

STANDARD ERROR ESTIMATION BY AN AUTOMATED BLOCKING METHOD

by

Marius Jonsson

THESIS

for the degree of

MASTER OF SCIENCE



Faculty of Mathematics and Natural Sciences
University of Oslo

June 2018

Abstract

Background: The sample mean $\bar{X} = \sum_{i=1}^n X_i$ is probably the most popular estimator of the expected value in all sciences, the standard error is given by the square root of the variance $V(\bar{X})$.

Purpose: This work aims at providing a stringent and modern treatment of the Blocking method (sensu Flyvbjerg & Petersen 1989), a popular way to compute $V(\bar{X})$ for correlated data.

Methods: Linear algebra, results from multivariate probability theory, real analysis and Fisherian statistical inference were used. The method validation was performed by Metropolis-Hasting-type sampling and autoregressive models.

Results: Here a new approach to estimate $V(\bar{X})$ for time series data is presented. The method applies to stationary data, such as Markov chains and other stationary time series. The complexity of the method is bounded above by $12n + O(\log_2 n)$ floating point operations, but this can be reduced to $n + O(1)$ in large computations. The convergence in relative error squared is better than $\propto n^{-1/2}$. The method is insensitive to the probability distribution of the observations. It is proven that only a small part of the correlation structure is relevant to the convergence rate of the method. From this, the 1989 Blocking method follows as a corollary. The result is also used to propose a hypothesis test to survey the relevant part of the correlation structure. The method is sufficiently robust to operate without supervision. An algorithm and sample code showing the implementation is available for Python, C++ and R. This code is available for download at github.com/computative/block. Method validation using autoregressive AR(1)- and AR(2)-processes and physics applications are included. This method has an accuracy similar to dependent bootstrapping, but scales in $O(n)$ -time. The method is easily adapted to multithread applications and time series larger than computing cluster memory.

Conclusions: By applying stringent mathematics, the Blocking method was automated, which will be helpful for all users of this method for computing standard errors of means generated from correlated data series.

Acknowledgements

The author is indebted to Professors Anders Rygh Swensen and Ørnulf Borgan (Department of mathematics, University of Oslo) for checking that the mathematics results are correct and for reviewing the attached manuscript. The author is also indebted to Professor Morten Hjorth-Jensen (Department of physics, University of Oslo). He introduced me to the blocking method, encouraged me to pursue the mathematics and has given invaluable support during the past two years. In particular, I am grateful for his checking that the physics content of the thesis is correct. Associate professor Henrik Flyvbjerg (Department of Micro- and Nanotechnology, Technical University of Denmark) helped by making it clear in which areas I could direct attention to attain the most useful results. Professor Galin Jones (School of Statistics, University of Minnesota) and Professor Richard Bradley (Department of mathematics, Indiana University) helped to get a special type of central limit theorem in place, making the results substantially slicker. Furthermore, I am grateful to Kenneth Ravn (Department of mathematics, University of Oslo) for countable favors. In particular for essential pattern-finding skills for proposition 8. Last, but not least, I am thankful to my father, Professor Bror Jonsson (Norwegian Institute for Nature Research) for much help and many suggestions.

Contents

1	Introduction	1
2	Method	7
2.1	Real analysis results	7
2.1.1	Functions on metric spaces	7
2.1.2	Banach spaces and the Fréchet derivative	10
2.2	Probability theory	12
2.2.1	Univariate theory	12
2.2.2	Multivariate probability theory	17
2.2.3	Preliminary probability theory	22
2.3	Applications of probability theory	28
2.3.1	Fisherian inference	28
2.3.2	Bayesian statistics	37
2.3.3	Shrinkage estimation	42
2.4	Time series	44
2.4.1	Stationarity	44
2.4.2	Markov chain Monte Carlo theory	46
2.5	Numerical estimation	51
2.5.1	Resampling methods	51
2.5.2	Manual blocking method	61
2.6	Physical models	66
2.6.1	Ising model and Metropolis algorithm	66
2.6.2	Two dimensional N -electron quantum dot and Metropolis-Hasting algorithm	70
3	Results: Deriving an automatic blocking method	75
3.1	The plan and preliminaries	75
3.2	Proof of the blocking method	76
3.3	Automating calculations	81
3.4	Covariance matrix of $\hat{\gamma}_i(\mathbf{1})$ and the matrix Σ_j	82
3.5	Algorithm	88
3.6	Test results	90
3.7	Multithread computing and memory limitations	93

Contents

4	Discussion	97
5	Conclusion and perspectives	101
	Appendix A	103
	Cited literature	107
	Index of notation	111
	Index	113

Chapter 1

Introduction

Reliable estimations of the variance of sample means

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

are essential in natural sciences (Riley and Hobson 2013). Here n denotes the number of random variables or observations X_i . This is because \bar{X} is a typical estimator of the expected value μ of X_i if the elements of $\{X_i\}_{i=1}^n$ are identically distributed. The variance of the mean is the expected squared error of the estimate. Already in 1867, Chebyshev (Devore and Berk 2012) explained this by showing that if the observations have variance σ^2 , and \bar{X} has finite non-zero variance $V(\bar{X})$, then for any real number $k > 0$

$$P(|\bar{X} - \mu| > k[V(\bar{X})]^{1/2}) \leq \frac{1}{k^2}. \quad (\text{Chebyshev's inequality}) \quad (1.1)$$

Throughout, $V(X)$ denotes the variance of a random variable X , and P is a probability measure. If the observations are independent and identically distributed, the variance of the mean is easily obtained by setting $V(\bar{X}) = \sigma^2/n$ (DeGroot and Schervish 2012), but for correlated data, the computation is more complicated (Shumway and Stoffer 2017). Here, however, I show that if there is some integer $d > 1$ such that $n = 2^d$, and X_1, \dots, X_n are observations from a stationary time series, then the complexity is essentially the same as that of the sample mean, and one can use an automated scheme to compute it.

In general, induced correlations in data can be complicated, and it can be difficult to deal with correlation correctly, let alone using this information to find realistic estimates of the standard error of \bar{X} . But, if the data are a time series, as is often the case in physics, it is possible to do this using the autocovariance function γ , and at the same time obtain realistic estimates of $V(\bar{X})$. To physicists, there are at least three methods that are popular. These are (1) *the Jackknife*,

(2) various types of *bootstrapping* and (3) by *blocking transformations* (Flyvbjerg and Petersen 1989). The first two have sound mathematical foundation built by mathematicians, but number (3) has not.

The use of blocking transformations refers to forming a new sample of data by taking the mean of every pair of subsequent observations. The ***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

$$B_i = \frac{1}{2} (A_{2i-1} + A_{2i}). \quad (1.2)$$

Such transformations are applied in many areas of probability theory, and Flyvbjerg and Petersen (1989) made popular a method where blocking transformations reduced the correlations of the data, and proposed a way to estimate the variance of the mean. However, Flyvbjerg and Pedersen (1989) leave much to be wanted in terms of mathematical rigor. These authors also stated that an automation of the blocking method would be useful. They claimed (without proof) to have an automation that works in cases where the autocovariance is $\gamma(h) \propto h^{-1}$. Here, h denotes observation lag. The type of automation in question will fail if γ is not injective¹ on \mathbb{N} , this is the case whenever γ is not strictly monotonous. I communicated with Flyvbjerg in the summer of 2017, and it became clear that it would be useful if solutions to these two problems were produced. This is precisely the aim of this thesis.

In the present study, the mathematics is developed and the automation, which is robust enough to operate without supervision, is given. Moreover the conditions of γ have become significantly milder. In their paper, Flyvbjerg and Petersen (1989) claim that the method works whenever $\gamma(h) \propto h^{-1}$, the conditions are now nearly as mild as they can be, namely that $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$. Requirements for the method to work has been planned in such a way that they are convenient to check for physicists with moderate understanding of probability theory. Thus, the method was developed with the end user in mind. But also current users of the blocking method were kept in mind: I recycle much of the philosophy of Flyvbjerg and Petersen (1989), although the mathematics is new. Despite this, no compromise is made on performance. The method scales as $12n + O(\log_2 n)$ in small calculations, but this is reduced to $n + O(1)$ in large calculations. The relevance of the method is clear by making it possible to speed up standard error calculations of the mean by factors of thousands or more, as compared to dependent bootstrapping or other methods of complexity $O(n^2)$ or $O(n \log n)$. I also present the algorithm and give sample code². The method validation is

¹See section 2.1 for definition of injective function

²Sample code for python, C++ and R is available from github at github.com/computative/block

empirical by use of examples that are natural test candidates. First by toy examples using autoregressive series because for such series, the variance $V(\bar{X})$ is known precisely and needs not to be estimated. This was useful to characterize the method performance. But there are also physics applications for users less familiar with linear processes. All of the above requires more mathematics than what is standard in undergraduate studies in physics. This brings me to the introduction of the second major goal of the thesis.

My supervisor, Professor Morten Hjorth-Jensen proposed that I write a thesis to serve as an introduction to probability theory with applications for physicists. The aim has been to adapt the material to suit master students of physics. That means substantial emphasis is placed on building a useful mathematical foundation for students to understand the results, often assuming acquaintance with concepts that are standard in physics education. This is useful for other reasons too: With the advent of machine learning in computational sciences, statistical methods are arguably becoming important to a larger audience of physicists according to Denby (2004). Thus, it is natural to present the results of this thesis in two ways. (1) As this submitted thesis, and (2) a shorter manuscript submitted for publication in a scientific, peer reviewed journal, which was submitted for publication on May 15, 2018. The latter is attached as appendix to the thesis, and can be readily understood by professionals already proficient in probability theory and applied probability theory. Yet, I present the theory in the methods chapter short for economic reasons. Thus, the method chapter introduces (a) a minimum required to understand the results, but at the same time, (b) will occasionally introduce some extra material. The latter puts important concepts in perspective, giving readers the broadest introduction at a low (theoretical) cost. My aim is to give students an optimal theoretical return, with the least time spent.

- Real analysis. A minimum amount of real analysis required to understand the results are given. In addition, a few basic concepts are introduced to make the discussion more concise. Third, some definitions useful in normed spaces are given, because this makes it easier to state sufficient requirements for consistency of the bootstrap estimator when resampling methods are discussed in section 2.5.1.
- Elementary probability theory is necessary. The thesis deals with time series data, i.e. data that have a certain structure called dependence. That makes it necessary to discuss multivariate probability theory, dependence and measure of dependence, because they are the tools to inscribe dependence in models, and probe that it is present in the data. It was also necessary to introduce the probability distributions that appear in the results.

- Statistical inference is necessary to analyze the results quantitatively, but here also presented to exhibit serious data analysis for students. Statistical inference is a key ingredient in the automation of the blocking method. In the results, Fisherian inference takes center stage, yet I allow for 5 pages of Bayesian statistics. The latter is for at least two reasons: Primarily, because in the thesis I try to give readers the optimal theoretical return with least amount of time spent, and at this point in the thesis, there are enough definitions in place to introduce Bayesian statistics with some sophistication. Second, because the discussion refers to Bayesian statistics, and I feel that Bayesian statistics is a natural way for continued work in this area. For the same reason, 2 pages on shrinkage estimators are offered.
- Time series is the general setting in which the method works. The method is relevant when the observations are sampled as a function of time. Two useful types of time series are Markov chains and autoregressive models, on which method validation is performed. So to understand the results, these are defined. Moreover, Markov chains and autoregressive models are widely used in natural sciences, so for many students, these are natural ingredients in a thesis that aims to give optimal theoretical return for the least time spent. In an economical manner, the theory of Markov chains is developed just enough to understand the Metropolis-Hasting theorem. However, extra emphasis is placed on this result, giving rigorous justification why the theorem and associated algorithm works. Integration in physics models are thoroughly laid out in the section on physics models.
- Numerical parameter estimation is given, because it is used in the results and useful to students. This is also natural because the trade off between the time spent and return for students is relatively large. In particular, this is so because there is sufficient background from previous sections to place these topics in context. Dependent bootstrap is used in the results, but introduction of non parametric independent bootstrap is a pedagogical approach to understand the dependent bootstrap. The Jackknife is given half a page as an introduction. I would have liked to introduce measure theory in the thesis, because it is impossible to prove the consistency of neither the bootstrap estimator nor the estimator for dependent bootstrap without it. However, it is not economical in light of the present results, or the goal of giving an economical introduction to applied probability theory and inference, so this was omitted. With the concepts and theory I give, it is possible to state the conditions for consistency of the bootstrap estimator for independent and identically distributed samples.
- Using the results of the Metropolis-Hasting theorem makes it possible to give transparent implementations of two textbook examples of physics models. The models chosen are the Ising model and an N -electron quantum

dot system. These are two common models encountered in physics. Thus, readers unfamiliar with these can see the connection with theory of Markov chains and the Metropolis-Hastings theorem. These models are also used in the method validation in the results. They are prerequisites for understanding the results of the thesis, and serve in application of much of the material covered in the thesis.

These are the requirements to understand the results. The methods section is a compilation of textbooks results, which are well known, and because of that, references are often omitted. This does not mean that these methods are my results. On the contrary, the method section is a collection of text book material that I have encountered in my studies. However, in cases where I give own examples in the methods section, I mark them by (†).

In summary, the results contain the natural consequences of equation (1.2). Thereafter, an automation, which is sufficiently robust to operate without supervision, is derived. 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.

Chapter 2

Method

2.1 Real analysis results

2.1.1 Functions on metric spaces

Mathematically inclined physicists may argue that for the general physicist, some knowledge of analysis is paramount, and perhaps more important than abstract algebra. Whilst abstract algebra has some tradition in theoretical physics, due to applications in for example quantum field theory (Peskin and Schroeder 1995), analysis is important because the use of functions saturates physics, where differentiation, integration and finally linear analysis play central rôles (Boas 2005). Some may argue that physicists should be concerned when their mathematical operations are valid.

The aim of this chapter is to present definitions and functions that will simplify the understanding of the results given later. Now, consider the following set and induced structure: Assume X is a set and $x, y, z \in X$. A **metric space** is a tuple (X, d) , where $d : X \times X \rightarrow [0, \infty)$ is a function with the properties that: (i) $d(x, y) \geq 0$ and zero if and only if $x = y$, (ii) $d(x, y) = d(y, x)$ and (iii) $d(x, z) \leq d(x, y) + d(y, z)$ for $x, y, z \in X$ (Lindstrøm 2017). The latter property is called the **triangular inequality**. The function d is called a **metric** (Lindstrøm 2017). A set $A \subseteq X$ is said to be **open** if it is the union of sets of the form $B(x; r) = \{y \in X \mid d(x, y) < r\}$ (Munkres 2000). The set $B(x; r)$ is called an **open ball**.

EXAMPLE 1. Let $\mathbb{N} = \{1, 2, 3, \dots\}$ denote the natural numbers, \mathbb{Q} the rational numbers and \mathbb{R} the real numbers. If x, y are points either in \mathbb{N} , \mathbb{Q} or \mathbb{R} , then $d(x, y) \equiv |x - y|$ is a metric on \mathbb{N} , \mathbb{Q} or \mathbb{R} . Here $|\cdot|$ denotes absolute value. So (\mathbb{N}, d) , (\mathbb{Q}, d) and (\mathbb{R}, d) are metric spaces. For proof,

(i) $|x - y| \geq 0$ and $|x - y| = 0$ if and only if $x = y$.

$$(ii) \quad |x - y| = |y - x|$$

(iii) We know $|x - y| \leq |x - z| + |z - y|$ is true for the absolute value according to calculus (Apostol 1961).

There are more interesting examples, as we shall see, but first we need some more definitions. ►

If f is any function between metric space X and Y , then we write $f : X \rightarrow Y$ to indicate that the **domain** of f is X and the **codomain** is Y . We say that f is **continuous** if for all $a \in X$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that $d_Y(f(a), f(x)) < \varepsilon$ for all $d_X(x, a) < \delta$ (Lindstrøm 2017). A function $f : X \rightarrow Y$ is called **surjective** if for every $y \in Y$ there exists an $x \in X$ such that $f(x) = y$. It is called **injective** if for all $x, y \in X$ such that $f(x) = f(y)$ we have $x = y$. If a function is a surjection and an injection¹, it called a **bijection** (Lindstrøm 2017).

EXAMPLE 2. The function $f : X \rightarrow Y$ given by a constant is continuous. Assume $f(x) = b \in Y$ for all $x \in X$. Pick $\varepsilon > 0$ and let $\delta = 1$. Then by axiom (i) for metric spaces, $d_Y(b, b) = 0$, so

$$d_Y(f(x), f(a)) = d_Y(b, b) = 0 < \varepsilon.$$

Since the function f equals the constant b , it is convention to say that f is **identically equal** b . The function $g : X \rightarrow X$ given by $g(x) = x$ is continuous. Pick $\varepsilon > 0$ and let $\delta = \varepsilon$, then

$$d_X(g(x), g(a)) = d_X(x, a) < \delta = \varepsilon.$$

The function g is called the **identity function**. Assume a function $h : X \rightarrow Y$ has the property that there exists a constant $C \geq 0$ such that

$$d_Y(h(x), h(y)) \leq C d_X(x, y) \quad \text{for all } x, y \in X.$$

Is this function continuous? Any such function is called a **Lipschitz function** (J. M. Lee 2013).

If $f, g : X \rightarrow \mathbb{R}$ are continuous functions, then $f + g$ is continuous: Let $\varepsilon > 0$ be given. That means there exists $\delta_f > 0$ and $\delta_g > 0$ such that $|f(x) - f(a)| < \varepsilon/2$ and $|g(x) - g(a)| < \varepsilon/2$ when $d_X(x, a) < \min(\delta_f, \delta_g)$. Let $\delta = \min(\delta_f, \delta_g)$ then

$$|f(x) + g(x) - (f(a) + g(a))| \leq |f(x) - f(a)| + |g(x) - g(a)| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

The proof that $f - g$ is continuous is similar. It turns out that if $f : X \rightarrow Y$ and $g : Y \rightarrow Z$ be continuous functions between metric spaces, then the composition

¹This means it is surjective and injective

$g \circ f : X \rightarrow Z$ is also continuous (Lindstrøm 2017).

It is common to use injective functions to measure the size of sets. If S is a set and there exists an injective function between S and \mathbb{N} , then the set S is called **countable**, and all finite sets are countable (Munkres 2000). Moreover, the set \mathbb{Q} is countable. ►

The **support** of f is the set $\{x \in X \mid f(x) \neq 0\}$. We write $\text{Supp } f$ to denote this set. If $A \subset X$ is any set, then we say that A is **connected** if A is not the union of two disjoint nonempty open sets (Munkres 2000). Naturally we can have functions with connected support, which is just to say that the set $\{x \in X \mid f(x) \neq 0\}$ is not the union of two disjoint nonempty open sets. Assume (X, d_X) and (Y, d_Y) are metric spaces and $f_n : X \rightarrow Y$ is a function for each $n \in \mathbb{N}$. Throughout, a sequence of the type $\{f_n\}_{n=1}^\infty$ is called a **sequence of functions**. We will be interested in the sense in which f_n can converge to some limit function, much in the sense of sequences of numbers. The definitions carry striking resemblance. Assume for every $\varepsilon > 0$ and $x \in X$ there exists some $N \in \mathbb{N}$ such that $d_Y(f(x), f_n(x)) < \varepsilon$ whenever $n \geq N$. If so we say that f_n **converges pointwise** to f (Lindstrøm 2017).

