehat{\sigma}_i^2$ and $\widehat{\gamma}_i(1)$ for $\mathbf{X}_i$ for each $i \in \{0, 1, \ldots, d-1\}$. Then, form M_i for all $i \in \{0, 1, \ldots, d-1\}$ using the estimates $\widehat{\sigma}_i^2$ and $\widehat{\gamma}_i(1)$. Using the results from the two previous sections, M_j becomes

$$M_j = \sum_{k=j}^{d-1} \frac{n_k \left[(n_k - 1)\widehat{\sigma}_k^2 \big/ \left(n_k^2\right) + \widehat{\gamma}_k(1)\right]^2}{\widehat{\sigma}_k^4}.$$

Pick some significance level α. It is convention in inference to let $\alpha = 0.05$, but it is possible to pick some other value. Then, compare M_k to $q_{d-k}(1-\alpha)$ for all k. Choose the smallest k such that $M_k \leqslant q_{d-k}(1-\alpha)$. Using this k, make the final estimate for the variance $\operatorname{var}(\overline{X}) = \widehat{\sigma}_k^2/n_k$.

The method has built in safety features (see Fig. 6). This is necessary because the method may operate without supervision. If the conditions above are not met, the method may fail. In case this happens, it is necessary to present a warning to the end user or application so they can take necessary action. Recall the conditions for the method (see Theorem 2): (i) the time series is strictly stationary, (ii) $\gamma(h) \to 0$ as $h \to \infty$, and (iii) the χ^2 approximation works. Therefore, if the method does not conclude that H_0 is true for any $k \leqslant d-1$, one of these are false, and the fault is caught with an `if` test. The conditions (i) and (ii) are either present or not by construction and, as such, end users will know whether these are satisfied or not. However, condition (iii) can fail if there are little data available. So, the bottom line is that the end user should check whether conditions (i) and (ii) are satisfied in order to guarantee that the method works reliably. If the end user is unsure if the time series is stationary, there exist statistical tests for stationary, such as the Dickey-Fuller test, which can be used. It is usually easy to check condition (ii) by estimating the autocovariance matrix by $\widehat{\gamma}(h)$ using a suitably small chunk of the data. See Fig. 6 for sample code of one implementation, or use flow chart of Fig. 2 for an overview.

The method is asymptotically unbiased in the way one would expect according to Proposition 3, which also gives an equation for the bias. Using that and Eq. (4) gives

$$\begin{aligned} \operatorname{Bias}(\widehat{\operatorname{var}}(\overline{X}); \operatorname{var}(\overline{X})) &\overset{(16)}{=} -\frac{\operatorname{var}(\overline{X})}{n_K} - e_K \\ &\overset{(4)}{=} -\operatorname{var}(\overline{X})\left(1 + \frac{1}{n_K}\right) + \frac{\sigma_K^2}{n_k}. \end{aligned} \tag{37}$$

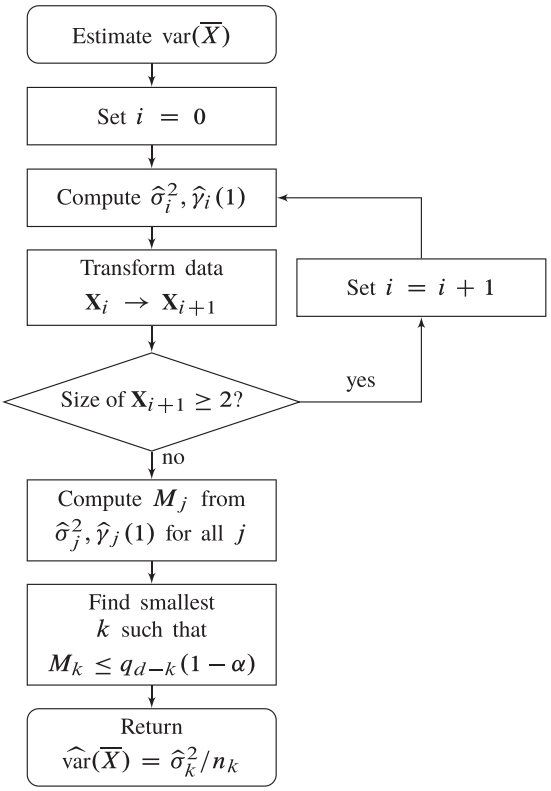

FIG. 2. Flow chart of algorithm. The idea is to return the estimate of var($\overline{X}$), as $\widehat{\sigma}_k^2/n_k$ for the smallest value of k such that there is no evidence that $\gamma_j(1) \neq 0$ for $j > k$. This is sensible because then, according to Corollary 1, there is no reason to believe that error e_k is reduced by further iterations of the method.

By estimating each quantity separately, this is the resulting estimator of the bias of $\widehat{\text{var}}(\overline{X})$ given var($\overline{X}$):

$$\widehat{\text{Bias}} \overset{(37)}{=} -\widehat{\text{var}}(\overline{X})\left(1+\frac{1}{n_K}\right)+\frac{\widehat{\sigma}_K^2}{n_K} \overset{(14)}{=} -\frac{\widehat{\text{var}}(\overline{X})}{n_K}.$$

The estimator can be used to survey the performance of the method if bias is critical to the application. In addition, it turns out to be relatively easy to compute the variance of the estimator $\widehat{\text{var}}(\overline{X})$. Summing the bias squared and the variance of an estimator yields the mean-squared error (MSE) [3].