EXAMPLE 3. The function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x) = x$ has support

$$\text{Supp } f = \{x \in \mathbb{R} \mid f(x) \neq 0\} = \{x \in \mathbb{R} \mid x \neq 0\} = \mathbb{R} \setminus \{0\}.$$

The function f does not have connected support. For the set $\text{Supp } f = \mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty)$ is clearly the union of two disjoint nonempty open sets, so it is not connected. The function $g(x) = 0$ is called the zero function of \mathbb{R} and it has support

$$\text{Supp } g = \{x \in \mathbb{R} \mid g(x) \neq 0\} = \{x \in \mathbb{R} \mid 0 \neq 0\} = \emptyset.$$

The empty set, \emptyset is connected since if it was the union of two disjoint nonempty open sets, then it would not have been the empty set, i.e. it would have contained an element. The function $h_n : \mathbb{R} \rightarrow \mathbb{R}$ given by $h_n(x) = e^{-nx}$ has the property that $h_n(x) > 0$ for all $x \in \mathbb{R}$ (Apostol 1961), so

$$\text{Supp } h_n = \{x \in \mathbb{R} \mid h_n(x) \neq 0\} = \{x \in \mathbb{R} \mid e^{-nx} > 0\} = \mathbb{R}.$$

There are not two disjoint open sets with union that equals \mathbb{R} (Munkres 2000). That means $\text{Supp } h_n$ is connected.

The sequence of functions $\{h_n\}$ converges pointwise to g on $(0, \infty)$. To see this, assume that $x \in (0, \infty)$ and assume $\varepsilon > 0$. We can without loss of generality assume $\varepsilon < 1$. Now define N to be the first integer strictly larger than $-\log(\varepsilon)/x$ and let $n \geq N$. Write

$$|h_n(x) - g(x)| = |e^{-nx} - 0| = |e^{-nx}| = e^{-nx} = \underbrace{(e^{-x})^n}_{\in (0,1)} < (e^{-x})^{-\log(\varepsilon)/x} = e^{\log \varepsilon} = \varepsilon,$$

which is the definition. ►

There is another kind of convergence which implies pointwise convergence called uniform convergence. It is convention to say that f_n **converges uniformly to** f if there is a function $f : X \rightarrow Y$ such that for all real numbers $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that if $n \geq N$, then $d_Y(f(x), f_n(x)) < \varepsilon$ for all $x \in X$ (Lindstrøm 2017).

EXAMPLE 4. Define the function $f_n : \mathbb{R} \rightarrow \mathbb{R}$ for all $n \in \mathbb{N}$ given by $f_n(x) = x/n$. The function f_n converges pointwise to zero but it does not converge uniformly: Pick $\varepsilon > 0$ and try to claim that there exists $N \in \mathbb{N}$ such that for all $x \in [0, \infty)$ we have

$$\varepsilon > \overset{!}{|f_n(x) - 0|} = \left| \frac{x}{n} \right| = \frac{x}{n} \quad \text{for all } n \geq N.$$

Now pick $y \in [0, \infty)$ so large that $y > \varepsilon N$. Then

$$\varepsilon < \frac{y}{N} = \left| \frac{y}{N} \right| = \left| \frac{y}{N} - 0 \right| = |f_N(y) - 0|,$$

a contradiction. So f_n does not converge uniformly on $[0, \infty)$ hence it cannot converge uniformly on $\mathbb{R} \supseteq [0, \infty)$. On the other hand, the function $g_n(x) = 1/n$ converges uniformly to zero on \mathbb{R} for if we let $N > 1/\varepsilon$ then by letting $n \geq N$,

$$|g_n(x) - 0| = \left| \frac{1}{n} - 0 \right| = \frac{1}{n} < \frac{1}{1/\varepsilon} = \varepsilon.$$

Since this expression does not depend on x , it must be true for all $x \in X$, which proves uniform convergence according to the definition. ►

2.1.2 Banach spaces and the Fréchet derivative

Assume $\{a_n\}_{n=1}^\infty$ is a sequence in a metric space X . We say that $\{a_n\}$ is a **Cauchy sequence** if for every $\varepsilon > 0$ there is a $N \in \mathbb{N}$ such that $d(a_m, a_n) < \varepsilon$ for all $m, n \geq N$ (J. M. Lee 2011). If there is some $a \in X$ such that for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(a_n, a) < \varepsilon$ for all $n \geq N$, then $\{a_n\}$ is a **convergent sequence** with **limit** a (J. M. Lee 2011). If $f : X \rightarrow Y$ is some continuous function between metric spaces and $\{a_n\}$ has limit a , then $\{f(a_n)\}$ has limit $f(a)$ (Lindstrøm 2017). If every Cauchy sequence in X is convergent, then we say that X is a **complete** metric space (Lindstrøm 2017). If X is complete and for every $\varepsilon > 0$ there exists a finite number of open balls $\{B(x_k; \varepsilon)\}_{k=1}^n$ such that $X \subseteq B(x_1; \varepsilon) \cup B(x_2; \varepsilon) \cup \cdots \cup B(x_n; \varepsilon)$, then X is called **compact** (Lindstrøm 2017). According to Lindstrøm (2017), the extreme value theorem from calculus extends to functions defined on compact metric spaces.

THEOREM 1 (Extreme value theorem). Assume that X is a nonempty, compact metric space and $f : X \rightarrow \mathbb{R}$ a continuous function. Then f has minima and maxima in X .

EXAMPLE 5. Assume that $X = \mathbb{Q}$ and use the metric $d(x, y) = |x - y|$, then define $a_n = \pi + 1/n$. Pick $\varepsilon > 0$ and assume $m, n \geq N$ is the first integer larger than $2/\varepsilon$. We can without loss of generality assume $n < m$. Then using the triangular inequality

$$|a_n - a_m| = \left| \pi + \frac{1}{n} - \pi - \frac{1}{m} \right| = \left| \frac{1}{n} - \frac{1}{m} \right| \leq \left| \frac{1}{n} \right| + \left| \frac{1}{m} \right| < \frac{2}{n} \leq \frac{2}{N} \leq \frac{2}{2/\varepsilon} = \varepsilon.$$

So $\{a_n\}$ is a Cauchy sequence. Is it convergent? According to the definition, the limit $a = \pi$ must be in X . But $X = \mathbb{Q}$ and does not contain π , so $\{a_n\}$ does not converge in X . This means that \mathbb{Q} is not complete since we've found a Cauchy sequence that does not converge. The sequence $\{a_n\}$ is however convergent in \mathbb{R} , and in fact all other Cauchy sequences in \mathbb{R} converge. So \mathbb{R} is complete by definition (Lindstrøm 2017).

Assume X is compact. Let $C(X, \mathbb{R})$ denote the set of continuous functions from (X, d_X) to (\mathbb{R}, d) then according to Lindstrøm (2017), $\rho(f, g) = \sup_{x \in X} \{|f(x) - g(x)|\}$ is a metric on $C(X, \mathbb{R})$. Pick $f, g \in C(X, \mathbb{R})$. First see that $\rho(f, g)$ is well defined since it is finite: As we proved, the difference of two continuous functions is continuous, moreover $|\cdot|$ is a continuous function so if $|h(x)| = |f(x) - g(x)| = |\cdot \circ (f - g)(x)|$ is a composition of continuous functions, so it is continuous. That means h has a maxima in X . Call this point $y \in X$ and write:

$$\rho(f, g) = \sup_{x \in X} \{|f(x) - g(x)|\} = \sup_{x \in X} \{h(x)\} = h(y) < \infty.$$

Verify now the axioms of metric spaces. The first two are more or less obvious. The triangular inequality is more difficult, but follows from the triangular inequality of the metric on \mathbb{R} : Suppose $f, g, h \in C(X, \mathbb{R})$, by the same argument as before it there is a point $y \in X$ such that $\rho(f, g) = |f(y) - g(y)|$. Use this and write

$$\begin{aligned} \rho(f, g) &= |f(y) - g(y)| = |(f(y) - h(y)) + (h(y) - g(y))| \\ &\leq |f(y) - h(y)| + |h(y) - g(y)| \leq \rho(f, h) + \rho(h, g), \end{aligned}$$

since $\rho(h, g)$ is the supremum of the set $\{|h(x) - g(x)| : x \in X\}$, as required. ►

Assume that $x \in X$ and X is a vector space with a norm $\|\cdot\|$. A tuple $(X, \|\cdot\|)$ is called a **normed space**. Every normed space is automatically a metric space (Lindstrøm 2017). This is because the axioms of the metric $d_1(x, y) = \|x - y\|$ is a direct consequence of the axioms of the norm (Lindstrøm 2017). If X is complete as viewed as a metric space under the metric d_1 , then we call X a **Banach space**.

EXAMPLE 6. The set \mathbb{R}^n together with the euclidan norm $\|(x_1, x_2, \dots, x_n)\| = (x_1^2 + x_2^2 + \dots + x_n^2)^{1/2}$ is a Banach space (McDonald and Weiss 2012). It turns

out that $\|f\|_\infty = \rho(f, 0)$, called the **sup-norm**, is a norm on the set $C(X, Y)$ as introduced in example 1. It also turns out that $(C(X, Y), \|\cdot\|_\infty)$ is a Banach space if Y is complete and X compact (McDonald and Weiss 2012). But you now know that \mathbb{R} is complete, so $(C(X, \mathbb{R}), \|\cdot\|)$ is a Banach space. There are many other examples of Banach spaces, see for McDonald and Weiss (2012). ►

If X and Y are normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, then a linear function $T : X \rightarrow Y$ between normed spaces is called bounded if there exists some $K \in \mathbb{R}$ such that $\|T(H)\|_Y \leq K\|H\|_X$ for all $H \in X$ (Rynne and Youngson 2007). Now, assume instead that X, Y are Banach spaces and that T is defined in an open set containing $B \in X$. We say that the limit of $T(H)$ as $H \rightarrow B$ exists and equals C if $B \in X$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|T(H) - C\| < \varepsilon$ for all $\|H - B\|_X < \delta$. We write $\lim_{H \rightarrow B} T(H) = C$ to denote this (Rynne and Youngson 2007). If there is some open set $U \subseteq X$, then we say T is **Fréchet differentiable** at $F \in U$ if there exists a bounded linear function $A : U \rightarrow Y$ such that

$$\lim_{H \rightarrow 0} \frac{\|T(F) - T(F + H) - A(H)\|_Y}{\|H\|_X} = 0. \quad (\text{van der Vaart 1998})$$

2.2 Probability theory

2.2.1 Univariate theory

Introduction to univariate probability theory is necessary because these tools are fundamental in order to understand the results given here. This is both the case when proof is given for the blocking method, but also the automated method which can be used to compute the standard errors quickly. After univariate probability theory has been introduced, it possible to define multivariate probability theory which can be used to study the interplay between random variables, such as those which a time series is comprised of.

Readers familiar with elementary quantum physics will know that if E is a random variable with probability density function (pdf) $g_E : U \rightarrow \mathbb{R}$, and $U = (a, b)$ is an interval of \mathbb{R} , we can suppose that a, b could equal $\pm\infty$ (i.e. $U = \mathbb{R}$). Then the probability $P(E \in (a, x))$, and E is in $(a, x) \subseteq (a, b)$ given by

$$\int_a^x g_E(e) de \equiv G_E(x). \quad (\text{DeGroot and Schervish 2012})$$

The function G_E is called the **cumulative distribution** or cdf of E (DeGroot and Schervish 2012). Since E is a random variable $G_E(b) = 1$, all probabilities integrate to 1. Using the cumulative distribution function, we can define the quantile function. All probability distributions, which have connected support,

have strict monotonuous cdf $G_E : \text{Supp } g_E \rightarrow [0, 1]$. (*Proof.* Suppose $p > q$ then $\int_a^p g_E(e)de - \int_a^q g_E(e)de = \int_q^p g_E(e)de \geq (p - q) \inf_{x \in [q, p]} G_E(x) > 0$ since $G_E > 0$ on its domain). So since G_E is strict monotonuous, there exists an inverse function for G_E (Lindstrøm 2006). We call it $G_E^{-1} : [0, 1) \rightarrow U$ or the **quantile function**, and sometimes the percentile function. If $\alpha \in [0, 1)$, we say that the **100 α -percentile of E** is the value $G_E^{-1}(\alpha)$ (Devore and Berk 2012). Thus the value α , is such that $P(E \in [a, \alpha)) = \alpha$.

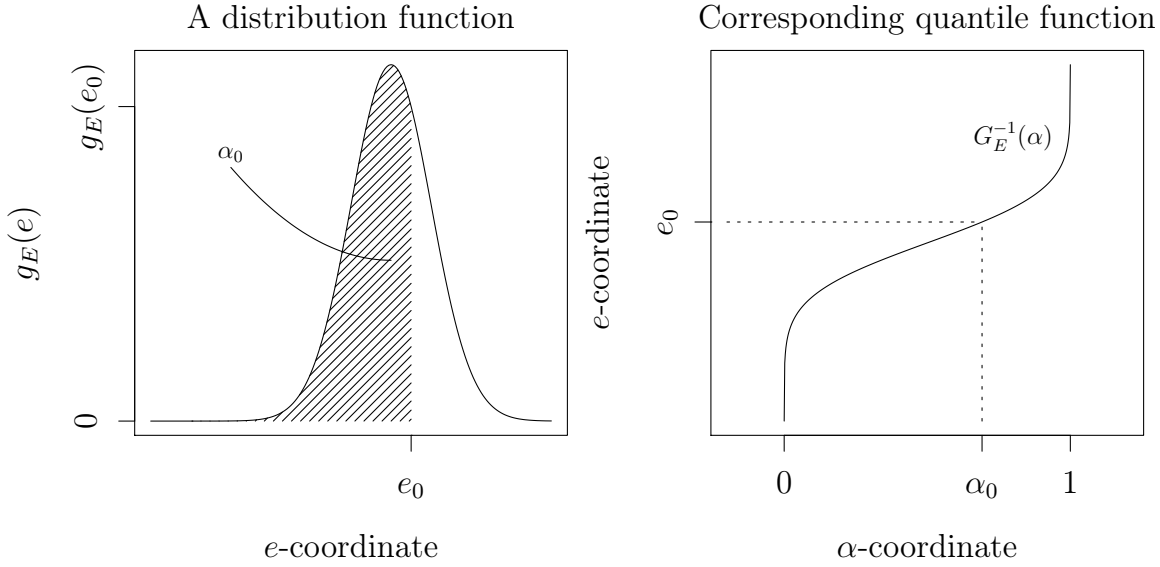


Figure 2.1: To the left: Consider some probability distribution g_E , some value $e_0 \in U$ such that $g_E(e_0) > 0$. We let $\alpha_0 = G_E(e_0)$. To the right: Conversely, the 100 α_0 -percentile is the value $e_0 = G^{-1}(\alpha_0)$. That means the 100 α -percentile is the e -coordinate to which one has to integrate to obtain the probability α_0 .

EXAMPLE 7 (Uniform distributed random variables). A random variable E is uniformly distributed on $[a, b]$ if the pdf of E is

$$g_E(e) = \begin{cases} 1/|b - a| & \text{if } e \in [a, b] \\ 0 & \text{else} \end{cases}. \quad (\text{Devore and Berk 2012})$$

That means that if $x \in [a, b]$, the cdf of E is

$$G_E(e) = \int_a^x g_E(e) \, de = \int_a^x \frac{1}{|b - a|} \, de = \frac{1}{b - a}(x - a).$$

Since g_E has connected support, G_E is strict monotonous. This can also be checked by differentiation. Suppose $x \in [a, b]$ then

$$\frac{\partial G_E}{\partial x}(x) = \frac{1}{b - a}, \quad (2.1)$$

is nowhere zero on $[a, b]$, so it is monotonous. That means the reverse function G_E^{-1} exists (Lindström 2006). As we know from calculus, this function is obtained by solving equation (2.1) for x :

$$G_E(e) = \frac{1}{b-a}(x-a) \quad \text{only if} \quad x = (b-a)G(x) + a, \\ \text{only if} \quad G_E^{-1}(\alpha) = (b-a)\alpha + a,$$

which is the quantile function of E . ►

In the rest of this section, U is an interval. Readers should be aware that if $f : \mathbb{R} \rightarrow \mathbb{R}$ is a continuous function, then we can compute the expected value of $f(E)$ by

$$\langle f(E) \rangle = \int_U f(e)g(e)de.$$

It is easy to understand that $f(E)$ is a random variable if E is one. Similarly, we can compute the standard deviation and variance of E . In statistics, some functions of random variables are more important than others. The average of random variables is one such function.

I will now introduce some new terminology. If E_1, E_2, \dots, E_n are random variables each with identical pdf g_E , then we say that E_1, E_2, \dots, E_n are **identically distributed** (DeGroot and Schervish 2012). If each E_j has expected value μ , then the average or **sample mean** of the E_j is $\mathbb{R}^n \rightarrow \mathbb{R}$ -function

$$E_1, E_2, \dots, E_n \mapsto \frac{1}{n}(E_1 + E_2 + \dots + E_n) \equiv \bar{E}. \quad (2.2)$$

Another important such function is the sample variance. Suppose that the true variance of each E_j is σ^2 . Then, we define the **sample variance** as the $\mathbb{R}^n \rightarrow \mathbb{R}$ -function

$$E_1, E_2, \dots, E_n \mapsto \frac{1}{n} \sum_{i=1}^n (E_i - \bar{E})^2 \equiv \hat{\sigma}^2. \quad (\text{Devore and Berk 2012}) \quad (2.3)$$

I claim that the functions defined in (2.2) and (2.3) are the most important functions to a statistician. The reason is that by computing \bar{E} and $\hat{\sigma}^2$, we receive estimates for the true values of μ and σ^2 . Let's make this precise. If A is a parameter in a probabilistic model we are interested in, we say that \hat{A} is an **estimator** of A if the absolute value $|\hat{A}(E_1, \dots, E_n) - A|$ is a small number (Devore and Berk 2012). And, if e_1, \dots, e_n are realizations or measurements of E_1, \dots, E_n , then $\hat{A}(e_1, \dots, e_n) \equiv a$ is an **estimate** of A . If the expected value of the estimator \hat{A} is A , then we say that \hat{A} is **unbiased** (DeGroot and Schervish 2012). We define the $\text{Bias}(\hat{A}; A) = \langle \hat{A} \rangle - A$, which measures how far from unbiased each estimator is. It is easy to see that if each of the E_j s are identically

distributed with expectation μ , then $\mu \bar{E}$ is an unbiased estimator of μ . Just write

$$\langle \bar{E} \rangle = \frac{1}{n} \langle E_1 + \cdots + E_n \rangle = \frac{1}{n} (\langle E_1 \rangle + \cdots + \langle E_n \rangle) = \frac{1}{n} (\mu + \cdots + \mu) = \frac{1}{n} n\mu = \mu.$$

Similarly, it is possible to see that $\hat{\sigma}^2$ is *not* an unbiased estimator of σ^2 even if the E_j s are indetically distributed (Devore and Berk 2012). It is however possible to define an estimator S^2 which is unbiased under mild conditions (Devore and Berk 2012), as we shall see later:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (E_i - \bar{E})^2 = \frac{n}{n-1} \hat{\sigma}^2. \quad (2.4)$$

Since we saw that \bar{E} and $\hat{\sigma}^2$ are random variables, one can ask what their variance is. The answer to this question can sometimes be difficult, and in fact the purpose of the thesis is to build an estimator of the variance of \bar{E} for one such case. However, it is possible to make some restrictions on the E_j s such that this is easy. To understand the problem better, we must define another concept called independence.

We've assumed that the probability distribution of each E_j is g_E . Using this we can compute that if $A_j \subseteq U$ then the probability $P(E_j \in A)$ that the realization e_j of E_j satisfies $e_j \in A_j$ is

$$P(E_j \in A_j) = \int_{A_j} g_E(e) de. \quad (\text{DeGroot and Schervish 2012}) \quad (2.5)$$

Consider now another question. What is the probability $P(E_1 \in A_1, E_2 \in A_2)$ if E_i and E_j are random variables, $e_1 \in A_1 \subseteq U$ and $e_2 \in A_2 \subseteq U$? Perhaps you'd guess that it is $P(E_1 \in A_1, E_2 \in A_2) = P(E_1 \in A_1)P(E_2 \in A_2)$? To find out if this is true, we define what it means that two random variables are independent. We will make a more useful definition in the next chapter, but motivate it here. We say that E_1, \dots, E_n are **independent** if having measured that $E_j = e_j$ does not reveal anything about the values of the other E_k s. It turns out that if E_1, E_2 are independent, then

$$P(E_1 \in A_1, E_2 \in A_2) = P(E_1 \in A_1)P(E_2 \in A_2). \quad (\text{DeGroot and Schervish 2012})$$

PROPOSITION 1. *In fact, if a_1, \dots, a_n are real (non-random) numbers and E_1, \dots, E_n are independent, then we get the following useful identities. Here, assume that*

$V(E_j)$ denotes the variance of E_j , then

$$V(a_1E_1 + a_2E_2 + \cdots + a_nE_n) = a_1^2V(E_1) + a_2^2V(E_2) + \cdots + a_n^2V(E_n) \quad (2.6)$$

$$\langle E_1 \cdot E_2 \cdot E_3 \cdots E_n \rangle = \langle E_1 \rangle \langle E_2 \rangle \langle E_3 \rangle \cdots \langle E_n \rangle \quad (2.7)$$

$$P(E_1 \in A_1, E_2 \in A_2, \cdots, E_n \in A_n) = P(E_1 \in A_1)P(E_2 \in A_2) \cdots P(E_n \in A_n). \quad (2.8)$$

Let's tackle the problem of calculating the variance $V(\bar{E})$ for the case that E_1, \cdots, E_n are independent and identically distributed (abbreviated iid).

$$\begin{aligned} V(\bar{E}) &\stackrel{(2.2)}{=} V\left(\frac{1}{n}(E_1 + \cdots + E_n)\right) = V\left(\frac{1}{n}E_1 + \cdots + \frac{1}{n}E_n\right) \\ &\stackrel{(2.6)}{=} \frac{1}{n^2}V(E_1) + \cdots + \frac{1}{n^2}V(E_n) = \frac{\sigma^2}{n}. \end{aligned} \quad (2.9)$$

The square root of this expression $V(\bar{E})^{1/2}$ is called the **standard error** of \bar{E} (Devore and Berk 2012).

EXAMPLE 8 (The bounded linear functional A). Assume $X = [a, b]$. Then according to Munkres (2000), X is a compact and connected subset of the real numbers and according to section 2.1, the set of continuous functions $C(X, \mathbb{R})$ is a Banach space under the sup norm $\|\cdot\|_\infty$ of example 6. Let $H(X)$ be the largest open set of cdfs on X . It is clear that $H(X) \subset C(X, \mathbb{R})$ because every differentiable function is continuous and every cdf is differentiable according to the Fundamental theorem of calculus (Apostol 1961). Hence if $F \in H(X)$, then there exists a continuous function $f = dF/dx$. If $\psi : X \rightarrow \mathbb{R}$ is another continuous function, then it is also bounded according to theorem 1. Thus, there exists some $K \in \mathbb{N}$ such that $|\psi(x)| \leq K$ for all $x \in X$.

Let us check if the function $A : H(X) \rightarrow \mathbb{R}$ given by

$$A(F) = \int_a^b \psi(x) dF(x) = \int_a^b \psi(x) f(x) dx,$$

is a bounded linear functional. Assume $F \in H(X)$ then $\|F\|_\infty = 1$ because

$$\|F\|_\infty = \sup_{x \in X} \{ |F(x)| \} = \sup_{x \in X} \left\{ \left| \int_a^x f(y) dy \right| \right\} = \left| \int_a^b f(y) dy \right| = |1| = 1.$$

Using this, it is easy to see that A is bounded. Just write:

$$|A(F)| = \left| \int_a^b \psi(x) f(x) dx \right| \leq \int_a^b |\psi(x) f(x)| dx \leq K \int_a^b f(x) dx = K \cdot 1 = K \|F\|_\infty.$$

To see that it is linear, let $\alpha, \beta \in \mathbb{R}$ and G be another cdf with g as pdf, then

$$\begin{aligned} A(\alpha F + \beta G) &= \int_a^b \psi(x) \frac{d(\alpha F + \beta G)}{dx}(x) dx \\ &= \alpha \int_a^b \psi(x) f(x) dx + \beta \int_a^b \psi(x) g(x) dx = \alpha A(F) + \beta A(G), \end{aligned}$$

which means that A is also linear. ►

2.2.2 Multivariate probability theory

Near the end of the previous section we introduced independence and discussed the probability that $E_1 \in A_1 \subseteq U$ and $E_2 \in A_2 \subseteq U$ was true. Here we want to explore similar concepts further, but we will work with n random variables E_j . In this section we will almost always assume that E_1, \dots, E_n are no longer independent, but still identically distributed. That means $P(E_1 \in A_1, E_2 \in A_2, \dots, E_n \in A_n) = P(E_1 \in A_1)P(E_2 \in A_2) \cdots P(E_n \in A_n)$ is generally false. But as you can imagine, there exists some function g which determines $P(E_1 \in A_1, E_2 \in A_2, \dots, E_n \in A_n)$, the probability that the measured value of $e_1 \in A_1$ and $e_2 \in A_2$ and \cdots and $e_n \in A_n$. This function g is called the **joint probability function** (Devore and Berk 2012) and is defined implicitly by

$$P(E_1 \in A_1, \dots, E_n \in A_n) = \int_{A_1} \int_{A_2} \cdots \int_{A_n} g(e_1, e_2, \dots, e_n) de_1 de_2 \cdots de_n.$$

In analogy to the univariate case, suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuous function, then we can use g to compute the expected value

$$\langle f(E_1, \dots, E_n) \rangle = \int_U \int_U \cdots \int_U f(e_1, e_2, \dots, e_n) g(e_1, e_2, \dots, e_n) de_1 de_2 \cdots de_n. \quad (2.10)$$

Using this formula we can define a new function that is similar to the variance for single random variables E . We say that the **covariance of E_i, E_j** is the function

$$E_i, E_j \mapsto \langle (E_i - \langle E_i \rangle)(E_j - \langle E_j \rangle) \rangle \equiv \text{Cov}(E_i, E_j). \quad (2.11)$$

Clearly, $\text{Cov}(E_i, E_j) = \text{Cov}(E_j, E_i)$. If you've taken quantum mechanics, you see that if $i = j$, then $\text{Cov}(E_i, E_i) = V(E_i)$ (use the definition of variance to see this). The covariance is important, because it is a measure of the linear relation between E_i and E_j for all i, j . By that, we mean that E_i and E_j has a tendency to vary in relation to one another. If the covariance $\text{Cov}(E_i, E_j)$ is positive, then we expect that a large value of E_i (relative to the expected value $\langle E_i \rangle$) tends to occur jointly with a large value of E_j (relative to $\langle E_j \rangle$). Or if a small value of E_i is measured (relative to the expected value $\langle E_i \rangle$), this tends to occur jointly

with a small value of E_j (relative to $\langle E_j \rangle$). That is, the sign and magnitude of E_i and E_j tends to be similar (Devore and Berk 2012).

On the other hand, if the covariance is negative, a large positive value of E_i tend to occur together with a large negative value of E_j . These are again measured relative to their expected value. To write this in mathematics is to say

$$\langle E_1 E_2 \rangle = \langle E_1 \rangle \langle E_2 \rangle + \text{Cov}(E_1, E_2), \quad (2.12)$$

$$V(E_1 \pm E_2) = V(E_1) + V(E_2) \pm 2 \text{Cov}(E_1, E_2). \quad (2.13)$$

Later, we will use linear algebra to handle a large number of random variables. We will form vectors of random variables $\mathbf{E} = (E_1, E_2, \dots, E_n)^\top$, called a **random vector** and say that the expected value of the vector and the covariance matrix are given by

$$\langle \mathbf{E} \rangle = \begin{bmatrix} \langle E_1 \rangle \\ \langle E_2 \rangle \\ \vdots \\ \langle E_n \rangle \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \text{Cov}(E_1, E_1) & \text{Cov}(E_1, E_2) & \cdots & \text{Cov}(E_1, E_n) \\ \text{Cov}(E_2, E_1) & \text{Cov}(E_2, E_2) & \cdots & \text{Cov}(E_2, E_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(E_n, E_1) & \text{Cov}(E_n, E_2) & \cdots & \text{Cov}(E_n, E_n) \end{bmatrix}, \quad (2.14)$$

respectively (Agresti 2015). The covariance matrix is important because it fully describes the correlation between the observations E_i . Since we want to use the covariances to estimate $V(\bar{X})$, this quantity is of central importance. We shall see later, that in some cases, it can be reduced to the autocovariance γ . Assume now that A is any non-random matrix and \mathbf{a} is a non-random vector in \mathbb{R}^n , then by using the definition of the variance and linearity of the expected value, we get

$$\langle \mathbf{E} - \mathbf{a} \rangle = \langle \mathbf{E} \rangle - \mathbf{a} \quad \text{and} \quad \langle A\mathbf{E} \rangle = A \langle \mathbf{E} \rangle \quad \text{and} \quad \langle \mathbf{E}^\top \mathbf{a} \rangle = \langle \mathbf{E}^\top \rangle \mathbf{a}. \quad (2.15)$$

In analogue to $V(E_i)$ we let $V(\mathbf{E})$ denote the covariance matrix. These quantities will take the role of the expected value and variance of single random variables in multivariate theories. Since we defined the covariance in terms of a product of random variables (see equation (2.11)), and we noted that $\text{Cov}(E_i, E_j) = \text{Cov}(E_j, E_i)$, it follows that Σ is symmetric (Mathai and Provost 1992). Note that the diagonal elements of Σ is precisely the variances of $E_1 \dots, E_n$. We show that Σ is positive semidefinite. Note that it follows from equations (2.11) and (2.14) that Σ is the outer product $\Sigma = \langle (\mathbf{E} - \langle \mathbf{E} \rangle)(\mathbf{E} - \langle \mathbf{E} \rangle)^\top \rangle$. Therefore, if $\mathbf{y} \in \mathbb{R}^n$

$$\begin{aligned} \mathbf{y}^\top \Sigma \mathbf{y} &= \mathbf{y}^\top \langle (\mathbf{E} - \langle \mathbf{E} \rangle)(\mathbf{E} - \langle \mathbf{E} \rangle)^\top \rangle \mathbf{y} \stackrel{(2.15)}{=} \langle \mathbf{y}^\top (\mathbf{E} - \langle \mathbf{E} \rangle)(\mathbf{E} - \langle \mathbf{E} \rangle)^\top \mathbf{y} \rangle \\ &= \langle [(\mathbf{E} - \langle \mathbf{E} \rangle)^\top \mathbf{y}]^\top (\mathbf{E} - \langle \mathbf{E} \rangle)^\top \mathbf{y} \rangle = \langle \|(\mathbf{E} - \langle \mathbf{E} \rangle)^\top \mathbf{y}\|^2 \rangle \geq 0, \end{aligned} \quad (2.16)$$

where we used the linearity of the expectation in the second step. Another property that is useful is, if A is some $m \times n$ matrix, then $V(A\mathbf{E}) = AV(\mathbf{E})A^\top$

(Agresti 2015). To see that this is true, just note that if A_{ij} are the elements of A , then the i th element of $A\mathbf{E}$ is

$$\begin{aligned} (A\mathbf{E})_i = \sum_{k=1}^n A_{ik}E_k \quad \text{only if} \quad V(A\mathbf{E})_{ij} &= \text{Cov} \left(\sum_{k=1}^n A_{ik}E_k, \sum_{l=1}^n A_{jl}E_l \right) \\ &= \sum_{k=1}^n \sum_{l=1}^n A_{ik} \underbrace{\text{Cov}(E_k, E_l)}_{=V(\mathbf{E})_{kl}} A_{lj}^T. \end{aligned} \quad (2.17)$$

Finally note that since $V(\mathbf{E}) = \langle (\mathbf{E} - \langle \mathbf{E} \rangle)(\mathbf{E} - \langle \mathbf{E} \rangle)^T \rangle$, it is immediate that $V(\mathbf{E} + \mathbf{a}) = V(\mathbf{E})$ and $V(\mathbf{E}) = \langle \mathbf{E}\mathbf{E}^T \rangle - \langle \mathbf{E} \rangle \langle \mathbf{E} \rangle^T$ (Mathai and Provost 1992). We summarize these findings in the following proposition:

PROPOSITION 2. *If \mathbf{E} is a random vector with covariance matrix $V(\mathbf{E}) = \Sigma$, and there are non-random variables $a \in \mathbb{R}$, $\mathbf{a} \in \mathbb{R}^n$ and $m \times n$ matrix A and then,*

- $\langle A\mathbf{E} \rangle = A \langle \mathbf{E} \rangle$ and $\langle \mathbf{E}^T \mathbf{a} \rangle = \langle \mathbf{E}^T \rangle \mathbf{a}$ and $\langle \mathbf{E} - \mathbf{a} \rangle = \langle \mathbf{E} \rangle - \mathbf{a}$
- $V(\mathbf{E}) = \Sigma$ is a positive semidefinite and symmetric $n \times n$ -matrix.
- The covariance matrix can be written $V(\mathbf{E}) = \Sigma = \langle (\mathbf{E} - \langle \mathbf{E} \rangle)(\mathbf{E} - \langle \mathbf{E} \rangle)^T \rangle = \langle \mathbf{E}\mathbf{E}^T \rangle - \langle \mathbf{E} \rangle \langle \mathbf{E} \rangle^T$.
- $V(A\mathbf{E}) = AV(\mathbf{E})A^T$ and $V(\mathbf{E} + \mathbf{a}) = V(\mathbf{E})$.

At the end of the previous section we said that we would make a more useful definition of independence.

Suppose g_{E_1}, \dots, g_{E_n} is the probability distribution of E_1, \dots, E_n , then we say that E_1, \dots, E_n are **independent** if the joint probability distribution g factors as

$$g(e_1, e_2, \dots, e_n) = g_{E_1}(e_1)g_{E_2}(e_2) \cdots g_{E_n}(e_n). \quad (\text{DeGroot and Schervish 2012})$$

If random variables are not independent, we say that they are **dependent**. Let's now prove this.

THEOREM 2. *If E_1, \dots, E_n are independent then $\text{Cov}(E_i, E_j) = 0$ whenever $i \neq j$, moreover the covariance matrix is diagonal.*

PROOF. Assume the hypothesis that E_1, \dots, E_n are independent is true. Assume $i \neq j$, then the pair E_i, E_j has a joint probability distribution g such that $g =$

$g_{E_i}g_{E_j} (*)$. Using this, compute the covariance

$$\begin{aligned}
\text{Cov}(E_i, E_j) &= \langle (E_i - \langle E_i \rangle)(E_j - \langle E_j \rangle) \rangle = \int_U \int_U \left((e_i - \langle E_i \rangle)(e_j - \langle E_j \rangle) \right) g(e_i, e_j) de_i de_j \\
&\stackrel{(*)}{=} \int_U \int_U \left((e_i - \langle E_i \rangle)(e_j - \langle E_j \rangle) \right) g_{E_i}(e_i) g_{E_j}(e_j) de_i de_j \\
&= \int_U (e_i - \langle E_i \rangle) g_{E_i} de_i \int_U (e_j - \langle E_j \rangle) g_{E_j} de_j \\
&= \left(\langle E_i \rangle - \langle E_i \rangle \right) \left(\langle E_j \rangle - \langle E_j \rangle \right) = 0^2 = 0.
\end{aligned}$$

Since we assumed $i \neq j$, this means that all the off-diagonal elements of Σ are zero, so Σ is diagonal. ■

The converse is not necessarily true; that is $\text{Cov}(E_i, E_j) = 0$ does not imply that E_i, E_j are independent. However, we will soon see that there exists some cases where this does hold.

Finally, we define conditional probability. Suppose the value $e_1 \in U$ is observed for E_1 . If E_1 and E_2 are not independent, then observing $E_1 = e_1$ reveals information about E_2 (this follows by the first definition given for independence). Observing $E_1 = e_1$ reveals information about E_2 , hence the probability distribution of E_2 , since the distribution contains all information about E_2 . Naturally then, the new probability distribution obtained after observing $E_1 = e_1$ is denoted, $g_{E_2|E_1=e_1}$, is different from g_{E_2} . To make this distinction explicit, we call g_{E_i} the **marginal distribution of E_i** , and $g_{E_2|E_1=e_1}$ is called the **conditional probability distribution of E_2 given $E_1 = e_1$** (Devore and Berk 2012). As you can imagine, since the probability distribution changes after observing $E_1 = e_1$, the probability that $E_2 \in A_2 \subseteq U$ is determined by

$$P(E_2 \in A_2 \mid E_1 = e_1) = \int_{A_2} g_{E_2|E_1=e_1}(e_2|e_1) de_2. \quad (2.18)$$

The random variable determined by $g_{E_2|E_1=e_1}$ is denoted by $E_2|E_1 = e_1$. We can also have more random variables than 2; to do this, it is convenient to use vector notation. Suppose $A_1, \dots, A_d \subseteq U$. If $\mathbf{E} = (E_1, \dots, E_d)$ is a d -dimensional random vector and we observed that $\mathbf{K} = \mathbf{k}$ and \mathbf{E} and \mathbf{K} have dependent components, then there is a probability distribution f such that

$$P(E_1 \in A_1, \dots, E_d \in A_d) = \int_{A_1} \int_{A_2} \dots \int_{A_d} f(e_1, e_2, \dots, e_d) de_1 de_2 \dots de_d. \quad (2.19)$$

The random variable determined by f is denoted by $\mathbf{E}|\mathbf{K} = \mathbf{k}$. To understand this better, consider an example.