Preliminaries 5. MSE is a risk function corresponding to the expected value of the squared error loss, and one of the most popular measures of the performance of an estimator. According to [3], if $\widehat{\theta}$ is any estimator and θ a parameter, then

$$\text{MSE}(\widehat{\theta};\theta) = \langle(\widehat{\theta}-\theta)^2\rangle = \text{Bias}^2(\widehat{\theta};\theta)+\text{var}(\widehat{\theta}). \quad (38)$$

►

Since an expression of the bias is available, it remains to determine the variance of $\widehat{\text{var}}(\overline{X})$. Working in a similar way as in Sec. III A, by first writing $\widehat{\sigma}_K^2$ as a quadratic form and then applying the result of Preliminaries 4, the following expression is obtained:

$$\text{var}\left(\widehat{\sigma}_K^2\right) = 2\left(\frac{\sigma_K^2}{n_K}\right)^2(n_K-1). \quad (39)$$

It is possible to estimate each quantity of the equation separately, obtaining the following estimator of the MSE of $\widehat{\text{var}}(\overline{X})$ given var($\overline{X}$):

$$\widehat{\text{MSE}} \overset{(38)}{=} [\widehat{\text{var}}(\overline{X})]^2\frac{2n_K-1}{n_K^2}. \quad (40)$$

It is now straightforward to implement these self-evaluation features in your estimation program. In the final part of this section, the announced upper bound on the computational complexity is derived.

An upper bound of the computational complexity of the method is $12n$. Consider the sample code in Fig. 6. The only contributions at order n are from the while loop. At iteration number i, the while loop can be computed using exactly $6n_i$ floating point operations. Using geometric series, the total floating point operations are

$$\text{cost} = \sum_{j=0}^{d-1} 6n_j = 6\times 2^d\sum_{j=0}^{d-1}2^{-j} = 12(n-1). \quad (41)$$

For time consuming computations that require multithread computing or time series so large that it comes in chunks, this bound can be reduced to $n + O(1)$ as will be shown in Sec. III D, but first consider these test results of the present implementation.

C. Test results

The method validation uses autoregressive models because Wold decomposition justifies their use in modeling stationary time series [6]. Moreover, var($\overline{X}$) can be be computed exactly for autoregressive models. This makes them ideal for our purpose.

Preliminaries 6. An *autoregressive model* of order p denoted AR(p) is a stochastic process $\{X_t\}_{t=1}^{\infty}$ such that

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

for $\phi_i \in \mathbb{R}$ and random variables ε_t that are independent, identically distributed with zero expected value and constant variance σ^2 for all t. The autoregressive models used have orders 1 and 2, they have autocovariance in closed form, and are stationary [6]. For all stationary processes, var($\overline{X}$) is given by the autocovariance γ, and for the AR(p)-processes of interest to us, γ is determined by the polynomial $P(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$. The autoregressive model is said to be *causal* if all the roots z_i of $P(z)$ satisfy $|z_i| > 1$. ►

Causal AR(1) and AR(2) process are easily parameterized to be asymptotic uncorrelated and stationary [6]. Two tests of the algorithm are presented. First, 6080 causal random AR(1) and AR(2) processes were generated. According to Preliminaries 6, this means the exact value of var($\overline{X}$) can be computed from the autoregressive coefficients ϕ_i for each of the AR(p) processes. The relative error squared, ϵ^2, converged to zero as a function of n/τ. Here, τ denotes the time constant of the autocorrelation function [τ is the smallest integer such that $|\gamma(h)| \leqslant \gamma(0)e^{-1}$ for all $h \geqslant \tau$]. Gamma regression is suitable because the observations of ϵ^2 are independent, identically gamma distributed, and the model is $\log(\epsilon^2) =$

TABLE II. Regression summary for the AR(p) processes: regression table for the AR(1) process (left) and AR(2) (right). If ϵ denotes expected relative error, the model was $\log(\epsilon^2) = \beta_0 + \beta_1 \log(n/\tau)$. The regression family is taken to be gamma, and fitted by maximum likelihood estimation using iterative reweighted least squares. The estimated values of β_j are given above along with standard errors. The p values are given for the null hypothesis that $\beta_j = 0$. Deviances explained are 50.65% and 65.39% for the AR(1) and AR(2) on 6078 degrees of freedom, respectively.

	AR(1)			AR(2)		
	Estimate	Std error	p value	Estimate	Std error	p value
β_0	0.7402	0.2592	0.00431	2.4566	0.0991	$<10^{-16}$
β_1	−0.5202	0.0271	$<10^{-16}$	−0.7022	0.0108	$<10^{-16}$

$\beta_0 + \beta_1 \log(n/\tau)$. The expected relative error squared is

$$\epsilon^2 = e^{\beta_0}\left(\frac{n}{\tau}\right)^{\beta_1}. \tag{42}$$

Maximum likelihood estimates of β_j and standard errors are given for the causal AR(1) and AR(2) processes in Table II. The table shows that if there are very little data available (say $n = \tau$), then the relative error $\epsilon = 0.7402^{1/2} = 0.861$ and $2.4566^{1/2} = 1.567$ for the AR(1) and AR(2) processes, respectively. It is also evident that the convergence rate of the AR(2) processes are faster than the AR(1) processes. Among the processes, the type of autocovariance was the main differentiator of the AR(p) processes. For $p = 1$, the autocorrelation is of the form $\gamma(h)/\sigma^2 = \phi^{-h}$, while in the case $p = 2$, there exist $z \in \mathbb{C}$ and $a, b, c \in \mathbb{R}$ such that $\gamma(h)/\sigma^2 = a|z|^{-h}\cos(hb + c)$. Furthermore, no effect is found of the sampling distribution ($p \geqslant 0.34$, t test), and therefore the difference in ϵ^2 between the AR(1) and AR(2) experiments is attributed to the $\gamma(h)$ according to Preliminaries 2. A plot of the regression analysis is given in Fig. 3 and for the regression summaries, see Table II.