EXAMPLE 9 (Multivariate normal distribution). *During studies of science one often encounter the univariate **normal probability distribution** with expectation μ and variance σ^2*

$$f(x; \mu, \sigma) = \frac{1}{(2\pi)^{1/2}\sigma} e^{-(1/2)(x-\mu)^2/\sigma^2}. \quad (2.20)$$

*Many phenomena are normal distributed, or approximately normal distributed making this distribution useful. Then, E_1, \dots, E_n are **multivariate normal distributed**, if all linear combinations of E_1, \dots, E_n are normal distributed. If E_i has expected value μ and variance σ^2 , then the marginal distribution of E_i is given by equation (2.20) (Agresti 2015). If $\mu = 0$ and $\sigma^2 = 1$, then we call the associated random variable **standard normal** distributed. If $\mathbf{E} = (E_1, \dots, E_n)^\top$. Then the elements of \mathbf{E} has joint probability distribution with expected value $\boldsymbol{\mu}$ and if the covariance matrix of \mathbf{E} , Σ is positive definite, then the probability distribution is the $\mathbb{R}^n \rightarrow \mathbb{R}$ function*

$$f(\mathbf{e}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{e} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{e} - \boldsymbol{\mu}) \right), \quad (2.21)$$

where $|\cdot|$ denotes the determinant (Agresti 2015). As such it is easy to understand that if \mathbf{E} is multivariate normal with expectation $\boldsymbol{\mu}$, then $\mathbf{E} - \boldsymbol{\mu}$ is multivariate normal with expectation $\mathbf{0}$ and covariance matrix Σ . It turns out that the multivariate normal has other interesting transformation properties, but they are not necessary to introduce in order to understand this thesis. However, if $\mathbf{E}_1 = (E_1, \dots, E_d)^\top$ and $\mathbf{E}_2 = (E_{d+1}, \dots, E_n)$, what is the conditional probability distribution for $\mathbf{E}_1 | \mathbf{E}_2 = \mathbf{e}_2$?

Suppose the covariance matrix for $(E_1, E_2)^\top$ is positive definite, then according to equation (2.14) it can be written in block-diagonal form

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad (2.22)$$

such that Σ_{11} and Σ_{22} are the covariance matrices of \mathbf{E}_1 and \mathbf{E}_2 respectively, whilst $\Sigma_{12} = \Sigma_{21}^\top$ contains the remaining elements of Σ . If Σ_{22} is invertible and we let $\boldsymbol{\mu}_i$ denote the expected value of \mathbf{E}_i then the conditional probability distribution for $\mathbf{E}_1 | \mathbf{E}_2 = \mathbf{e}_2$ is given by equation (2.21), but with expected value

$$\boldsymbol{\mu}' = \langle \mathbf{E}_1 | \mathbf{E}_2 = \mathbf{e}_2 \rangle = \boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{e}_2 - \boldsymbol{\mu}_2)$$

and covariance matrix $\Sigma' = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.$

Thus, the conditional distribution of a pair of multivariate normal random variables is itself multivariate random normal (Agresti 2015). ►

The purpose of this is to introduce conditional probability. This is a way of relating the three types of probabilities we introduced above. We say that **conditional probability** is the relation

$$P(E_1 \in A_1, \dots, E_d \in A_d) = P(E_d \in A_d | E_{d-1} \in A_{d-1} \dots E_1 \in A_1) \quad (2.23) \\ \times P(E_{d-1} \in A_{d-1} \dots E_1 \in A_1).$$

By iteratively use of the definition, we obtain the **product rule** (DeGroot and Schervish 2012) of probability:

$$P(E_1 \in A_1, \dots, E_d \in A_d) = P(E_d \in A_d | E_{d-1} \in A_{d-1} \dots E_1 \in A_1) \quad (2.24) \\ \times P(E_{d-1} \in A_{d-1} | E_{d-2} \in A_{d-2} \dots E_1 \in A_1) \dots P(E_1 \in A_1)$$

Let's consider how one could make this consistent. Let us consider the implications:

$$\int_{A_1} \int_{A_2} \dots \int_{A_d} f(e_1, e_2, \dots, e_d) de_1 de_2 \dots de_d \stackrel{(2.18)}{=} P(E_1 \in A_1, \dots, E_d \in A_d) \\ \stackrel{(2.18)}{=} P(E_d \in A_d | E_{d-1} \in A_{d-1} \dots E_1 \in A_1) P(E_{d-1} \in A_{d-1} | E_{d-2} \in A_{d-2} \dots E_1 \in A_1) \\ = \int_{A_1} \int_{A_2} \dots \int_{A_d} g_{E_d | E_{d-1}=e_{d-1} \dots E_1=e_1}(e_d | e_{d-1}, \dots, e_1) f(e_{d-1}, \dots, e_1) de_1 de_2 \dots de_d$$

One way to obtain consistency would be to require

$$f(e_1, e_2, \dots, e_d) = g_{E_d | E_{d-1}=e_{d-1} \dots E_1=e_1}(e_d) f(e_{d-1}, \dots, e_1). \quad (2.25)$$

In fact, this is the right answer ².

At the end I will introduce a convenient notation. Suppose \mathbf{E} is a multivariate random variable with covariance matrix Σ and expected value $\boldsymbol{\mu}$. Then, we write $\mathbf{E} \sim N(\boldsymbol{\mu}, \Sigma)$ to denote this. If \mathbf{E} is only approximately multivariate normal, we write $\mathbf{E} \approx N(\boldsymbol{\mu}, \Sigma)$ to denote this (Devore and Berk 2012).

2.2.3 Preliminary probability theory

In the previous subsection we saw that if two random variables are independent, then their covariance is zero. We say that if the covariance of two random variables E_1, E_2 is zero, then they are **uncorrelated**. In the special case that E_1, \dots, E_n are multivariate normal, 'uncorrelatedness' is the same as independence. In fact, since the normal distribution is so common, some physicists may think that these are identical. In strict mathematical sense, that is false, but sometimes it is a useful approximation, as section 2.5 shows. We prove

²If you want to prove the converse, see chapter 7 of McDonald and Weiss (2012)

THEOREM 3. Suppose E_1, \dots, E_n are multivariate normal random variables with expectation $\boldsymbol{\mu}$ and covariance matrix Σ . Then E_1, \dots, E_n are independent if and only if they are uncorrelated.

PROOF. Implication to the right is exactly the statement of theorem 2, so it suffices to show implication to the left. Define real numbers $\sigma_i^2 \equiv V(E_i)$. Suppose each E_1, \dots, E_n are uncorrelated. That means $\text{Cov}(E_i, E_j) = 0$ if $i \neq j$ and $\text{Cov}(E_i, E_j) = V(E_i)$ if $i = j$. So the covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$. Since Σ is a diagonal matrix, it is invertible, and the distribution function of $\mathbf{E} = (E_1, \dots, E_n)^\top$ exists according to example 9, and is given by

$$\begin{aligned} f(\mathbf{e}) &= (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{e} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{e} - \boldsymbol{\mu}) \right); \quad \mathbf{e} = (e_1, \dots, e_n)^\top \\ &= (2\pi)^{-n/2} \prod_{i=1}^n [\sigma_i^{-1}] \exp \left(-\frac{1}{2} \sum_{i=1}^n (e_i - \mu_i)^2 / \sigma_i^2 \right); \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top \\ &= \prod_{i=1}^n \underbrace{\frac{1}{(2\pi)^{1/2} \sigma_i} \exp \left(-\frac{(e_i - \mu_i)^2}{2\sigma_i^2} \right)}_{\equiv f_i(e_i; \mu_i, \sigma_i)} \equiv \prod_{i=1}^n f_i(e_i; \mu_i, \sigma_i), \end{aligned}$$

is a product of marginal distributions of the form (2.20). So E_1, \dots, E_n are independent. ■

At the end of the last section, we introduced the normal distribution. In applications, we want to juggle random variables in a similar way as we juggle real (non-random) variables in ordinary analysis. As you can imagine, a random variable raised to a power appears all the time in applications. It turns out that determining the properties of a random variable raised to a power is a surprisingly difficult task. In fact, entire books are dedicated to the problem. See for example Mathai and Provost (1992). The most important case is if Z is a standard normal random variable (Devore and Berk 2012). Then, we say that $X = Z^2$ is **chi-square distributed with one degree of freedom** (later: dof). A sum of n independent chi-square random variables with 1 dof, $\sum_{i=1}^n (Z_i)^2$, is said to be **chi-square distributed with n degrees of freedom** and its pdf is

$$g(x) = \begin{cases} \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{if } x \leq 0 \end{cases}. \quad (\text{Agresti 2015}) \quad (2.26)$$

Here, $\Gamma : [0, \infty) \rightarrow [1, \infty)$ denote the Gamma-function. See figure 3.1 for a plot of this pdf. We conveniently write $X \sim \chi_n^2$ to mean that X is chi-square distributed with n dof. Or if X is only approximately chi-square we write $X \approx \chi_n^2$. It is useful to note that

$$\langle X \rangle = n \quad \text{and} \quad V(X) = 2n \quad (\text{Agresti 2015}) \quad (2.27)$$

Another very useful property of the chi-square distribution is given in the following theorem

THEOREM 4. *Consider $\mathbf{E} = (E_1, \dots, E_n)^\top$ is multivariate normal with expected value $\boldsymbol{\mu}$ and invertible covariance matrix Σ , then $(\mathbf{Y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \sim \chi_n^2$.*

PROOF. (Step 1) We show that Σ^{-1} has a principal square root. We have seen that Σ is symmetric, therefore it has orthogonal diagonalization $\Sigma = PDP^\top$ (Lay 2012). Moreover since we've also seen that it is positive semidefinite, all of its eigenvalues $d_i \geq 0$ are non-negative. Pick $\mathbf{x} \in \mathbb{R}^n$, then

$$\begin{aligned} \mathbf{x}^\top \Sigma^{-1} \mathbf{x} &= \mathbf{x}^\top (PDP^\top)^{-1} \mathbf{x} = \mathbf{x}^\top \underbrace{(P^\top)^{-1}}_{=P} D^{-1} \underbrace{P^{-1}}_{=P^\top} \mathbf{x} \\ &= (P^\top \mathbf{x})^\top D^{-1} \underbrace{P^\top \mathbf{x}}_{\equiv \mathbf{y}} = \sum_{i=1}^n y_i^2 d_i^{-1} \geq 0. \end{aligned}$$

So Σ^{-1} is positive semidefinite. Also, since Σ is symmetric, Σ^{-1} is symmetric since $(\Sigma^{-1})^\top = (\Sigma^\top)^{-1} = (\Sigma)^{-1}$. This proves that Σ^{-1} has decomposition $\Sigma^{-1} = (\Sigma^{-1})^{1/2} (\Sigma^{-1})^{1/2}$, where $(\Sigma^{-1})^{1/2}$ is a symmetric, semidefinite matrix.

(Step 2) Define $\mathbf{Z} = (\Sigma^{-1})^{1/2} (\mathbf{Y} - \boldsymbol{\mu})$, the vector \mathbf{Z} is a linear combination of the elements of $\mathbf{Y} - \boldsymbol{\mu}$, so since $\mathbf{Y} - \boldsymbol{\mu}$ is multivariate normal according to example 9, then \mathbf{Z} is also multivariate normal according to example 9. Let's determine the expected value and covariance matrix of \mathbf{Z} , since this determines the distribution of \mathbf{Z} according to equation (2.21). Using linearity of the expectation we obtain

$$\langle \mathbf{Z} \rangle = \langle (\Sigma^{-1})^{1/2} (\mathbf{Y} - \boldsymbol{\mu}) \rangle \stackrel{(2.15)}{=} (\Sigma^{-1})^{1/2} (\langle \mathbf{Y} \rangle - \boldsymbol{\mu}) = (\Sigma^{-1})^{1/2} (\boldsymbol{\mu} - \boldsymbol{\mu}) = \mathbf{0}.$$

The covariance matrix is

$$\begin{aligned} V((\Sigma^{-1})^{1/2} (\mathbf{Y} - \boldsymbol{\mu})) &= V((\Sigma^{-1})^{1/2} \mathbf{Y}) \stackrel{(2.17)}{=} (\Sigma^{-1})^{1/2} V(\mathbf{Y}) (\Sigma^{-1})^{1/2} \\ &= (\Sigma^{-1})^{1/2} \Sigma (\Sigma^{-1})^{1/2}. \end{aligned} \tag{2.28}$$

So, if the matrix $(\Sigma^{-1})^{1/2}$ is invertible, then $V((\Sigma^{-1})^{1/2} (\mathbf{Y} - \boldsymbol{\mu})) = I_n$ is the identity matrix because

$$V((\Sigma^{-1})^{1/2} (\mathbf{Y} - \boldsymbol{\mu})) (\Sigma^{-1})^{1/2} \stackrel{(2.28)}{=} (\Sigma^{-1})^{1/2} \Sigma (\Sigma^{-1})^{1/2} (\Sigma^{-1})^{1/2} = (\Sigma^{-1})^{1/2}.$$

And if $V((\Sigma^{-1})^{1/2} (\mathbf{Y} - \boldsymbol{\mu})) = I_n$ is the identity matrix, then the proof is complete since $(\mathbf{Y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{Z}^\top \mathbf{Z}$ is the sum of squares of independent standard normal random variables according to example 9 and theorem 3. We prove this by contradiction. Assume that $(\Sigma^{-1})^{1/2}$ is not invertible, so it does not have full rank. Then, there is a column \mathbf{a}_i of $(\Sigma^{-1})^{1/2}$ that is a linear combination of

the other columns of $(\Sigma^{-1})^{1/2}$. Without loss of generality, assume $i = 1$ that is $\mathbf{a}_1 = \sum_{j=2}^n b_j \mathbf{a}_j$. But then

$$\begin{aligned}\Sigma^{-1} &= (\Sigma^{-1})^{1/2}(\Sigma^{-1})^{1/2} = \left[(\Sigma^{-1})^{1/2} \mathbf{a}_1 \ \dots \ (\Sigma^{-1})^{1/2} \mathbf{a}_n \right] \\ &= \left[\sum_{j=2}^n b_j (\Sigma^{-1})^{1/2} \mathbf{a}_j \ \dots \ (\Sigma^{-1})^{1/2} \mathbf{a}_n \right],\end{aligned}$$

and so Σ^{-1} does not have full rank. But this is a contradiction since Σ is invertible by hypothesis. ■

We have another result which will be useful later. But first we need a small lemma

LEMMA 1. *If $X_3 = X_1 + X_2$ and $X_1 \sim \chi_\mu^2$ and $X_3 \sim \chi_\nu^2$ and X_1 and X_2 are independent, then $X_2 \sim \chi_{\nu-\mu}^2$ if $\nu > \mu$.*

For proof, see chapter 6 of Devore and Berk (2012). Consider now the theorem:

THEOREM 5. *If X_1, X_2, \dots, X_n are independent with variance σ^2 , identically normal distributed, then*

$$\frac{S^2(n-1)}{\sigma^2} \sim \chi_{n-1}^2 \quad \text{and so} \quad V(\hat{\sigma}^2) = 2 \frac{\sigma^4(n-1)}{n^2}.$$

PROOF. First, inspect the following quantity:

$$\begin{aligned}\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 &= \sum_{i=1}^n \left(\frac{X_i - \bar{X} + \bar{X} - \mu}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 + n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 - 2 \left(\frac{\bar{X} - \mu}{\sigma} \right) \frac{1}{\sigma^2} \underbrace{\left[-n\bar{X} + \sum_{j=1}^n X_j \right]}_{=0} \\ &= \underbrace{\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2}_{=(n-1)S^2/\sigma^2 \text{ (eqn 2.4)}} + \left(\frac{\bar{X} - \mu}{\sigma/n^{1/2}} \right)^2.\end{aligned}\tag{2.29}$$

We see that $(n-1)S^2/\sigma^2$ is the difference of two chi square, random variables with n and 1 degrees of freedom, respectively (Devore and Berk 2012). To see this, just note that

$$\frac{\bar{X} - \mu}{\sigma/n^{1/2}} \sim N(0, 1) \quad \text{and} \quad \frac{X_i - \mu}{\sigma} \sim N(0, 1),$$

and use the defining property of chi square random variables (being the square of a standard normal random variable). Note also that $(n-1)S^2/\sigma^2$ and $(\bar{X} - \mu)^2/(\sigma/n^{1/2})^2$ are independent. To see this, note that they are uncorrelated since

$$\text{Cov}(X_i - \bar{X}, \bar{X}) = \text{Cov}(X_i, \bar{X}) - V(\bar{X}) = \frac{1}{n} \sum_{j=1}^n \text{Cov}(X_i, X_j) - \frac{\sigma^2}{n} = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$$

Since X_i are normal distributed, this means that they are also independent according to theorem 3. Hence, squared sums of such terms must be independent. That means we can apply lemma 1 using

$$X_3 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2, \quad X_2 = \frac{S^2(n-1)}{\sigma^2}, \quad X_1 = \left(\frac{\bar{X} - \mu}{\sigma/n^{1/2}} \right)^2.$$

This proves the first part of the theorem. To get the variance formula use equation (2.4), and since $V(X_2) = 2(n-1)$ according to lemma 1 and equation (2.27) we have

$$2(n-1) = V(X_2) = V\left(\frac{S^2(n-1)}{\sigma^2}\right) \stackrel{(2.4)}{=} V\left(\frac{n-1}{\sigma^2} \frac{n}{n-1} \hat{\sigma}^2\right) = \frac{n^2}{\sigma^4} V(\hat{\sigma}^2).$$

Divide through by n^2/σ^4 , and the theorem follows. ■

Since linear algebra clearly has applications to probability theory, and since quadratic forms appear in many applications, it is natural to ask what the expected value and covariance of two quadratic forms. In general, the latter is a difficult question (Mathai and Provost 1992), but in the case that the random variables are multivariate normal, it is easier. In the following, assume $\text{Tr}(A)$ denotes the trace of a square matrix A .

THEOREM 6. *If A, B are symmetric $n \times n$ -matrix, suppose $\mathbf{E} \sim N(\boldsymbol{\mu}, \Sigma)$ is multivariate normal with expectation $\boldsymbol{\mu}$ and covariance matrix Σ . Assume also that Σ has decomposition $\Sigma = QQ^\top$ for some $n \times r$ matrix Q of rank r , then*

$$\begin{aligned} \langle \mathbf{E}^\top A \mathbf{E} \rangle &= \text{Tr}(A\Sigma) + \boldsymbol{\mu}^\top A \boldsymbol{\mu} \quad \text{and} \quad \text{Cov}(\mathbf{E}^\top A \mathbf{E}, \mathbf{E}^\top B \mathbf{E}) = 2\text{Tr}(A\Sigma B\Sigma) \\ &\quad + 4\boldsymbol{\mu}^\top A\Sigma B \boldsymbol{\mu}. \end{aligned}$$

One of the most important theorems in all probability theory is the central limit theorem. I give two versions of the theorem and one example.