Second, two standard textbook physics applications were studied. The variance of the mean energy was estimated using the Flyvbjerg and Petersen [7] blocking method and the automated blocking method. The estimates were compared with dependent bootstrapping using geometric simulation

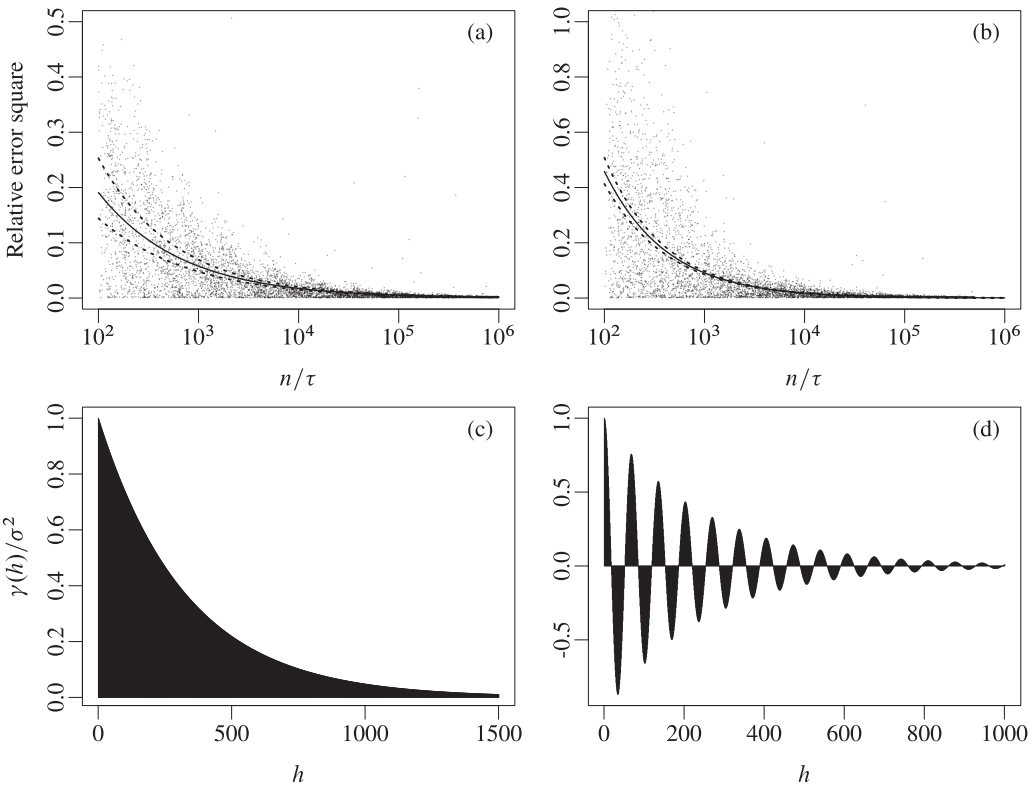

FIG. 3. Relative error squared of two autoregressive models versus observations per time autocorrelation time constant. There is exponential convergence rate for two common correlation structures in natural sciences. In the tail of the distribution, there are outlayers that failed the probabilistic test. The outlayers are not of major concern because (1) they are rare and (2) induced mistakes are small in magnitude. (a), (c) Represent test results for an AR(1) time series with an autocorrelation that is positive with exponential decay, typical of Metropolis-type Markov chains, where it is expected that the observations correlate positively. The process was distributed Gamma(1,1). (b), (d) Represent test results for an AR(2) process. The autocorrelation has exponential decay, but oscillates. The process was distributed multivariate standard normal. It was found that the method was insensitive to the distribution of the observations ($p \geqslant 0.34$). Consequently, the difference in behavior of the method is attributed to $\gamma(h)$, as explained by Corollary 1. The expected relative error squared ϵ^2 was modeled by gamma regression, $\log(\epsilon^2) = \beta_0 + \beta_1 \log(n/\tau)$. Deviance explained was 50.65% and 65.39% for AR(1) and AR(2) on 6078 degrees of freedom, respectively. Dashed lines give 95% confidence intervals of the expected relative error squared. The plots indicate that it is reasonable to expect the first digit of the method was correct for some $n \gtrsim 20\tau$, and two digits correct for some $n \gtrsim 25\,000\tau$.