THEOREM 7 (Central limit theorem). *If E_1, E_2, \dots are independent, identically distributed random variables, and $\bar{E} = (1/n) \sum_{i=1}^n E_k$ denote the sample mean of the first n random variables. Then*

$$\frac{\bar{E} - \langle \bar{E} \rangle}{V(\bar{E})^{1/2}} \sim N(0, 1) \quad \text{as } n \rightarrow \infty.$$

The sequence converges in distribution to the standard normal distribution.

Commit the following thought experiment to mind:

EXAMPLE 10 († Galilean mean distribution). *According to Viviani (1717), Galileo dropped particles of distinct masses from the Leaning Tower in Pisa between the years 1589 and 1592 and discovered that their descent times did not depend on the falling masses. Conduct now a thought experiment; imagine that he was also interested in the probability distribution of the mean decent time of the particles. He allegedly recorded the decent times of a large number of particles D_1, \dots, D_n . To make the following thought experiment interesting, imagine the particles were of an ideal gas, such that one particle did not interact with any other; thereby our intuition says the observations D_1, \dots, D_n were independent. In fact, n was so large that there was a divisor of n denoted by $k \in \mathbb{N}$ such that $n/k \in \mathbb{N}$ was also a large number. He computed the sample means of the drop times*

$$\bar{D}_i = \frac{1}{k} \sum_{j=ik+1}^{(i+1)k} D_j, \quad \text{for } i = 0, 1, \dots, \frac{n}{k} - 1.$$

That means k takes the rôle of the size of each sample mean \bar{D}_i . Upon drawing a histogram counting the number of \bar{D}_i , he would have discovered that \bar{D} was normally distributed (see figure 2.2) because theorem 7 says this is true.

Imagine that he performed a second experiment. The gas was no longer an ideal gas, and according to the first definition of independence, the decent times D_i were no longer independent. Instead, they were prepared in a way such that there exists a chi-square distributed random variable $X \sim \chi_1^2$ making each decent time $D_i = X$ for all $1 \leq j \leq n$. If $Z \sim N(0, 1)$, we can calculate (using that the following integrand is symmetric and recalling that any probability distribution is normalized):

$$\begin{aligned} P(Z \in (-\infty, 0]) &\stackrel{(2.5)}{=} \int_{-\infty}^0 f(x) dx \stackrel{(2.20)}{=} \int_{-\infty}^0 \frac{e^{-(1/2)x^2}}{(2\pi)^{1/2}} dx \\ &= \frac{1}{2} \left(\int_{-\infty}^0 \frac{e^{-(1/2)x^2}}{(2\pi)^{1/2}} dx + \int_{-\infty}^0 \frac{e^{-(1/2)x^2}}{(2\pi)^{1/2}} dx \right) \\ &\stackrel{u=-x}{=} \frac{1}{2} \left(\int_{-\infty}^0 \frac{e^{-(1/2)x^2}}{(2\pi)^{1/2}} dx + \int_{\infty}^0 \frac{e^{-(1/2)(-u)^2}}{(2\pi)^{1/2}} (-1) du \right) \\ &= \frac{1}{2} \underbrace{\int_{-\infty}^{\infty} \frac{e^{-(1/2)x^2}}{(2\pi)^{1/2}} dx}_{=1} = \frac{1}{2}. \end{aligned}$$

On the other hand, since $\bar{D} = (1/n) \sum_{i=1}^n D_i = (1/n) \sum_{i=1}^n X = (1/n)nX = X$,

$$P\left(\frac{\bar{D} - \langle \bar{D} \rangle}{V(\bar{D})} \in (-\infty, 0]\right) \stackrel{(2.27)}{=} P\left(\frac{X - 1}{2} \in (-\infty, 0]\right) = P\left(\frac{X - 1}{2} \leq 0\right). \quad (2.30)$$

The probability of this equals the probability that $X \leq 1$ (\dagger). To see this, just solve the inequality. So

$$\begin{aligned} P\left(\frac{\overline{D} - \langle \overline{D} \rangle}{V(\overline{D})} \in (-\infty, 0]\right) &\stackrel{(\dagger)(2.30)}{=} P(X \leq 1) \stackrel{(2.26)}{=} \int_{-\infty}^0 0 \, dx + \int_0^1 \frac{x^{1/2-1} e^{-x/2}}{2^{1/2} \Gamma(1/2)} dx \\ &= \int_0^1 \frac{x^{-1/2} e^{-x/2}}{(2\pi)^{1/2}} dx \geq \int_0^1 \frac{x^{-1/2}}{(2\pi)^{1/2}} \left(1 - \frac{x}{2}\right) dx \\ &= \frac{1}{2} \underbrace{\left(\frac{100}{18\pi}\right)^{1/2}}_{>1} > \frac{1}{2}. \end{aligned}$$

We may recall that $\Gamma(1/2) = \pi^{1/2}$ and used Lagrange's remainder theorem (Lindström 2006) in the 3rd to last step³. This proves that the distribution that Galileo found in the second experiment could not possibly have been the normal distribution. In fact, figure 2.2 reveals what he found. ►

Remember this example. It shows why the central limit theorem is one of the most important theorems of probability theory and also that the conclusion can be false once the hypothesis is false. As you can imagine, it is very useful whilst working on asymptotics (Mathai and Provost 1992). There exists many extensions of the central limit theorem where some requirements are relaxed. We will be interested in one such extension. But before we can introduce this theorem, we need to cover more material. The extended central limit theorem is presented in a seminal paper by Ibragimov (1975) and is given in section 2.4 where time series are presented and discussed.

2.3 Applications of probability theory

2.3.1 Fisherian inference

You may have heard of Bayesian statistics. But a significantly more famous school of statistics is called Fisherian- or frequentist statistics. Ironically, in my experience, these names are rarely used. Instead, we just say *inference*, or statistical inference (Efron 1986). As this is the most common method of statistical inference, you've undoubtedly heard of many of the applications, which we will cover.

But why do we want to learn about Fisherian statistics? First of all, this allows us to understand the results of the present thesis. And in the physical

³The function $\exp : \mathbb{R} \rightarrow \mathbb{R}$ is continuously differentiable at all orders on $[0, 1]$, so according to Lagrange's remainder theorem; for each $x \in [0, 1]$ there is some number $c(x) \in (0, x) \subseteq (0, 1)$ such that $\exp(-x/2) = 1 - x/2 + c(x)^2/8 > 1 - x/2$.

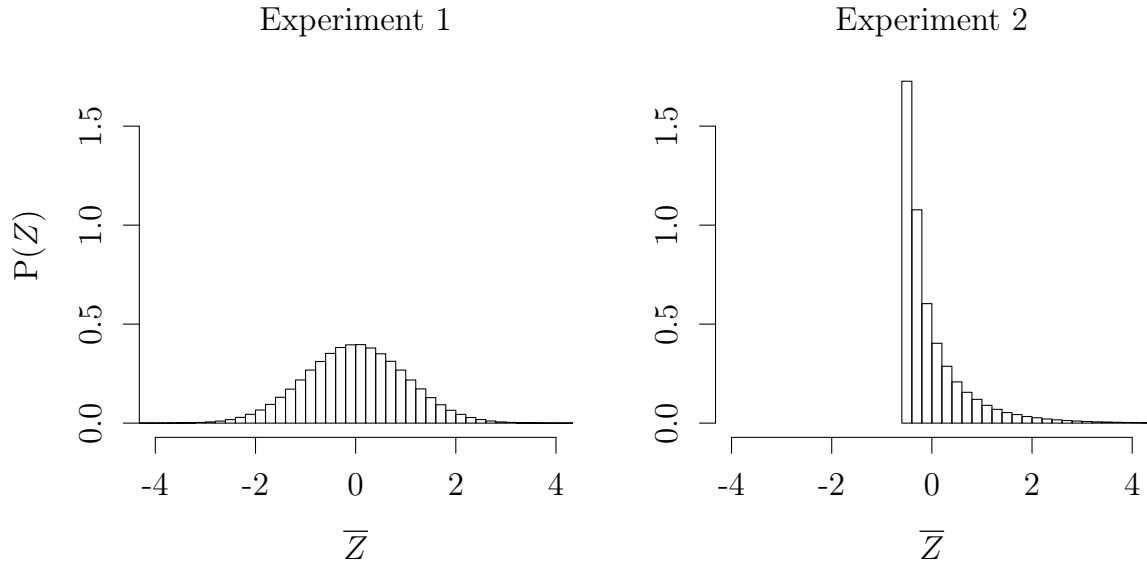


Figure 2.2: To the left: The empirical distribution Galileo found in the thought experiment for the standardized mean $\bar{Z} = (\bar{D} - \langle \bar{D} \rangle) / V(\bar{D})^{1/2}$. In the first experiment, the empirical distribution is similar to the standard normal distribution. This is because the conditions of theorem 7 was satisfied. To the right: The second experiment. Here, the independence of the observations X_i, X_j were violated since $\text{Cov}(X_i, X_j) = \text{Cov}(U, U) = V(U) = 2$. Therefore, the conclusion of the theorem was false, and the asymptotic distribution that Galileo found is dissimilar to the normal distribution.

sciences, the applications are widespread. Suppose you are working on a theory of physics and you want to find out if there is physical evidence of support. You conduct an experiment, and the theory predicts what you will find. Statistical inference has tools which can tell you the probability that your theory is wrong Devore and Berk (2012). If the probability that your theory is wrong is sufficiently small, then other physicists are going to conclude that your theory is right.

In practice, almost all physical theories contain constants that we want to estimate from experimental data. The art of estimating is called **point estimation**.

EXAMPLE 11 (†). *We are conducting an experiment to determine the expected position μ of a particle along the x -axis of a coordinate system. We intend to make n observations indexed by $1 \leq i \leq n$. Suppose the associated state function $|\psi|^2$ is Gaussian. This means the measured position, X_i of the particle is $X_i \sim N(\mu, \sigma^2)$ according to example 9. According to the discussion in the previous chapter, the sample mean and sample variance defined in equations (2.2) and (2.3) produce educated guesses at the values μ and σ^2 .* ►

You may be curious how we came up with the clever definitions of equations (2.2) and (2.3). Is there a recipe that can be used to find the best estimator for any such state function and probability distribution? The answer is complicated, but if the observations X_1, \dots, X_n are independent and identically distributed, there is a theory of Fisherian statistics that provides some clear answers (Devore and Berk 2012).

Suppose F is any probability distribution and $X_1, \dots, X_n \sim F(\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$ which is to be estimated. Suppose we let $f_X(x; \boldsymbol{\theta})$ denote the marginal pdf of F and $f(\mathbf{x}; \boldsymbol{\theta})$ the joint pdf. Assume also that X_1, \dots, X_n are independent and identically distributed. We define the **likelihood** function $L(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta})$. The likelihood is somewhat analogous to a probability density function in the sense that value $\hat{\boldsymbol{\theta}}$, which maximizes the L , is the value of $\boldsymbol{\theta}$ that we believe is the best estimate for $\boldsymbol{\theta}$. We believe it is the best estimator because there were development in the 20th century that proved that $\hat{\boldsymbol{\theta}}$ has some of the best properties of any possible estimator. The function $\hat{\boldsymbol{\theta}}$ is said to be a **maximum likelihood estimator**. Statistical methods, which use maximum likelihood estimators, are typically called **maximum likelihood** methods (DeGroot and Schervish 2012). If f is continuously differentiable, we call the vector $\mathbf{s}(\boldsymbol{\theta}) = \nabla \log L(\boldsymbol{\theta})$ the **score function**. As you probably predicted, we find the maximum of L by solving $\mathbf{0} = \mathbf{s}$ for $\hat{\boldsymbol{\theta}}$. We make this definition because any maximum of L is a maximum of $\log L$. To see that this is true, suppose $\hat{\boldsymbol{\theta}}$ is a maximum of L . Then $L(\hat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ in the domain of L by definition. But since the logarithm is monotonous on \mathbb{R} , $\log L(\hat{\boldsymbol{\theta}}) \geq \log L(\boldsymbol{\theta})$.

Maximum likelihood estimators have appealing properties (Devore and Berk 2012), which are responsible for the success of Fisherian inference:

- If $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimator, it is asymptotically normal distributed under mild conditions⁴.
- $\lim_{n \rightarrow \infty} \langle \hat{\boldsymbol{\theta}} \rangle = \boldsymbol{\theta}$
- $\lim_{n \rightarrow \infty} V(\hat{\boldsymbol{\theta}}) = I^{-1}$ where I is the matrix with elements consisting of the expected values of all partial derivatives of $-\mathbf{s}$.
- The estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal distributed.
- The estimator is an asymptotic **minimal variance unbiased estimator**, acronymed MVUE (Devore and Berk 2012). That means if $\hat{\boldsymbol{\phi}}$ is an unbiased estimator of $\boldsymbol{\theta}$, $V(\hat{\boldsymbol{\phi}}_i) \geq V(\hat{\boldsymbol{\theta}}_i)$ for all $1 \leq i \leq n$ when $n \rightarrow \infty$.

The normality follows from the theorem 7 (Sweeting 1980). In the previous section we sloppily said that maximum likelihood estimators are the best estimators of all. What we meant is to say that they are MVUE (Devore and Berk 2012). Consider the next few examples to understand the applications.

EXAMPLE 12. Assume $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ independent identically distributed. According to example 9

$$\begin{aligned} L(\mu, \sigma) &= f(x_1, \dots, x_n; \mu, \sigma) = (2\pi)^{-n/2} |\sigma^2 I_n|^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{e} - \boldsymbol{\mu})^\top (\sigma^2 I_n)^{-1} (\mathbf{e} - \boldsymbol{\mu}) \right) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp \left(\sum_{i=1}^n -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) = \prod_{i=1}^n \frac{1}{(2\pi)^{1/2} \sigma} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \end{aligned}$$

That means the log likelihood $\log L$ is given by

$$\log L(\mu, \sigma^2) = \sum_{i=1}^n -\frac{1}{2} \log 2\pi - \log \sigma - \frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}.$$

Let's find the maximum likelihood estimators for μ and σ^2 by solving $\mathbf{s} = \mathbf{0}$. Since f is continuously differentiable, all partial derivatives exists and are continuous, so

$$\begin{aligned} \mathbf{0} = \mathbf{s}(\mu, \sigma^2) &= \nabla \log L(\mu, \sigma) = \left(\frac{\partial \log L}{\partial \mu}, \frac{\partial \log L}{\partial \sigma^2} \right) \\ &= \left(\frac{n}{\sigma^2} (\bar{X} - \mu), \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^2} \right). \end{aligned}$$

⁴Suppose $D_{\mathbf{s}}$ is the Jacobian matrix of \mathbf{s} . Sufficient requirements are that \mathbf{s} is differentiable and $D_{\mathbf{s}}$ is invertible. The expected value of each component of $(D_{\mathbf{s}})_{ij}$ must be continuous and $\hat{\boldsymbol{\theta}}$ must be an interior point in the domain of L . Moreover $\|\mathbf{s}\|$ and $\|(D_{\mathbf{s}})_i\|$ must be integrable and $\langle \|(D_{\mathbf{s}})_i\| \rangle < \infty$ for all $1 \leq i \leq n$.

Solving for μ, σ gives two stationary points of the log likelihood. These are indeed the maximum of the likelihood, as you can check. And so by solving the equation for μ, σ we obtain the maximum of $\log L$:

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2.$$

You've seen these estimators before in section 2.2.1, they are the sample mean and sample variance! Moreover we know that for large n they are normal distributed, unbiased and have smaller variance than any other estimator from Fisherian inference. ►

EXAMPLE 13 (Confidence intervals). Assume θ is some scalar parameter which we wish to estimate. A **100(1 - α) percent confidence interval for θ** is a subset $[a, b] \subset \mathbb{R}$ such that the probability that the true value of θ is in $[a, b]$ is 100(1 - α)% (Devore and Berk 2012). According to the above, $\hat{\theta} \approx N(\theta, V(\hat{\theta}))$. If $z_{\alpha/2}$ denotes the $\alpha/2$ quantile in the standard normal distribution, by example 9,

$$1 - \alpha \approx P \left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{V(\hat{\theta})^{1/2}} \leq z_{\alpha/2} \right)$$

only if $1 - \alpha \approx P \left(\hat{\theta} + z_{\alpha/2} V(\hat{\theta})^{1/2} \geq \theta \geq \hat{\theta} - z_{\alpha/2} V(\hat{\theta})^{1/2} \right),$

where we solved the inequality for θ . Since $z_{\alpha/2}$ is known for all $\alpha \in [0, 1]$, this shows that if we can estimate $V(\hat{\theta})$, then letting

$$a = \hat{\theta} - z_{\alpha/2} V(\hat{\theta})^{1/2} \quad \text{and} \quad b = \hat{\theta} + z_{\alpha/2} V(\hat{\theta})^{1/2}$$

solves the problem. This is appealing since you can now present the estimate for the position of the particle in example 11 as well as uncertainty limits of that estimate for significance levels α ! If on the other hand, θ is a vector, it is possible to construct confidence regions of \mathbb{R}^n using for example theorem 4 in analogy to what will be done below in combination with numerics. ►

Experts see that the above example implicitly answers some questions of the type: "Given some hypothesis A and appropriate data, how can I determine if A is false?". If you do not yet see this link, we shall make it explicit at the end of this section. But you do understand that this is interesting, because if you want to find out if another hypothesis B is true, in some sense we can choose $B = \neg A$, the negation of A . Statisticians call this area of statistics **hypothesis testing** (DeGroot and Schervish 2012). But before we formulate some precise results, let's make precise what we mean by a "hypothesis" and "appropriate data".

Hypothesis testing works backward relative to the way people often think. Often we will believe that some hypothesis H_a is true. If we find no contradictions, many people will believe it is true, but if we find a contradiction we conclude that not H_a must be true. Hypothesis testing works differently. When doing hypothesis testing we opt a default viewpoint H_0 , which is the safe view that we learnt nothing new (we will make this precise in a moment). Only if the observations contains sufficient burden of proof, we will conclude that not H_0 is true. We say that H_0 is called the **null hypothesis** denoted by H_0 if it is a statement about whether a mathematical quantity can be found in a set. It is best to pick H_0 such that it has no burden of proof (DeGroot and Schervish 2012). We say that an **alternative hypothesis**, denoted by H_a , is the negation of H_0 and has the burden of proof. To make it more clear what we mean, we make an example

EXAMPLE 14 (†). *Gravitational lensing is a demonstration of general relativity. We will assume that prior to Eddington and Dyson May 1919 experiment, measuring the deflection of light by the sun, the general consensus was that light was not bent by the presence of mass. However in the abstract of Dyson, Eddington, and Davidson (1920): "IX. A determination of the deflection of light by the Sun's gravitational field, from observations made at the total eclipse of 29 May 1919" we read that the authors were interested in the following hypothesis:*

1. *The path is uninfluenced by gravitation.*
2. *The energy or mass of light is subject to gravitation in the same way as ordinary matter. If the law of gravitation is strictly the Newtonian law, this leads to an apparent displacement of a star close to the sun's limb amounting to $0''.87$ outwards.*
3. *The course of a ray of light is in accordance with Einstein's generalised relativity theory. This leads to an apparent displacement of a star at the limb amounting to $1''.75$ outwards.*

If Eddington and Dyson had used modern statistical inference, one way to analyze the data correctly could have been done in the following three steps:

- (a) *First propose that (1) is the null hypothesis. This makes sense since we assumed that this was the general consensus before 1919, therefore there is no burden of proof to assume true. A sensible choice of alternative hypothesis according to the definition above would be that "Light is deflected by gravitation". This works because it implies that the null hypothesis is false and since the consensus was that light is not deflected by gravity, there is profound burden of proof.*

- (b) After discovering that (1) was false, it would have been natural to propose two new hypothesis: To take (2) as the new null hypothesis since Newtonian gravity was the classical theory in this context. Therefore, there was no burden of proof to assume it was true given that light is deflected by gravity. The alternative hypothesis, being the negation of the null hypothesis is that "Newtonian gravity does not explain the apparent displacement". This hypothesis bears the burden of proof since it is non-classical given that light is deflected by gravity.
- (c) After discovering that (1) and (2) are false, it remains to test whether (3) could be discarded. We propose (3) as the final null hypothesis. This is because given that light is deflected by gravity, but the gravity is not Newtonian, the natural candidate for the truth was general relativity, hence it does not bear the burden of proof. We take the negation of (3) to be the alternative hypothesis.

Statements of the form "A does not bear the burden of proof", may surprise you. Perhaps because you even disagree! In which case, it is possible to make different choice of null and alternative hypothesis. The implications of making a poor choice of a hypothesis, is that you will make it more difficult to quantify your uncertainty later on. Hence, you think about your audience when making choices between hypotheses. If you make choice of H_0 such that your audience agree that H_0 does not carry the burden of proof, then the audience will be convinced later when you present the conclusion of the test along with the uncertainty. ►

Next we discuss what we mean by 'appropriate data'. Suppose we make observations X_1, \dots, X_n . If there is some $\mathbb{R}^n \rightarrow \mathbb{R}$ -function such that $f(X_1, \dots, X_n)$ has known distribution when H_0 is true, then we say that $f(X_1, \dots, X_n)$ is a **test statistic** (Devore and Berk 2012) .

EXAMPLE 15 (†). In the previous example, example 14, we proposed that the first two hypothesis (a) were

$$\begin{aligned} H_0 &: \text{The path is uninfluenced by gravitation} \\ H_a &: \text{Light is deflected by gravitation} \end{aligned}$$

Suppose \bar{X} denotes the mean deflection of light from measurements X_1, \dots, X_n of n stars. According to our intuition, each measurement is independent because the light from one star does not influence the light from another star to cause any appreciable effect relevant to the measuring apparatus. Moreover, if the null hypothesis is true, there is no deflection of light, so the variation in measurements are only due to error of measurement relative to zero degrees deflection. Since the same measuring apparatus was used to measure the deflection angles, they are

identically distributed. Therefore, the conditions to use theorem 7 are true; that means

$$Z_1 = \frac{\bar{X}}{V(\bar{X})^{1/2}} \stackrel{H_0}{\approx} N(0, 1)$$

is asymptotic standard normal distributed when H_0 is true. Therefore, we say that \bar{X} is a test statistic for the hypothesis H_0 since 7 says that the distribution is known. Note that although this is an approximation, according to Devore and Berk (2012): If the number of observations (here number of stars) is larger than about 40, statisticians would say that the approximation was satisfactory.

In example 11, where we considered the point estimation of the position and variance of position of the particle. Suppose our theory tells us that we should expect the standard deviation to be σ_0 . So we propose the hypothesis

$$H_0 : \sigma = \sigma_0 \quad \text{versus} \quad H_a : \sigma \neq \sigma_0.$$

The observations X_1, \dots, X_n in example 11 are normal distributed. So according to example 12, the estimator $\hat{\sigma}^2$ is a maximum likelihood estimator, hence asymptotic unbiased, normal distributed according to the properties of maximum likelihood estimators on page 31. So if the variance $V(\hat{\sigma}^2)$ can be determined, then

$$Z_2 = \frac{\hat{\sigma}^2 - \sigma_0}{V(\hat{\sigma}^2)^{1/2}} \stackrel{H_0}{\approx} N(0, 1)$$

is an approximate test statistic. It is an easy exercise for readers to prove this by computing the expected value and variance of Z . ►

We are now ready to do hypothesis testing. If there is a test statistic, T available for the hypothesis H_0 , then according to the definition, its pdf, f is known whenever H_0 is true. Since we know f under H_0 , we can read off from f which value we are likely to observe. If we observe a value of T which is sufficiently unlikely, we conclude that H_0 is false. We say sufficiently, because we decide on a confidence level which is so unlikely that it convinces us that if it is observed, H_0 is false. If you struggle to understand the logic, then this is just the contrapositive (Lindstrøm 2017) of the following statement

$$H_0 \text{ is true} \implies \text{We observe a probable value of } T.$$

If you've taken an introduction to propositional logic you conclude that the contrapositive is:

$$H_0 \text{ is not true} \iff \text{We observe an improbable value of } T.$$

But how improbably is sufficiently improbable before we conclude that H_0 is not true? In ordinary inference, a sufficiently improbable value is the 95%-percentile.

In particle physics on the other hand, the 99.99997%-percentile is chosen. If Φ^{-1} is the quantile function, then $\Phi^{-1}(0.9999997) = 5$. Since a standard normal random variable has standard deviation of $\sigma = 1$, this means $\Phi^{-1}(0.9999997) = 5 \cdot 1 = 5\sigma$. This is the reason physicists talk of 5σ significance. As we shall see later, this is jargon.

EXAMPLE 16 (†). *Let's see what Eddington and Dyson would have concluded about General Relativity had they used modern statistical inference in 1920. Example 14 has three sets of hypothesis to be tested. According to Dyson, Eddington, and Davidson (1920), $\bar{X} = 1''.98$ and the standard error is $V(\bar{X})^{1/2} = 0''.12$ were the estimates .*

- (a) *Here the test statistic is Z_1 as given in example 15. Hence the observed value is $Z_1 = \bar{X}/V(\bar{X}) = 1''.98/0''.12 = 16.5$. So Eddington would have concluded with 16.5σ significance that light is indeed deflected by gravity!*

In the following, assume that μ_0 is the predicted deflection angle according to (b) Newtonian gravity and (c) General relativity. According to Dyson, Eddington, and Davidson (1920), the value of μ_0 is thus (b) $\mu_0 = 0''.87$ and (c) $\mu_0 = 1''.75$. We will follow similar argumentation as in example 15 and use the following test statistic:

$$Z_3 = \frac{\bar{X} - \mu_0}{V(\bar{X})} \stackrel{H_0}{\approx} N(0, 1).$$

- (b) *According to the values supplied by Dyson, Eddington, and Davidson (1920), the value under the new null hypothesis is a $Z_3 = 9.25\sigma$ result.*
- (c) *According to the values supplied by Dyson, Eddington, and Davidson (1920), the value under the final null hypothesis is a $Z_3 = 1.92\sigma$ result.*

So using modern statistical inference, Eddington would have concluded that he could disregard that "light is not deflected by gravity" and "light follows newtonian gravity" on 5σ significance. However, he would not have been able to discard General relativity, even at 2σ ! ►

What does the significance measure? I briefly discussed that. I will also rectify the confusion induced by the informal jargon as pointed out in example 16. The **significance** is the probability of discarding H_0 given that H_0 is true (DeGroot and Schervish 2012). This is the reason why it was important to choose H_0 and H_a in the particular way we did. It would have been irresponsible to assume that H_0 had the burden of proof, for then we could end up accepting H_0 even if there was little proof of support (DeGroot and Schervish 2012). As such, the significance does not measure the probability that H_a is true, it only tells us the probability of making a mistake when discarding H_0 (Devore and Berk 2012). Physicists may say, however that there is 5σ is the significance of B . What the

physicist mean is that since the cumulative function is strictly monotonous, there is a bijection between «the values of the test statistic T », and «the values α such that $\leq \alpha$ » in the case T is standard normal. So in light of example 15, the physicist has already applied theorem 7.

As readers now understand, the significance is a number that is determined prior to the experiment. It is irresponsible to choose significance levels after the experiment had been conducted (Devore and Berk 2012): If chosen after, one might have used a suitable significance level to claim discovery whenever one wanted. Therefore, we decide the significance level, on which everyone agrees, prior to starting the experiment. The second best one can do, is to evaluate the *p-value*. The *p-value* is a measure that tells us the smallest significance level α which would still imply that we could conclude that H_a is true (DeGroot and Schervish 2012). It is the standard measure to express your certainty in your results. Small *p-values* of ($p \ll 0.05$) indicate that it is relatively safe to make a decision based on the inference (DeGroot and Schervish 2012). This is because it says that the probability of discarding H_0 when H_0 is true, is small.

EXAMPLE 17. *Looking back at Eddingtons experiment, we found 16.5σ "significance" for the hypothesis*

H_0 : *The path is uninfluenced by gravitation*

H_a : *Light is deflected by gravitation.*

*I write "significance" to emphasize that this is jargon. The correct way, as shown above, would be to say that the significance $\alpha = 1 - 0.99999971335$ implies that $\Phi^{-1}(0.99999971335) = 5\sigma$ in the standard normal distribution. The definition of the *p-value* is the smallest significance that will still yield discovery, so it is $p = 1 - \phi^{-1}(16.5) < 10^{-16}$. ►*

Before wrapping up this subsection, I will return to the remark made in the beginning. Hypothesis testing is often reduced to seeing if the value of some parameter θ lies in some interval. The boundaries of the interval are the $1 - \alpha$ quantiles in the distribution of the test statistic whenever H_0 is true. But recall example 13. This sounds like the definition of a confidence interval. In fact, computing a confidence interval is performing a hypothesis test, and hypothesis testing is often the act of checking if the observed value of the test statistic fall in the confidence interval of the test statistic as computed if H_0 was true (Agresti 2015).

2.3.2 Bayesian statistics

After introducing Fisherian inference, it is easy to motivate Bayesian inference, and to define the word statistic. According to Devore and Berk (2012), a *statistic*

is any quantity whose value can be calculated from sample data. In the previous chapter, you saw that Fisherian statistics was concerned with guessing the value of θ by calculating the statistic $\hat{\theta}$. My understanding is that we refer to the branch of probability theory that makes decisions based on data. Bayesian statistics is the branch of statistics where our belief about the statistic is not deterministic in the sense that there is a true, unknown value of θ that we want to guess. Instead in **Bayesian statistics** we believe that the statistic θ is described by degree of certainty and uncertainty (Gelman et al. 2014). What do we mean by this? In the previous chapter, we said that the likelihood function was in some sense analogous to a probability distribution for θ . In Bayesian statistics, however, we express our belief about θ in terms of a probability distribution. The distribution describes the probability that the value of θ lies in a volume of \mathbb{R}^n . The advantage of this, as you will see, is that we can combine our knowledge of θ given by the observations X_1, \dots, X_n with other types of knowledge we may have about θ (Gelman et al. 2014).

EXAMPLE 18. (†) Suppose you measured the position of the particle from example 11 at the position $x_1 = \mu_0$. According to quantum mechanics (and in particular the Schrödinger equation itself), wave function collapse ensures that immediate subsequent measurements, x_2, \dots, x_n of the position of the particle will be near μ_0 . A physicist conducting such an experiment knows that subsequent measurements will be near μ_0 , but Fisherian statistics does not know quantum physics, and so the estimator \bar{X} (which is the maximum likelihood estimator according to example 12) does not take this knowledge into account. Bayesian statistics, on the other hand, takes prior information that is known a priori into account, and therefore the Bayes estimate will be much better. This is especially true if n is a small number, as the reader later will see. ►

THEOREM 8 (Bayes theorem). Suppose A_1, A_2 are subsets of \mathbb{R} and E_1, E_2 are random variables, then

$$P(E_1 \in A_1 | E_2 \in A_2) = \frac{P(E_2 \in A_2 | E_1 \in A_1)P(E_1 \in A_1)}{P(E_2 \in A_2)}. \quad (2.31)$$

Moreover if $p(e_1|e_2)$ is the conditional probability distribution of E_1 and E_2 and $p(e_i)$ is the marginal probability distribution of E_i for $i \in \{1, 2\}$, then

$$p(e_1|e_2) = \frac{p(e_2|e_1)p(e_1)}{p(e_2)}. \quad (2.32)$$

The proof is easy. To get the first formula, just use the definition of conditional probability twice:

$$\begin{aligned} P(E_1 \in A_1 | E_2 \in A_2)P(E_2 \in A_2) &\stackrel{(2.24)}{=} P(E_1 \in A_1, E_2 \in A_2) = P(E_2 \in A_2, E_1 \in A_1) \\ &\stackrel{(2.24)}{=} P(E_2 \in A_2 | E_1 \in A_1)P(E_1 \in A_1). \end{aligned}$$

Now divide by $P(E_2 \in A_2)$ on each side of the equality. In the second step we used that the probability of any number of events is symmetric with respect to the probability measure P . Your intuition should confirm that this is true. Think for example of the toss of two dice. The probability that the eyes of the first die happen to be in A_1 and the second in A_2 is the same as the probability that the second die is in A_2 and the first is in A_1 . We have not shown this explicitly. If you want to see the proof, this follows from the axioms of measure theory (see for example chapter 7 of (McDonald and Weiss 2012)). Proof of the second equality is proven similarly by using equation (2.25).

Although theorem 8 looks innocent, its implications are profound (Gelman et al. 2014). For example, often it is an effective way to translate a probability we do not necessarily know how to calculate, to one that we can deal with by partitioning the set of outcomes, on which we have more control. This becomes even more interesting when we talk about time series later. For now, consider instead: Bayes theorem gives rise to all of Bayesian statistics (Gelman et al. 2014).

Now, is a good time to explain how a Bayesian thinks about the world and make some definition. A Bayesian thinks of the parameter θ as a random variable. As you understand, this contrasts the Fisherian philosophy, where we were interested in testing what the true value of θ was. Suppose we have observed \mathbf{x} that are realizations of X_1, \dots, X_n . That means using Bayes theorem above

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}.$$

The distribution $p(\theta)$ is called **prior distribution**. When doing Bayesian statistics, we let 'the prior' contain all the information known about θ prior observing \mathbf{x} . This is the kind of information an expert would produce, since we assume the expert has all prior information about θ . The probability distribution $p(\mathbf{x}|\theta)$ is the pdf of \mathbf{x} , which is known whenever the distribution of X_1, \dots, X_n is known. A Bayesian will call this the **likelihood**. In fact, it is identical to the likelihood from Fisherian statistics. The distribution $p(\mathbf{x})$ is the marginal distribution of \mathbf{x} that is either known or obtained by integrating the numerator with respect to θ to ensure normalization of $p(\theta|\mathbf{x})$. We call $p(\theta|\mathbf{x})$ the **posterior distribution** and contains the information when combining the prior information with the estimate by observing \mathbf{x} (Gelman et al. 2014).

EXAMPLE 19. (†) *In example 18 we explained that Fisherian inference doesn't know quantum physics, but we promised that, in some sense, Bayesian statistics would be able to make better estimates, since the physicist knew quantum mechanics, thus be able to incorporate his prior information in the analysis. A Bayesian physicist would suppose that the prior was constructed from the Schrödinger equation and containing the initial information which was obtained by observing the*

particle at μ_0 . Subsequent measurements of the position of the particle will yield much better estimates.

Since the posterior distribution contains information from the likelihood and the prior, statisticians will think of a uniform prior (containing no information) as an equivalent treatment to Fisherian statistics (Gelman et al. 2014). To experts it should be clear, that this is by no means a rigorous assertion, but in applications it works well. ►

EXAMPLE 20. (†) Suppose two physicists are measuring the position X of a particle in a 1-dimensional square well of length a . So any measurement $X \in [0, a]$. At time $t = 0$ the particle is measured at $\mu = a/2$. We assume the particle is not moving much. The physicists want to make the best possible estimate for the expected position after $t = 0$.

Lemma. (†) If a particle is measured at μ such that $\Psi(0, x) = \delta(x - \mu)$ in an infinite square well at $t = 0$, then $\langle X \rangle = \mu$ for all $t \geq 0$.

Proof. See appendix.

Imagine now that the physicists are trying to estimate the expected position based on one observation at $t > 0$ without knowledge of the wave function. One of the physicists is a Bayesian, the other a Fisherian frequentist. The Bayesian will be much more able to incorporate her/his knowledge of wave function collapse to make better estimate than the frequentist. Let X denote the measurement at $t > 0$. Since the physicists do not have the wave function by hypothesis, they make no assumption about the particles whereabouts a priori, and say that the probability of observing the particle at X given μ is uniform distributed:

$$p(X|\mu) = \frac{1}{a} \quad \text{for } x \in [0, a] \quad \text{only if} \quad 0 \stackrel{!}{=} \frac{\partial}{\partial \mu} l(\mu) = \frac{\partial}{\partial \mu} \log \left(\frac{1}{a} \right) \\ = -\frac{\partial}{\partial \mu} \log(2\mu) = -\frac{1}{\mu}.$$

The likelihood does not have a maximum in \mathbb{R} , hence the frequentist cannot use maximum likelihood estimation. But she/he can use Chebychev's inequality (see equation 1.1). Therefore, the Fisherian frequentist takes the mean $\bar{X} = (1/1)X = X$ as her/his estimate of the position. The Bayesian uses the posterior mean, $\hat{\mu} = \langle \mu|X \rangle$, as her/his estimator. Unlike the frequentist, the Bayesian is able to encode her/his prior knowledge that the particle was observed at μ , and proposes a Gaussian prior centered at μ :

$$p(m) = Ae^{-10^2(m-\mu)^2} \quad \text{such that} \quad A = \int_0^a e^{-10^2(m-\mu)^2} dm \quad \text{for } x \in [0, a].$$

Consequently, the Bayesian estimator is obtained by integration using the definition of the expected value by

$$\begin{aligned}\langle \mu | X \rangle &= \int_0^a m \frac{p(m)p(X|m)}{p(X)} dm = \frac{1}{p(X)} \int_0^a mp(m) \frac{1}{a} dm = \frac{1}{ap(X)} \int_0^a mp(m) dm; \\ p(X) &= \int_0^a p(m)p(X|m) dm = \frac{1}{a} \int_0^a p(m) dm.\end{aligned}$$

Using numerical integration it turns out that the Bayesian obtained the exact answer by using the information contained in the prior⁵. See table 2.1 for a comparison of the consequences of the observation on the final estimate. The

Table 2.1: A frequentist and Bayesian physicist use inference to estimate the expected position of a particle based on exactly one observation. According to the theorem above, the exact answer is μ , which is also obtained by the Bayesian. The Fisherian frequentist use the mean, a common frequentist estimator to estimate the position. The reason the Bayesian is able to estimate correctly, is that she/he is able to encode her/his knowledge of wave function collapse in the prior distribution. The estimates are given for a variety of values of X , which shows that the frequentist estimate is not only wrong, but essentially meaningless when the value of X is far from μ . To purists, examples such as these are the pinnacle of Bayesian statistics because it shows it's still possible to do estimation when there are few observations available; and the frequentist fails.