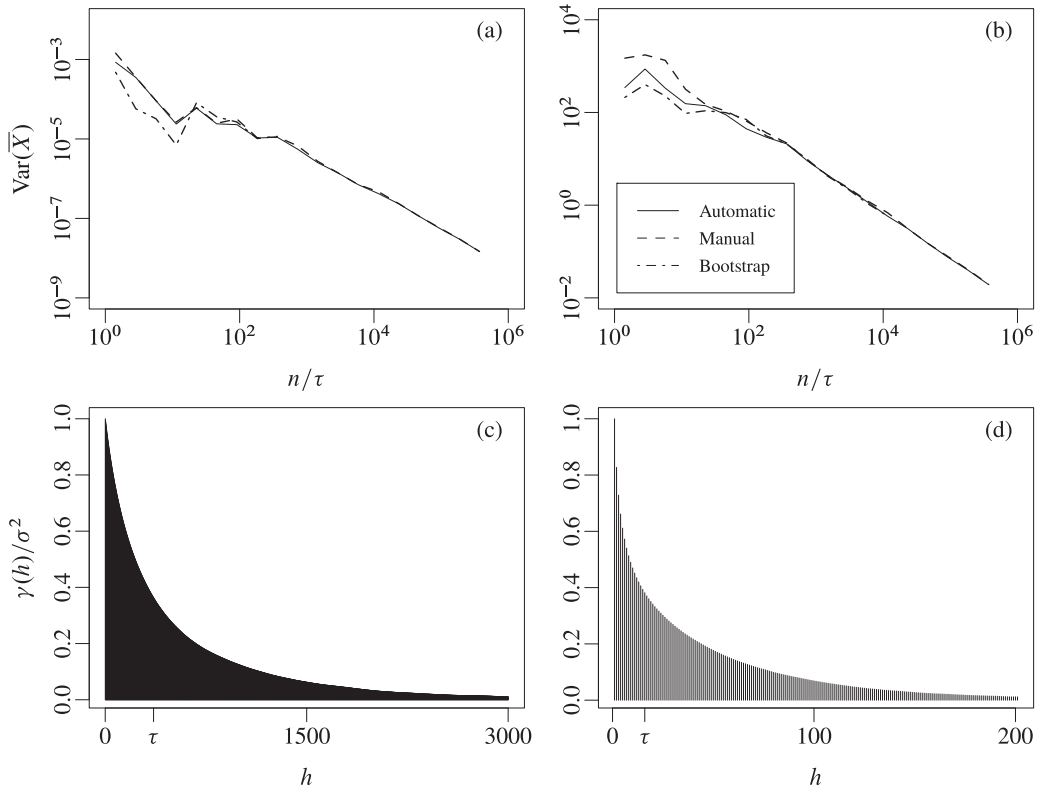

FIG. 4. Case study of two textbook physics applications for mean variance estimation. The variance of the mean energy was estimated using manual and automatic Blocking methods and compared with dependent bootstrapping using geometric simulation. The estimates are plotted in (a) and (b), while the autocorrelation is depicted in (c) and (d). (a), (c) Represent test results for a two electron quantum dot with trail energy of a Slater-Jastrow–type state function for a theory of a harmonic oscillator potential with Coulomb repulsion ($\omega = 1$). The importance sampling/Hastings-Metropolis theorem implementation implies acceptance rate >99.999. The autocorrelation time constant was $\tau = 360$. It is clear that $\mathrm{var}(\overline{X}) = 1.46 \times 10^{-8}$ at 10^8 samples according to all three models, and by the above discussion, the error of the variance estimate is correct to the second digit. (b), (d) Represent test results for the Ising model for a 20×20 grid of spins with periodic boundary conditions at temperature $T = 2.4$. The energy was sampled from a Boltzman distribution at significance level 95% according to a χ^2 test using a Metropolis-type Markov chain. The autocorrelation time constant was measured at $\tau = 16$, so at 10^6 samples $\mathrm{var}(\overline{X}) = 1.49 \times 10^{-1}$ according to all three methods, and there is reason to believe that the first two digits are correct.

type [18]. In either application, the autocorrelation functions bore resemblance of the AR(1) autocorrelation (in light of Corollary 1). The first application was an n-electron quantum dot with trail energy of a Slater-Jastrow–type state function for a theory of a harmonic oscillator potential with Coulomb repulsion. The angular frequency was $\omega = 1$. Importance sampling/Hastings-Metropolis theorem was used together with a Fokker-Plank–type prior [19]. The implementation has acceptance rate >99.999% for each proposed state. The autocorrelation time constant was $\tau = 360$, and the time until the observations were close to uncorrelated was $h \approx 4\tau$. The second application was an implementation of the Ising model using a 20×20 grid of spins with periodic boundary conditions at temperature $T = 2.4$ [20]. The energy was sampled from a Boltzman distribution at significance level $\alpha = 0.05$ according to a χ^2 goodness-of-fit test [3] using a stationary, time-reversible Markov chain constructed using the Metropolis algorithm [21]. The autocorrelation time constant was measured to be $\tau = 16$. Figure 4 plots the results.

D. Multithread computing and memory limitations

If the time series is sufficiently large, it is common to store the time series in smaller chunks, rather than in one file, or in memory all at once. Such can happen if the computing facility memory is smaller than the time series, or the application generating the time series runs on multiple threads. This is typically the case when the time series is generated by a Markov chain on multithread clusters. As shown above, it is possible to reduce the size of the data by applying blocking transformations on each chunk until the chunks are small enough to be imported onto a single node or personal computer. Suppose the amount of memory that can be allocated on this node is 2^d real numbers.

Assume now that the total length of the time series is $n = 2^D$, divided into 2^k smaller chunks of length 2^{D-k}. Let $\mathbf{X}$ denote any such vector containing a chunk of the time series. It is important that within such a chunk, the order of the observations is preserved. Now on the chunk, apply blocking transformations $D - d$ times and form $\mathbf{X}_{D-d} = T_{D-d}T_{D-d-1}\ldots T_1\mathbf{X}$ where T_i is defined in Eq. (22). The size of $\mathbf{X}_{D-d}$ is exactly $2^{D-k}/2^{D-d} = 2^{d-k}$. The same procedure is executed on each of the 2^k chunks, and thus the total data are of all the transformed chunks is $2^k 2^{d-k} = 2^d$, as required. On the parent node, or personal computer doing the final estimate, write the data to memory by concatenating