Measured X	Exact answer $\langle X \rangle$	Frequentist answer \bar{X}	Bayesian answer $\hat{\mu}$
$(4/5)^1 \mu$	μ	$(4/5)^1 \mu$	μ
$(4/5)^2 \mu$	μ	$(4/5)^2 \mu$	μ
$(4/5)^4 \mu$	μ	$(4/5)^4 \mu$	μ
$(4/5)^8 \mu$	μ	$(4/5)^8 \mu$	μ

example shows what one would expect. Although neither of the physicists knew that the true expected value was μ , Bayesian inference allows users to combine different kinds of information, thus obtaining much better estimates when the amount of data are small. ►

We explained that Fisherian inference had two components. They were hypothesis testing and point estimation. In fact, Fisherian inference is comprised of many more topics. But the topics we introduced were some of the most theoretical economical methods I could think of, moreover they were sufficient to understand the results of this thesis. Bayesian statistics has two braches which are analogous to point estimation and hypothesis testing. But this time the theory is more

⁵It is possible to obtain this analytically by first showing that the intregrand is a symmetric function about the point μ .

elegant than in Fisherian inference, and we essentially use the same procedure to address both branches. It turns out that since we require the theory to be more elegant than Fisherian inference, the mathematics is less appealing, and we frequently resort to numerics.

Consider first Bayesian decision theory on the set $A \neq \emptyset$. Suppose we want to do inference on the parameter $\theta \in T$ based on the observation $y \in Y \neq \emptyset$. Define the set of **decision functions** \hat{A} to be the set of $Y \rightarrow A$ functions. Clearly, since Y, A are nonempty, \hat{A} is nonempty. We say that $L : T \times \hat{A}$ is a **loss function** if it is real-measurable with finite conditional expectation with respect to y given θ . The **risk function**, R is the expected loss given the action $\hat{a} \in \hat{A}$ and $\theta \in T$ if it has finite expected value for all $\hat{a} \in \hat{A}$ with respect to the measure of θ and given by:

$$R(\hat{a}, \theta) = \langle L(\theta, \hat{a}) | \theta \rangle = \int_{\text{Supp } p} L(\theta, \hat{a}) p(y | \theta) dy.$$

So far there is nothing Bayesian in the treatment. But since the risk function is the **Bayes risk**, is any bounded function $\text{BR} : \hat{A} \rightarrow \mathbb{R}$ and given by

$$\text{BR}(\hat{a}) = \langle R(\hat{a}, \theta) \rangle = \int_{\text{Supp } p} R(\hat{a}, \theta) p(\theta) d\theta.$$

The **minimum Bayes risk** is the number $\text{MBR} = \inf_{\hat{a} \in \hat{A}} \text{BR}(\hat{a})$. This number exists since it is the infimum of a nonempty set of real numbers that are bounded below. The **Bayes solution** is the set

$$B = \{\hat{a} \in \hat{A} \mid \text{BR}(\hat{a}) = \text{MBR}\},$$

it is not clear that the cardinality of B is finite, but when talking of Bayesian inference, it is common to make constructions such that B has a parameterization or is finite.

Similarly we conduct Bayesian inference on an estimator $\hat{\theta} \in \hat{T}$. Loss function $L : T \times \hat{T} \rightarrow \mathbb{R}$. Risk function is the expected loss given the estimator $\hat{\theta} \in \hat{T}$ and $\theta \in T$. The Bayes risk is now a $\hat{\theta} \rightarrow \mathbb{R}$ -function and MBR is taken over \hat{T} instead of $\hat{\theta}$ with the same properties as before.

2.3.3 Shrinkage estimation

Suppose that $\hat{\theta}$ is any estimator of θ . In the chapter on Fisherian inference we briefly said that maximum likelihood estimators were amongst the best estimators of all. To make such a statement, we would like an ordering relation (Munkres 2000) on the set of all estimators of θ . We do not have such an ordering, but we do

have a function from the set of estimators of $\boldsymbol{\theta}$ to \mathbb{R} which can be used to make an informal ordering. The function is called the **mean squared error**, MSE given by $\text{MSE}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \langle \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \rangle$ (Devore and Berk 2012) with the property that

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) &= \langle \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \rangle = \langle \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{\theta}} \rangle - \langle \hat{\boldsymbol{\theta}} \rangle^\top \langle \hat{\boldsymbol{\theta}} \rangle + \langle \hat{\boldsymbol{\theta}} \rangle^\top \langle \hat{\boldsymbol{\theta}} \rangle - 2 \langle \hat{\boldsymbol{\theta}} \rangle^\top \boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\theta} \\ &= \langle \hat{\boldsymbol{\theta}}^\top \hat{\boldsymbol{\theta}} \rangle - \langle \hat{\boldsymbol{\theta}} \rangle^\top \langle \hat{\boldsymbol{\theta}} \rangle + \|\text{Bias}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\|^2 \\ &= \langle \text{Tr}(\hat{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}^\top) \rangle - \text{Tr} \langle \hat{\boldsymbol{\theta}} \rangle \langle \hat{\boldsymbol{\theta}} \rangle^\top + \|\text{Bias}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\|^2 \\ &= \text{Tr} V(\boldsymbol{\theta}) + \|\text{Bias}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})\|^2. \end{aligned} \quad (2.33)$$

Linearity of the expectation and cyclic permutation of the elements of the trace were used in the penultimate step. So the MSE measures the bias and the sum of the variances of each component of $\hat{\boldsymbol{\theta}}$. This property explicitly motivates the definition we gave for MVUE. Moreover, it is a standard measure (Devore and Berk 2012) of the usefulness of estimators amongst statisticians, and makes precisely what we mean when we say that one estimator is better than the other.

Now suppose $\mathbf{y} \sim N(\boldsymbol{\theta}, \sigma^2 I)$. Then according to the above, the log likelihood is

$$\log L(\boldsymbol{\theta}) = \sum_i^n -\frac{1}{2} \log 2\pi - \log \sigma - \frac{1}{2} \frac{(y_i - \theta_i)^2}{\sigma^2},$$

so the score function and the maximum likelihood estimators are

$$0 = \mathbf{s}_i = (\nabla L(\boldsymbol{\theta}, \sigma^2))_i = -\frac{1}{2} 2 \frac{(y_i - \theta_i)}{\sigma^2} (-1) = \frac{y_i - \theta_i}{\sigma^2} \quad \text{only if} \quad \mathbf{y} = \hat{\boldsymbol{\theta}}.$$

Assume now that $n \geq 3$ and define now the **James-Stein estimator** $\check{\boldsymbol{\theta}}$ by

$$\begin{aligned} \check{\boldsymbol{\theta}} &= \left(1 - \frac{(n-2)\sigma^2}{\|\hat{\boldsymbol{\theta}}\|^2} \right) \hat{\boldsymbol{\theta}} \quad \text{only if} \quad \langle \check{\boldsymbol{\theta}} \rangle = \left\langle \left(1 - \frac{(n-2)\sigma^2}{\|\hat{\boldsymbol{\theta}}\|^2} \right) \hat{\boldsymbol{\theta}} \right\rangle \\ &= \boldsymbol{\theta} - \underbrace{(n-2)\sigma^2 \left\langle \frac{\hat{\boldsymbol{\theta}}}{\|\hat{\boldsymbol{\theta}}\|^2} \right\rangle}_{\neq \mathbf{0} \text{ for any } n \geq 3} \neq \boldsymbol{\theta}. \end{aligned}$$

So $\check{\boldsymbol{\theta}}$ is biased according to the definition. Stein (1956) surprised with the proof that a biased, ad-hoc non-maximum likelihood estimator would uniformly dominate the maximum likelihood estimator in mean square error. Furthermore James and Stein (1961) strengthened the results and showed that $\text{MSE}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) \geq \text{MSE}(\check{\boldsymbol{\theta}}; \boldsymbol{\theta})$ for all $n \geq 3$. This is known as **Stein's paradox** (Efron and Morris

1977). Note that whenever $(n-2)\sigma^2 < \|\hat{\theta}\|$, each component of the estimator is shrunk relative to the maximum likelihood estimator. Equation (2.33) revealed that since the estimator is unbiased, and we know that the maximum likelihood estimator is asymptotic unbiased, it follows that $\text{Tr}V(\check{\theta})$ must be a small number since the James-Stein estimator worked so well. The idea is that if any estimator is multiplied by a number $0 \leq k < 1$, the variance is reduced because $V(\hat{\theta}) \equiv V(k\hat{\theta}) = k^2V(\hat{\theta}) < V(\hat{\theta})$. Any such estimator is called a ***shrinkage estimator***. Shrinkage estimation is the set of tools which can be used to strike the ideal compromise between *variance deflation* and *biased inflation*. Experts will see that shrinkage estimates are implicit in Bayesian statistics. This is because in Bayesian statistics, the prior weights the posterior estimate according to the observations. Therefore, unless we choose a uniform prior, Bayesian statistics have built in shrinkage estimation, even though we are not able to factor the shrinkage out as a multiplicative factor, as we saw with the James-Stein estimator: Recall example 20. There, the shrinkage of the estimate was toward the expected estimate μ , in contrast to the frequentist estimate, which presented the observation X as the final answer.

2.4 Time series

2.4.1 Stationarity

Suppose X_1, X_2, \dots is a sequence of random variables, as would be the case if each X_i is the observation from an experiment. There are many ways to gather data. It is possible to imagine X_i as the observation of some experiment at each time point i . If that is the case, then X_1, X_2, \dots is called a ***time series***. In fact, time series data are so plentiful in physics and science that they are interesting to study in their own right. For the purpose of this thesis, I will present properties of time series data from (i) experiments, and in particular, Markov chains because they are useful in physics.

We have already seen the widespread usefulness of identically distributed random variables. We will now introduce a slightly stronger concepts. We will say that a time series is ***weakly stationary*** or ***stationary*** for brevity if (1) there exists a constant $\mu \in \mathbb{R}$ such that $\langle X_i \rangle = \mu$ for all i and (2) the covariance $\text{Cov}(X_i, X_j)$ only depends on the difference $|i - j| = h$. It is automatic that a stationary time series is identically distributed, as you can check. We define now a stronger type of stationarity. We will say that X_1, X_2, \dots is ***strictly stationary*** if the cdf of any set of the type $\{x_i, x_{i+1}, \dots, x_k\}$ has the same cdf as the shifted set $\{x_{i+j}, x_{i+j+1}, \dots, x_{k+j}\}$ for all i, j, k . Any strictly stationary time series is stationary if it has finite variance, as you can check (Brockwell and Davis 2016). In the case that X_1, X_2, \dots is stationary, the covariance $\text{Cov}(X_i, X_j)$ only

depends on the difference $|i - j|$. In this case, we might as well consider the function $\gamma(h) = \text{Cov}(X_i, X_{i+h})$. This function is called the **autocovariance**. Since X_i must have equal variance σ^2 due to stationarity (to see this just take $h = 0$), we can define $\rho(h) = \gamma(h)/\sigma^2$. This function is called the **autocorrelation** (Shumway and Stoffer 2017). Just as we did for the covariance, we can make an autocovariance matrix Σ consisting of the covariances of X_i, X_j . Define also the **sample covariance** by

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-h} (X_i - \bar{X})(X_{i+h} - \bar{X}). \quad (2.34)$$

We will be particularly interested in the autocovariance of one type of time series, namely the autoregressive model time series.

An **autoregressive model** of order p , denoted $\text{AR}(p)$, 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 are independent, identically distributed with zero expected value and constant variance σ^2 for all t . The autoregressive models used are order 1 and 2, have autocovariance in closed form and are stationary (Shumway and Stoffer 2017). Therefore $V(\bar{X})$ is known for $\text{AR}(1)$ and $\text{AR}(2)$ -processes because for all stationary processes $\gamma(h) = \gamma(-h)$ it follows from the definition that

$$\begin{aligned} V(\bar{X}) &= V \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \text{Cov} \left[\sum_{i=1}^n X_i, \sum_{j=1}^n X_j \right] \\ &= \frac{1}{n^2} \left[n\gamma(0) + (n-1)\gamma(1) + \cdots + \gamma(n-1) \right. \\ &\quad \left. + (n-1)\gamma(-1) + (n-2)\gamma(-2) + \cdots + \gamma(1-n) \right] \\ &= \frac{\sigma^2}{n} + \frac{2}{n} \sum_{h=1}^{n-1} \left(1 - \frac{h}{n} \right) \gamma(h) \quad \text{since } \gamma(0) = \sigma^2. \end{aligned} \quad (2.35)$$

Each $\text{AR}(p)$ process has a characteristic polynomial $P(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$. For $\text{AR}(1)$ processes the autocovariance function, it turns out that the autocovariance functions for $\text{AR}(1)$ and $\text{AR}(2)$ ⁶ are given by

$$\gamma(h) = \sigma^2 \frac{\phi^h}{1 - \phi^2} \quad \text{and} \quad \gamma(h) = \sigma^2 a r^{-h} \cos(h\theta + b)$$

⁶Note that, in the thesis we only consider $\text{AR}(2)$ models where the roots z_1 and z_2 of P satisfy $z_1 = z_2^*$ and $|z_i| > 1$.

respectively (Shumway and Stoffer 2017). Here r and θ are such that $z_i = re^{i\theta}$ is any root of the characteristic polynomial P . After r and θ are determined, we determine a and b by $\gamma(0) = \sigma^2$ and the initial conditions

$$\gamma(h) = \phi_1\gamma(h-1) + \phi_2\gamma(h-2) \quad \text{for } h \in \{1, 2\}.$$

This means that after γ is determined for AR(1) and AR(2) processes, we can compute $V(\bar{X})$ exactly using the formula (2.43) (Flyvbjerg and Petersen 1989).

A stationary time series X_1, X_2, \dots typically contains dependent random variables. In section 2.2.3, it was promised that the central limit theorem could be extended to dependent random variables. Since X_1, X_2, \dots is stationary, the autocovariance exists. In the case that $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ it is convention to say that the time series is **asymptotic uncorrelated** or **ρ -mixing** in the literature (Bradley 1987). The idea is that observations from the time series separated by sufficiently large difference in time are nearly independent. Some authors call it *weak independence* for this reason. Moreover, a seminal paper by Ibragimov 1975 shows that a strictly stationary time series, which is asymptotic uncorrelated, has a central limit theorem:

THEOREM 9 (Ibragimov theorem). *Assume X_1, X_2, \dots is a strictly stationary times series which is asymptotic uncorrelated. If $\lim_{n \rightarrow \infty} V(\sum_{i=1}^n X_i) = \infty$, $V(X_i) < \infty$ and $\langle |X_i|^{2+k} \rangle < \infty$ for some $k > 0$, then*

$$\frac{\bar{X} - \langle \bar{X} \rangle}{V(\bar{X})^{1/2}} \sim N(0, 1) \quad \text{as } n \rightarrow \infty.$$

This theorem was one of the major results from 20th century probability theory because it says that the central limit theorem continue to hold under mild conditions. According to Bradley (1987), the conditions $\lim_{n \rightarrow \infty} V(\sum_{i=1}^n X_i) = \infty$ and $V(X_i) < \infty$ are not restrictive and are standard in this context. Therefore, the most restrictive conditions are ρ -mixing and strict stationarity. Fortunately, these are precisely the conditions required for the present thesis.

2.4.2 Markov chain Monte Carlo theory

Markov chains are frequently used by physicists to generate random variables from probability distributions that are difficult to obtain analytically. This is due to theorem 14, which is called the Hastings-Metropolis theorem. Markov chains themselves are a simplification of an arbitrary stochastic process, where we assume that the transition probability of the Markov chain only depends on its present configuration. As such one can say that Markov chains are suitable to model processes with short term memory. Despite this, we shall see that Markov chains induce enough structure that they are worthwhile to study. Suppose

$\{X_n\}_{n=1}^\infty$ is a time series and X_n can take values from a set S . Then, we call S the **state space** of X_n . If in addition

$$P(X_{n+1} = i | X_n = j, X_{n-1} = j_1, \dots, X_0 = j_n) = P(X_{n+1} = i | X_n = j) \\ \text{for all } i, j, j_k \in S \text{ and } n \in \mathbb{N},$$

then we say that $\{X_n\}$ is a **Markov chain** (Ross 2014). We will concern ourselves with strictly stationary Markov chains on a countably infinite state space, so we can define

$$P(X_{n+1} = j | X_n = i) = P_{ij} \quad \text{for all } n, i, j.$$

The numbers P_{ij} are called the **transition probabilities** of X_n (Ross 2014). Suppose P_{ij} is the transition probability matrix of a Markov chain X_n and S is the state space. Let $P_{ij}^n = P(X_{k+n} = j | X_k = i)$. These numbers are called the **n -step transition probabilities**, and denote the probability that the Markov chain transitions from $i \in S$ to $j \in S$ in exactly n steps. We say that $i \in S$ and $j \in S$ **communicate** if there exists $n \in \mathbb{N}$ such that $P_{ij}^n > 0$. Each subset of S such that alle state communicate are called a **class**. If S contains exactly one class, then we say that X_n is **irreducible** (Ross 2014). In other words, a Markov chain is irreducible if and only if it is possible to reach every state from every other state.

Suppose $i \in S$ and define $\pi_i = P(X_n = i)$. These numbers are called the **stationary probabilities** of X_n . Let $f_i = P(X_n = i \text{ for some } n \in \mathbb{N} | X_0 = i)$ denote the probability that the Markov chain returns to state i if it starts in state $i \in S$. The state i is called **recurrent** if $f_i = 1$ and **transient** if $f_i < \infty$. It turns out that if $A \subset S$ is a class, and there exists $i \in S$ such that i is recurrent/transient, then $j \in A$ is also recurrent/transient. Assume that the state $j \in S$ is recurrent, $X_0 = j$ and define m_j to the expected number of transitions before X_n returns to j . We say that j is **null recurrent** if $m_j = \infty$ and we say j is **positive recurrent** if $m_j < \infty$ (Ross 2014).