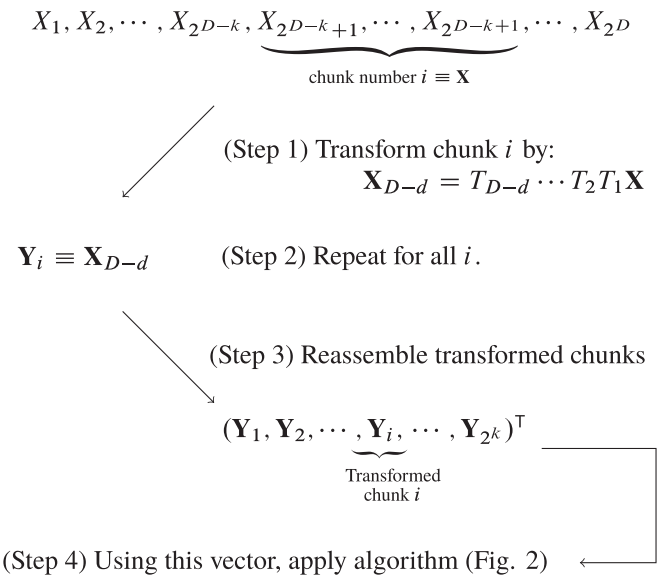

FIG. 5. In the case that the time series is too large for memory, or is so large that it is not saved in a single file, it is possible to reduce the size of each chunk of the time series by repeated use of the matrices T_j. This is convenient in the case that the time series is generating by a multithreaded program. The procedure is as follows: Choose one of the chunks of the time series, number i. (Step 1) Transform chunk number i by applying the matrix $T_{D-d}T_{D-d-1}\ldots T_2T_1$. Define Y_i to be the result of this transformation. (Step 2) Repeat for all $j \neq i$. (Step 3) Concatenate all the chunks after they have been formed into one long vector $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{2^k})^{\mathsf{T}}$. This vector will now have size 2^d, thus is small enough to (Step 4) be handled on a single node by using the algorithm of Fig. 2. See Sec. III D for more details including the definitions of the numbers D, d, and k.

each of the 2^{d-k} blocks end to end into a long vector of size 2^d, then perform the ordinary algorithm as it is given in Fig. 2. Figure 5 provides an overview of the procedure.

The computational cost of this is low. Performing the transformations T_i as it is done in the code of Fig. 6 requires precisely n_{i-1} floating point operations, as you can check. So, the total number of floating point operations is computed using geometric series:

$$\sum_{j=1}^{D-d} n_{j-1} = \sum_{j=0}^{D-d-1} 2^{D-k-j} = \frac{2^D - 2^d}{2^{k-1}}. \tag{43}$$

According to Eq. (41), it is necessary to add the $12(2^d - 1)$ floating point operations that the parent node must spend at the end, so the total cost is bounded above by

$$\text{cost} \overset{(43)}{\leqslant} \underbrace{2^D}_{=n} \frac{1}{2^{k-1}} + 2^d \left(12 - \underbrace{\frac{1}{2^{k-1}}}_{\leqslant 1}\right)$$
$$\leqslant n + O(1) \quad \text{as } n \to \infty.$$

```
# data vector must be of size 2^d for some integer d
X = loadtxt("data.txt")
n = len(X); d = log2(n); mu = mean(X); i = 0
s, gamma = zeros(d), zeros(d)
# Chi-square percentiles. More values in appendix
q = array([6.634897, ... , 50.892181])
# Get autocovariance and variance for all X_i
while n >= 2:
   # estimate variance and autocovariance of X_i
   x = X - mu
   gamma[i] = sum(x[1:n]*x[0:(n-1)])/n
   s[i] = sum(x**2)/n
   # perform blocking transformation
   y = zeros(n/2)
   for j in arange(0, n/2):
      y[j] = 0.5*( X[2*j] + X[2*j+1] )
   X = y; n = n/2; i = i + 1
# Generate the test statistic M_j
M = zeros(d)
for j in arange(0,d):
   n = 2**(d-j)
   M[j] = n*((n-1)*s[j]/n**2 + gamma[j])**2/s[j]**2
# elements reversed twice such cumsum is correct
M = cumsum(M[::-1])[::-1]
# Determine the smallest k such that H_0 is true
for k in arange(0,d):
   if(M[k] < q[k]):
      break
if(k >= d - 1):
       print "Warning: Add more data."
# and the answer is
nk = 2**(d-k)
answer = s[k]/nk
```

FIG. 6. PYTHON implementation of the algorithm. The code is purposefully verbose to aid implementation in languages of lower level of abstraction, such as C. In practice, the implementation can be optimized and shrunk to about 10–15 lines of code. The most recent implementations for PYTHON, C++, and R are available (see [8]).

The reason it is possible to rejoin the time series by putting it end to end is the same reason dependent bootstrapping works: As long as the chunks are large enough, the resampling of putting the observations end to end does not change γ. See for example [18,22]

Analogously, the total mean can also be computed in chunks since it splits up into a mean of means. Define mean of chunk number j by $\widehat{\mu}_j = (1/2^{D-k}) \sum_{i=(j-1)2^{D-k}}^{j2^{D-k}} X_i$, then write

$$\overline{X} = \frac{1}{2^D} \sum_{i=1}^{2^D} X_i = \sum_{j=1}^{2^k} \frac{1}{2^D} \underbrace{\sum_{i=(j-1)2^{D-k}}^{j2^{D-k}} X_i}_{2^{D-k}\widehat{\mu}_j} = \frac{1}{2^k} \sum_{j=1}^{k} \widehat{\mu}_j. \tag{44}$$

That implies the total mean of the time series is just the mean of all the means. All in all, there is no problem in splitting the whole time series in chunks since both statistics of interest are recovered at the end.

If the data are generated by a program, it is a time saver to compute the estimates on each thread at the same time the

program is generating data. If you choose to do so, precision is maximized by making the chunks as large as possible. Working in this way saves considerable time because then the data do not have to be read back into memory for post-processing later.

IV. DISCUSSION

This study provides the following four new contributions: (a) rigorous proof of the Flyvbjerg and Petersen [7] blocking method conforming to the standard of modern mathematics. The results give prospects for new research with relevance to any blocking method. (b) An automated blocking method is provided. It works for a variety of autocovariance functions. (c) Autoregressive models were chosen to provide error estimates. These account for error due to both (i) the method and (ii) the sampling. (d) Integration of the blocking method for multithread computing or extremely large time series. The contributions include proof of the behavior of the method. Furthermore, Proposition 1 outlines an approach to (i) estimate the standard error more efficiently and (ii) provide economical error estimates, both in terms of computation simplicity and in precision. This method is simple to explain and implement (requires no more than 10–20 lines of code), and will appeal to those using the Flyvbjerg and Petersen [7] blocking method, because it works under more general conditions and maintains the simplicity of the original method.

Several authors have attempted to give justification for the use of the blocking methods. Best known is the work of Flyvbjerg and Petersen [7] providing motivational mathematics to explain the idea of Blocking transformations for standard error estimation. They claim that there exists "an obvious fixed point" but gives no proof [7]. For mathematically interested readers, this can present a distraction since fixed points of any function $T_j : A \to B$ are defined when $A = B$ [4,23,24]. In this context, $A \neq B$ since $A = \mathbb{R}^{n_i}$ and $B = \mathbb{R}^{n_i/2}$ (see Sec. I). Thus, in mathematical sense, there is no fixed point present. Instead, the justification given in the results is the following: For the blocking method, the variables subject to k blocking transformations $\mathbf{X}_k$ form a stationary time series if $\mathbf{X}$ is stationary. This means that it is possible to express the error e_k as a function of $\gamma_k(h)$. From this and the transformation properties of the $\gamma_j(h)$, it follows that the behavior of the truncation error is given by $\{\gamma_k(1)\}$. See Proposition 1. This may come as a surprise because this implies that the behavior of the method is determined by the set of $\{\gamma_k(1)\}_{k=0}^{d-1}$. Flyvbjerg and Petersen [7] appear unaware of this because they state that the blocking method converges if $\gamma(h) \propto 1/h$ [7], which does not capture the essence, as was proved in Theorem 1. In fact, their blocking method works whenever $X_1, X_2, \ldots$ is asymptotic uncorrelated, as Theorem 1 makes precise.

Proposition 1 proves that the blocking method is applicable under more general conditions than assumed by Flyvbjerg and Petersen [7]. First, because $\gamma(h)$ need not be proportional to $1/h$ [note Proposition 1, which places no restriction on $\gamma(h)$, although finite variance is required]. Therefore, $\gamma(h)$ can have any shape. Second, Flyvbjerg and Petersen [7] constrain $\gamma(h)$, exactly n degrees of freedom (since γ is a function $\gamma : \{0, 1, \ldots, n-1\} \to \mathbb{R}$), while the results only constrain $\{\gamma_k(1)\}$, exactly $\log_2(n) = d$ degrees of freedom. Theorem 1 may also have theoretical interest in statistical mathematics. The sum $s_k = \sum_{h=1}^{n_k-1}(1 - h/n_k)\gamma_k(h)$, for $k = 0$, appears frequently in the study of time series [6], Proposition 1 and Lemma 1 together imply that $(2/n_k)s_k$ is determined up to a constant[1] by the set $\{\gamma_k(1)\}$. Moreover, according to Theorem 1, $\gamma_k(h)$ converges uniformly to zero on $\mathbb{N}$. In this way, blocking transformations are intimately linked to s_k. This provides interesting prospects for further work: for physics, it is possible to provide realistic error estimates and improve the method substantially if $\{\gamma_k(1)\}$ is estimated more accurately [since $e_k = (2/n_k)s_k$]. However, elegant solutions probably require non-Fisherian statistics. Perhaps Bayesian statistics can be used because estimation is difficult for large k simply because when k is large, then the sample size n_k available for estimation is small. For example, a suitable shrinkage estimator may be particularly useful. See, for example, Stein's phenomenon [25] and applications such as those by Schäfer and Strimmer [26]. Another proposed application of Proposition 1 is the proof of Corollary 1, which explains the behavior of the blocking method [7], and why the automatic blocking method works. The corollary shows that the estimates $\mathrm{var}(\overline{X})$ improve with each blocking transformation, until (i) the variables $\mathbf{X}_k$ become uncorrelated or (ii) there exist j such that the covariances $\gamma_k(1) = 0$ for all $k \geqslant j$. In case (i) the truncation error $e_k = 0$ since if the components of $\mathbf{X}_k$ are uncorrelated, then $\mathrm{cov}((\mathbf{X}_k)_i, (\mathbf{X}_k)_j) = \sigma_k^2\delta_{ij}$ by definition, so $\gamma_k(h) = 0$ for $h \geqslant 1$ and, hence,

$$e_k = \frac{2}{n_k}\sum_{h=1}^{n_k-1}\left(1 - \frac{h}{n_k}\right)\underbrace{\gamma_k(h)}_{=0} = 0.$$

Proposition 2 strengthens this statement to include case (ii) because it shows that γ_k converges uniformly and identically to zero on $\mathbb{N}$ whenever $\gamma(h) \to 0$ as $h \to \infty$.

There are several methods that automates computation of $\mathrm{var}(\overline{X})$. In physics, the most well known automated method for standard error estimation is perhaps dependent bootstrapping [18]. Dependent bootstrapping is useful when n is small or if n is large and the required precision is small. The main advantage of bootstrapping methods over the present method is their flexibility. Since bootstrap methods are popular, much is known about their applications, for example, Parr [27] has shown that Frechét differentiability is a sufficient condition that the independent bootstrap works. According to Politis and Romano [18], dependent bootstrapping has asymptotically valid procedures even for multivariate parameter spaces. However, high precision estimates for large data sets are often needed. Flyvbjerg and Petersen [7] proposed an alternative method to automate computation. They proposed an automation by providing a confidence interval to test for normality of $\mathbf{X}_k$ using $\widehat{\sigma}^2$. Typically, this method works well since $\gamma(h) \propto 1/h$ is commonly used in physics. However, this is not always

[1]That constant is σ_0^2, as can be proved by iteratively using Proposition 1.

the case, and for automations operating without supervision, it is possible to provide improvements. For example, stability of the method depends on the shape of the covariance $\gamma(h)$. This can fail for certain types of correlation structures, for example, oscillatory AR(2)-like processes introduced here. The automation works for causal AR(p) processes for any p, and places no assumption on the shape of $\gamma(h)$. In addition, this paper provides updates that make it convenient to use the method for multithread computing. Another alternative method is the Gamma method proposed by Wolff [28]. The Gamma method works well with correct setup, with errors claimed to be lower than those of the automatic blocking method proposed here. Wolff [28] claims that the Gamma method works for other types of correlation structures than exponential decaying types. But, it may be necessary to set up the method's integration window manually. Wolff provides suitable tools for the purpose, and explains that it seems impossible to design automatic windowing that is adequate in all possible cases. As such, it is possible to introduce a fully automated method. In contrast to Wolff [28] recommendation, Lee *et al.* [29] are proponents of a method that Wolff has called binning (which is essentially a blocking method). Lee *et al.* proposed inequalities that can be used to automate calculations. However, this approach requires estimates that may or may not be available. The automated blocking method has none of the complications mentioned above.

The proposed automation uses exactly two approximations; namely, that (1) Σ_j is diagonal and (2) calculated to leading order in $1/n_k$. In practice, approximation (2) can be avoided because an exact expression for the covariance matrix is provided by Proposition 5. However, the method will often fail when approximation (1) is dropped. Consequently, I recommend keeping it as a minimum. The problem is that inversion of Σ_j can become difficult because the condition number often become very large as n increases. However, using this approximation is not a problem because Proposition 6 guarantees that the expected error of approximation (1) is precisely zero. And, if n_k is large, then (2) is not a problem because then $(1/n_k)^j \ll (1/n_k)^i$ for all $i < j$. The bottom line is that M_j must be computed using a diagonal covariance matrix $\mathrm{diag}(s_j, s_{j+1}, \ldots, s_{d-1})$ but it is up to the end user whether he/she sets $s_i = \mathrm{var}(\widehat{\gamma}_i(1))$ using Corollary 3 or expanding to leading in order by setting $s_i = \sigma_i^4/n_i$ for all $j \leqslant i \leqslant d-1$. In this paper I opted for $s_i = \sigma_i^4/n_i$ because the tests I have performed showed found no effect of higher precision ($p = 0.72$, t test).

Using AR(p) process for method validation is natural in this context because it is possible to quantify both the error due to (i) the method and (ii) due to the sampling of the data. It would have been impossible to encompass the error due to sampling if the estimates had been compared to high precision estimates from another industry standard method. The error estimates are empirical rather than analytical, but one drawback is that it is only possible to validate the method on a finite number of problems. AR(1) and AR(2) processes were chosen because Wold decomposition says that the random component of any time series can be expressed as an autoregressive model [6]. AR(1) and AR(2) correlation functions are the two most common ones encountered in modeling of time series. The two textbook cases, quantum dot and the Ising model, show that their correlation structures were similar to the AR(1) processes. Using Eq. (42), the results show that the accuracy is as follows: With almost no data available, end users can expect that the estimates are of correct order of magnitude (since $\epsilon^2 = e^{\beta_0}$ if $n = \tau$). The expected accuracy increases to produce the first correct digit already at circa $n = 10^4$ and 10^5 observations for the Ising model and quantum dot, respectively. While it is expected that the second digit is also correct if n circa 10^6 and 10^8 for the Ising model and quantum dot, respectively. This means that the convergence of the relative error to zero is slower than the claimed value for the Gamma method [28]. However, unlike the estimates due to Wolff [28], Table II gives regression results, thus providing a measure on all sources of error (even the errors made by the end users in sampling of the data). In practice, the physics application shows that the estimates are similar to those of dependent bootstrapping and the Flyvbjerg and Petersen [7] blocking method, regardless of n (see Fig. 4), is fully automatic and works in $O(n)$ time.

V. CONCLUSION AND PERSPECTIVES

A rigorous proof of the blocking method (Flyvbjerg and Petersen [7]) is a main result of this study. That method has become one of the industry standards for estimating standard errors $\mathrm{var}(\overline{X})^{1/2}$ of the mean whenever the number of observations is large. Second, the proof gives an automated implementation that eliminates the need for human intervention. The method uses Fisherian inference to propose a hypothesis test that can be used to determine the estimate of the standard error. The method has complexity $O(n)$, and works for all common covariance structures in natural sciences. This should first and foremost appeal to researchers in computational physics, but also in other sciences, since the study conforms to the standard rigor of modern mathematics and introduces terminology standard in the other sciences. By being automated and complexity $O(n)$, the present method is less expensive than other methods for standard error estimation of the mean used in computational physics (source code is available from [8]).

The paper proposes prospects for more research. Proposition 1 shows that the behavior of any blocking method is determined by the set $\{\gamma_k(1)\}_{k=0}^{d-1}$. However, more advanced estimation is needed to use the result for efficient estimation of $\mathrm{var}(\overline{X})$. The problem is that for large k, the data available to estimate $\gamma_k(1)$ are small and, consequently, any classical Fisherian estimation is inappropriate. Accordingly, shrinkage estimation or Bayesian estimation may be used. The result is interesting for applications because the truncation error of blocking methods can be expressed in terms of $\{\gamma_k(1)\}$. Therefore, professional error bounds may be provided by developing the mathematics further. Or, better yet, it may be possible to estimate the errors, which would provide significant benefits to end users. Furthermore, it is probably possible to relax the requirements of Theorem 2 because work is constantly being done on central limit theorems. Finally, it would be useful to classify all the Markov chain Monte Carlo methods that are common in computational physics (see for

example [30]), such that it is more clear for which methods Theorem 2 continues to hold.

ACKNOWLEDGMENTS

The author is indebted to Professor M. Hjorth-Jensen (Department of Physics, University of Oslo) and Associate Professor H. Flyvbjerg (Department of Micro- and Nanotechnology, Technical University of Denmark) for valuable support during the development of the results. Thanks go to Professor Ø. Borgan and Professor A. Rygh Swensen (Department of Mathematics, University of Oslo) for helpful comments to drafts of the manuscript. Professor G. L. Jones (School of Statistics, University of Minnesota) and Professor R. C. Bradley (Department of Mathematics, Indiana University) had technical comments that were helpful and the author is grateful to K. Ravn (University of Oslo) for contributing essential pattern-finding skills for Proposition 5. Finally, I am grateful to two anonymous reviewers for their helpful comments and suggestions.

APPENDIX

Lemma 6. The sequence $\{f_k\}$ satisfies the following properties:

(1) $f_k(i) \leqslant i$ for all $1 \leqslant i \leqslant 2^{k+1} - 1$.
(2) $\sum_{i=1}^{2^{k+1}-1} f_k(i) = 2^{2k}$.
(3) $f_{k+1}(i) = f_k(i) + 2f_k(i - 2^k) + f_k(i - 2^{k+1})$.

Proof. The first property is obvious. For the second property, use the arithmetic series formula. Write

$$\sum_{i=1}^{2^{k+1}-1} f_k(i) = 1 + 2 + \cdots + 2^k + (2^k - 1) + \cdots + 2 + 1$$

$$= 2^k + 2\sum_{i=1}^{2^k-1} i = 2^k + 2\frac{2^k - 1}{2}2^k = 2^{2k}.$$

For the third property, use induction. The base case is satisfied for $k = 0$, then,

$$1 = f_1(1) = f_0(1) + 2f_0(1 - 1) + f_0(1 - 2) = 1 + 0 + 0,$$

$$2 = f_1(2) = f_0(2) + 2f_1(2 - 1) + f_0(2 - 2) = 0 + 2 + 0,$$

$$1 = f_1(3) = f_0(3) + 2f_2(3 - 1) + f_0(3 - 2) = 0 + 0 + 1.$$

For the induction step, suppose hypothesis is true for k. If $0 \leqslant i \leqslant 2^{k+1}$, then $f_{k+1} = i$. Moreover, either $0 \leqslant i \leqslant 2^k$ or $2^k \leqslant i \leqslant 2^{k+1}$. If the former is true, then

$$f_k(i) + 2f_k(i - 2^k) + f_k(i - 2^{k+1}) = i + 2 \times 0 + 0 = i.$$

If the latter is true, then

$$f_k(i) + 2f_k(i - 2^k) + f_k(i - 2^{k+1})$$
$$= 2^{k+1} - i + 2(i - 2^k) + 0 = i.$$

The other cases are proved similarly. ■

[1] K. F. Riley and M. P. Hobson, *Essential Mathematical Methods for the Physical Sciences* (Cambridge University press, Cambridge, 2011).
[2] The Editors, Phys. Rev. A **83**, 040001 (2011).
[3] J. L. Devore and K. L. Berk, *Modern Mathematical Statistics with Applications*, 2nd ed. (Springer, London, 2012).
[4] J. N. McDonald and N. A. Weiss, *A Course in Real Analysis*, 2nd ed. (Academic, Oxford, 2013).
[5] M. H. DeGroot and M. J. Schervish, *Probability and Statistics*, 4th ed. (Pearson Education, Harlow, 2014).
[6] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications with R Examples*, 4th ed. (Springer, Cham, 2017).
[7] H. Flyvbjerg and H. Petersen, J. Chem. Phys. **91**, 461 (1989).
[8] PYTHON, R, and C++ code is available: www.github.com/computative/block.
[9] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 3rd ed. (Springer, Basel, 2016).
[10] A. Agresti, *Foundations of Linear and Generalized Linear Models*, 1st ed. (Wiley, Hoboken NJ, 2015).
[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. (CRC Press, Boca Raton, FL, 2014).
[12] I. A. Ibragimov, Teor. Veroyatnost. i Primenen. **20**, 135 (1975).
[13] R. C. Bradley, Rocky Mountain J. Math. **17**, 95 (1987).
[14] B. Efron, Am. Stat. **40**, 1 (1986).
[15] A. Mathai and S. B. Provost, *Quadratic Forms in Random Variables* (Marcel Dekker, New York, 1992).
[16] D. C. Lay, *Linear Algebra and Its Applications*, 4th ed. (Addison-Wesley, Boston, 2012).
[17] M. Hazewinkel, *Encyclopedia of Mathematics* (Kluwer, Dordrecht, 1993).
[18] D. N. Politis and J. P. Romano, J. Am. Stat. Assoc. **89**, 1303 (1994).
[19] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, 2nd ed. (Springer, Berlin, 1985).
[20] M. Plischke and B. Bergersen, *Equilibrium Statistical Physics*, 3rd ed. (World Scientific, New Jersey, 2006).
[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[22] D. N. Politis and H. White, Econom. Rev. **23**, 53 (2006).
[23] M. Hazewinkel, *Encyclopedia of Mathematics* (Kluwer, Dordrecht, 1989).
[24] E. J. Borowski and J. M. Borwein, *Dictionary of Mathematics* (Collins, London, 1989).
[25] C. Stein, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Contributions to the Theory of Statistics*, Vol. 1 (Univ. of Calif. Press, 1956), pp. 197–206.
[26] J. Schäfer and K. Strimmer, Stat. Appl. Genet. Mol. Biol. **4**, 1 (2005).
[27] W. C. Parr, Stat. Probab. Lett. **3**, 97 (1985).
[28] U. Wolff, Comput. Phys. Commun. **156**, 143 (2004).
[29] R. M. Lee, G. J. Conduit, N. Nemec, P. López Ríos, and N. D. Drummond, Phys. Rev. E. **83**, 066706 (2011).
[30] G. L. Jones, Probab. Surveys **1**, 299 (2004).