PROPOSITION 3. *Any Markov chain, which is irreducible and recurrent, has $\pi_j = 1/m_j$. So an irreducible, null recurrent Markov chain has $\pi_i = 0$ for all $i \in S$. If it is irreducible and positive reccurent, then $\pi_i > 0$.*

Thus, in the case that the Markov chain is null recurrent, it does not have a stationary distribution. It turns out that it does not have a stationary distribution in the case that it is transient either:

THEOREM 10. *An irreducible Markov chain has stationary probabilities if and only if it is positive recurrent.*

A Markov chain X_n is said to be **periodic** if it can only return to each state in a multiple of $d > 1$ steps. A Markov chain that is irreducible and not periodic is called **aperiodic**. It turns out that a periodic Markov chain does not have stationary probabilities while an aperiodic Markov chain does (Ross 2014). A Markov chain, which is aperiodic and positive recurrent, is called **ergodic**. Let Y_n be related to X_n by $Y_{n-k} = X_{n+k}$ for all $k \in \mathbb{Z}$. If Y_n is a Markov chain, it is called the **time reversed** Markov chain of X_n (Ross 2014).

THEOREM 11. *Assume $\{X_n\}_{n=1}^\infty$ is an irreducible Markov chain with transition probabilities P_{ij} . Then it has stationary distribution $\{x_i\}_{i \in S}$ if and only if there exists unique $x_i > 0$ for all $i \in S$ such that*

$$x_i = \sum_{j \in S} x_j P_{ji} \quad \text{and} \quad \sum_{i \in S} x_i = 1.$$

Moreover, stationary distributions are unique.

Proof of the above is perhaps most easily obtained from Grimmett and Welsh (2014).

PROPOSITION 4. *Assume $\{X_n\}$ is an ergodic Markov chain with transition probabilities P_{ij} , stationary distribution π_i and we define Y_n by $Y_{n-k} = X_{n+k}$ for all $k \in \mathbb{Z}$, then Y_n is a Markov chain with transition probability matrix*

$$Q_{ij} = \frac{\pi_j P_{ji}}{\pi_i}.$$

PROOF. Since X_n is a Markov chain,

$$P(X_{m+k} = i | X_{m+k-1} = j, X_{m+k-2} = j_2, \dots) = P(X_{m+k} = i | X_{m+k-1} = j),$$

by definition. That means X_{m+k} and X_m are independent for all $k > 1$. But that means X_m and X_{m+k} are independent for all $k > 1$ (just changing the order of the random variables X_i). That means

$$P(X_m = i | X_{m+1} = j, X_{m+2} = j_2, \dots) = P(X_m = i | X_{m+1} = j)$$

with the implication that

$$\begin{aligned} P(Y_m = i | Y_{m-1} = j, Y_{m-2} = j_2, \dots) &= P(X_m = i | X_{m+1} = j, X_{m+2} = j_2, \dots) \\ &= P(X_m = i | X_{m+1} = j) = P(Y_m = i | Y_{m-1} = j) \end{aligned}$$

So Y_n is a Markov chain. To get the transition probability matrix we use Bayes theorem, equation (2.31), and write:

$$\begin{aligned} Q_{ij} &= P(Y_n = j | Y_{n-1} = i) = \frac{P(Y_{n-1} = i | Y_n = j) P(Y_{n-1} = j)}{P(Y_{n-1} = i)} \\ &= \frac{P(X_{n+1} = i | X_n = j) \pi_j}{\pi_i} = \frac{P_{ji} \pi_j}{\pi_i} \end{aligned}$$

as required. ■

THEOREM 12. Suppose $\{X_n\}_{n=1}^\infty$ is an irreducible Markov chain with transition probabilities P_{ij} and there exists numbers $x_i > 0$ for all $i \in S$ such that $\sum_{i \in S} x_i = 1$. If there is a transition probability matrix Q_{ij} such that

$$x_i Q_{ij} = x_j P_{ji}, \quad (2.36)$$

then X_n is ergodic, has a time reversed Markov chain Y_n with transition probabilities Q_{ij} . Moreover x_i are the stationary probabilities of X_n and Y_n .

PROOF. First see that the Markov chain is positive recurrent. Sum each side of equation (2.36) and use that Q_{ij} is a transition probability matrix, thus summing it over j equals one:

$$x_i Q_{ij} = x_j P_{ji} \quad \text{only if} \quad x_i = x_i \sum_{j \in S} Q_{ij} = \sum_{j \in S} x_j P_{ji},$$

only if $\{x_i\}_{i \in S}$ are the stationary probabilities of X_n according to theorem 11. But that means it is positive recurrent according to theorem 10. Since the Markov chain is irreducible and positive recurrent it is also aperiodic, since periodic Markov chains do not have stationary distributions that means X_n is not periodic as explained above. Moreover, it is recurrent, so X_n is ergodic. According to proposition 4, it has a reversed Markov chain Y_n with transition probability matrix Q_{ij} such that $x_i Q_{ij} = x_j P_{ji}$. ■

An ergodic Markov chain with transition probability P_{ij} is said to be **time reversible** if the reversed transition probabilities $Q_{ij} = P_{ij}$.

THEOREM 13 (Detailed balance). Assume $\{X_n\}_{n=1}^\infty$ is an ergodic Markov chain with transition probability P_{ij} and stationary probabilities π_i . Then X_n is time reversible if and only if

$$\pi_i P_{ij} = \pi_j P_{ji}$$

for all $i, j \in S$.

PROOF. The Markov chain is ergodic, so the time reversed Markov chain has transition probability matrix Q_{ij} such that $Q_{ij} = \pi_j P_{ji} / \pi_i$. Assume first that the Markov chain is time reversible, then $Q_{ij} = P_{ij}$, so $\pi_i P_{ij} = \pi_j P_{ji}$ is immediate. Conversely, assume that $\pi_i P_{ij} = \pi_j P_{ji}$ for all $i, j \in S$. Since the Markov chain is ergodic $\pi_i > 0$ for all $i \in S$ according to proposition 3, so we can write $P_{ij} = \pi_j P_{ji} / \pi_i$. That means

$$Q_{ij} = \frac{\pi_j P_{ji}}{\pi_i} = P_{ij},$$

so the Markov chain is time reversible. ■

THEOREM 14 (Hastings-Metropolis theorem). *Suppose that $C\pi_i > 0$ is a discrete probability distribution. If q_{ij} is any irreducible, positive recurrent transition probability matrix, and X_n is a Markov chain with transition probability matrix*

$$P_{ij} = \begin{cases} \alpha_{ij}q_{ij}, & j \neq i \\ q_{ii} + \sum_{k=0}^{\infty} q_{ik}(1 - \alpha_{ik}) & j = i \end{cases}, \quad \text{where} \quad \alpha_{ij} = \min \left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1 \right), \quad (2.37)$$

then X_n is time reversible with stationary distribution $C\pi_i$.

PROOF. Assume that the hypothesis is true. Then clearly $\sum_{i \in S} C\pi_i = 1$. Notice that if

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} = 1,$$

then there is nothing to prove since then $\alpha_{ij} = 1$ and $\alpha_{ji} = 1$, and therefore

$$\pi_i P_{ij} = \pi_j P_{ji} \quad (2.38)$$

is automatic. Hence it suffices to prove (2.38) for the two cases

$$\frac{\pi_j q_{ji}}{\pi_i q_{ij}} > 1 \quad \text{and} \quad \frac{\pi_j q_{ji}}{\pi_i q_{ij}} < 1,$$

separately. Suppose first that $\pi_j q_{ji} > \pi_i q_{ij}$ (\dagger). Write:

$$\begin{aligned} \pi_i P_{ij} &\stackrel{(2.37)}{=} \pi_i q_{ij} \alpha_{ij} \stackrel{(2.37)(\dagger)}{=} \pi_i q_{ij} \cdot 1 = \pi_i q_{ij} \frac{\alpha_{ji}}{\alpha_{ji}} = \alpha_{ji} \pi_i q_{ij} \frac{1}{\alpha_{ji}} \\ &\stackrel{(\dagger)}{=} \alpha_{ji} \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \alpha_{ji} \pi_j q_{ji} \stackrel{(2.37)}{=} \pi_j P_{ji}. \end{aligned}$$

In the case that $\pi_j q_{ji} < \pi_i q_{ij}$ (\ddagger), write

$$\pi_i P_{ij} \stackrel{(2.37)}{=} \pi_i q_{ij} \alpha_{ij} \stackrel{(2.37)(\ddagger)}{=} \pi_i q_{ij} \frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \pi_j q_{ji} = \pi_j q_{ji} \cdot 1 \stackrel{(2.37)(\ddagger)}{=} \pi_j q_{ji} \cdot \alpha_{ji} = \pi_j P_{ji},$$

which means X_n is ergodic, has a time reversed Markov chain Y_n with transition probabilities $Q_{ij} = P_{ij}$ and stationary probability π_i by theorem 12. That means it is time reversible by theorem 13. ■

The **Hastings-Metropolis algorithm** is just the act of generating the Markov chain X_n from theorem 14 (Ross 2014). We first propose an algorithm which does this, afterwards we check that it is indeed X_n : Assume X_0 is given. Suppose the Markov chain is in state i at time n . Now draw some state j from S according to the probability distribution q_{ij} . Then set $X_{n+1} = j$ with probability α_{ij} . If we do not set $X_{n+1} = j$, then set $X_n = i$. Repeat for all $n \in \mathbb{N}$. Let us check that this Markov chain is X_n :

Suppose the Markov chain is in state i at time n . The Markov chain moves from i to j *precisely* when (1) j is drawn, and (2) given that j is drawn, a transition to j happens with probability α_{ij} . Using the definition of conditional probability (*) we get:

$$\begin{aligned} P_{ij} &= P(X_{n+1} = j \mid X_n = i) = P(\text{state } j \text{ is drawn and transition to } j \mid X_n = i) \\ &\stackrel{(*)}{=} P(\text{transition to } j \mid X_n = i, \text{state } j \text{ is drawn})P(\text{state } j \text{ is drawn} \mid X_n = i) \\ &= q_{ij}\alpha_{ij}. \end{aligned} \tag{2.39}$$

It only remains to show that P_{ii} is as in the theorem. P_{ii} is the probability that if $X_n = i$, then it does not move to $j \in S$ for any $j \neq i$. We use the property that $P(A) = 1 - P(\text{not } A)$ and that if A_i is a sequence of mutually exclusive events, then $P(A_i \mid \text{for one } i) = \sum_i P(A_i)$ and obtain⁷

$$\begin{aligned} P_{ii} &= P(X_{n+1} = i \mid X_n = i) = 1 - P(X_{n+1} \neq i \mid X_n = i) \\ &= 1 - P(X_{n+1} = k \text{ for some } k \neq i \mid X_n = i) = 1 - \sum_{j \neq i \in S} P(X_{n+1} = k \mid X_n = i) \\ &\stackrel{(2.39)}{=} 1 - \sum_{k \neq i \in S} q_{ik}\alpha_{ik}, \end{aligned}$$

which proves that the Markov chain generated this way is indeed X_n . In the case that the probability distribution g_{ij} is uniform, we call the procedure the **Metropolis algorithm** (Ross 2014).

2.5 Numerical estimation

2.5.1 Resampling methods

Efron (1987) explains that resampling methods 'scramble' the observations which describe the parameter θ in some way. The purpose of scrambling the data is to obtain useful estimates of the probability distribution of the estimator $\hat{\theta}$. This is often done if deriving the distribution of $\hat{\theta}$ by analytical means is impossible or inconvenient. The significance of this is reflected in that Efron's original paper has more than 16 000 citations by early spring 2018. Although these citations have come from all the sciences, a lot of work has been done by statisticians and mathematicians. On 'Web of Science', a search for the topic *bootstrap* returns nearly 6 500 papers in journals on statistics and probability theory alone. A similar search on 'Scopus' returns more than 7 000 papers in the field of mathematics. In addition, there has been a renaissance in the study of resampling methods in

⁷These two properties should be known from high school

the 21st century, with more than 6 000 papers in just 18 years in mathematics. Part of the reason is that, even though the ideas which will be presented here seem innocent and simple, the required mathematics is deep. In fact, there exists conjectures too deep for present mathematics (Efron 1987). This will become apparent to us because often we will only give intuitive explanations for why the methods are valid. We could have done substantially more with measure theory in place, but this is not economical in light of the present results. However, using our introduction to real analysis, it is possible to state and understand a few results in some detail. See for example theorem 16.

Two famous resampling methods are *the independent bootstrap* and *the jackknife*. It would make most sense to start by discussing the independent bootstrap, because the jackknife method follows by making a linearization of the parameters of interest (Efron 1987; Efron 1979). As such, the jackknife is a special case of the independent bootstrap (Efron 1987). Still, the jackknife was made popular prior to the independent bootstrap. And as the popularity of the independent bootstrap soared, new variants, such as *the dependent bootstrap*⁸ or stationary bootstrap were introduced, see for example Politis and Romano (1994) or Politis and White (2006). There also exists textbooks on the subject. The mathematical complexity of the latter variants is also greater, and consequently it is pedagogical to introduce the methods in this order.

The Jackknife and independent bootstrap work for independent, identically distributed random variables (Efron 1987). If these conditions are not satisfied, the methods will fail. This is important for the results of the thesis, because here the variables are dependent, and we will need the dependent bootstrap. Yet, it should be said that if the data are independent, identically distributed, and we only want to estimate the variance of \bar{X} (which often is the case), then there is no need for bootstrapping. For if X_1, X_2, \dots, X_n are independent identically distributed and come from an unknown distribution F , then the standard error is easily computed by taking the square root of the following expression: (2.9)

$$V(\bar{X}) \stackrel{(2.9)}{=} \frac{\sigma^2}{n} \approx \frac{\hat{\sigma}^2}{n} \stackrel{(2.3)}{=} \frac{1}{n^2} \sum_{i=1}^n (X_i - \bar{X})^2.$$

And consequently, these methods are most useful when the data are dependent or the estimator is not the sample mean.

⁸We will only consider non-parametric bootstrap, but there exists a popular variant called parametric bootstrap, which assumes knowledge of the probability distribution of the observations

The Jackknife

The Jackknife works by making many replicas of the estimator $\hat{\theta}$. Since the jackknife is a resampling method, we explained that this happens by scrambling the data in some way. When using the jackknife, this is done by systematically leaving out one observation from the vector of observed values $\mathbf{X} = (X_1, X_2, \dots, X_n)$ (Tukey 1958). Let \mathbf{X}_i denote the vector

$$\mathbf{X}_i = (X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_n),$$

which equals the vector \mathbf{X} with the exception that observation number i is left out. Using this notation, define $\hat{\theta}_i$ to be the estimator $\hat{\theta}$ computed using \mathbf{X}_i . According to Efron (1987), to get an estimate for the bias and standard error of $\hat{\theta}$, use the following estimators for each component of $\hat{\theta}$:

$$\widehat{\text{Bias}}(\hat{\theta}, \theta) = (n-1) \left(-\hat{\theta} + \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i \right) \quad \text{and} \quad \hat{\sigma}_{\hat{\theta}}^2 = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_i - \frac{1}{n} \sum_{j=1}^n \hat{\theta}_j \right)^2.$$

Sample code is contained in figure 2.3 and is available for download from the url github.com/computative/resample.

The independent bootstrap

Many authors explain the bootstrap in an algebraic manner, similar to the way it was originally proposed by Efron (1987). In this thesis, I give a different view of the method; building our intuition upon Parr (1985): Since $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is a function of random variables, $\hat{\theta}$ itself must be a random variable. Thus it has a pdf, call this function $p(\mathbf{t})$. The aim of the bootstrap is to estimate $p(\mathbf{t})$ by the relative frequency of $\hat{\theta}$. You can think of this as using a histogram in the place of $p(\mathbf{t})$. If the relative frequency closely resembles $p(\mathbf{t})$, then using numerics, it is straight forward to estimate all the interesting parameters of $p(\mathbf{t})$ using point estimators. If the probability density function of X_i , $p(x)$, had been known, then it would have been straight forward to do this by: (1) Drawing lots of numbers from $p(x)$, suppose we call one such set of numbers $(X_1^*, X_2^*, \dots, X_n^*)$. (2) Then using these numbers, we could compute a replica of $\hat{\theta}$ called $\hat{\theta}^*$. By repeated use of (1) and (2), many estimates of $\hat{\theta}$ could have been obtained. The idea is to use the relative frequency of $\hat{\theta}^*$ (think of a histogram again) as an estimate of $p(\mathbf{t})$.

But unless there is enough information available about the process that generated X_1, X_2, \dots, X_n , $p(x)$ is in general unknown. Therefore, Bradley Efron (1979) asked the natural question: What if we replace $p(x)$ by the relative frequency of the observation X_i ; if we draw observations in accordance with the relative frequency of the observations, will we obtain the same result in some

```

# jack.py

def jack(data, stat):
    n = len(data); t = zeros(n); inds = arange(n); t0 = time()
    # 'jackknifing' by leaving out an observation for each i
    for i in range(n):
        t[i] = stat(delete(data,i) )

    return t

# define a function which returns your chosen estimator theta-hat
def stat(data):
    theta-hat = mean(data)
    return theta-hat

# boot returns the bootstrap sample
t = jack(X, stat)

```

Figure 2.3: The code follows the algorithm outlined in the text. Consider first the function `jack()`. In the `for`-loop, this function repeatedly estimates the function called `statistic()` under the resampled data by systematically leaving out one observation from the data. The function `stat()` is passed as an argument to `jack()`. The array `t` is eventually returned, which contains all the estimates $\hat{\theta}$, and can be plotted or analysed in other ways, such as by calling `std(t)` from `numpy` to estimate the standard error of $\hat{\theta}$. The function `std(t)` is just the estimator $\hat{\sigma}^2$.

```

# boot.py

def boot(data, statistic, R):
    t = zeros(R); n = len(data); inds = arange(n); t0 = time()

    # non-parametric bootstrap
    for i in range(R):
        t[i] = statistic(data[randint(0,n,n)])

    return t

# define a function which returns your chosen estimator theta-hat
def stat(data):
    theta-hat = mean(data)
    return theta-hat

t = boot(X, stat, 2**9)

```

Figure 2.4: The code follows the algorithm 2.6. Consider first the function `boot()`. In the `for`-loop, this function repeatedly estimates the function called `statistic()` under the resampled data `data[randint(0,n,n)]`. The function `statistic()` is passed as an argument to `boot()`. The array `t` is eventually returned, which contains all the estimates $\hat{\theta}$, and can be plotted or analysed in other ways, such as by calling `std(t)` from `numpy` to estimate the standard error of $\hat{\theta}$. The function `std(t)` is just the estimator $\hat{\sigma}^2$.

asymptotic sense? The answer is yes. The paper of Efron (1979) gave little in the way of general theory (Efron 1979). In contrast, he gave computational examples showing that in many cases, it was reasonable. It is standard to make a tweak that also speeds up computation: Instead of generating the histogram for the relative frequency of the observation X_i , just draw the values $(X_1^*, X_2^*, \dots, X_n^*)$ with replacement from the vector \mathbf{X} . The end result is exactly the same, as bootstrapping is a demonstration of.

See figure 2.5 for a pictorial explanation and figure 2.6 for a concise summary and the final algorithm. Sample code for python is contained in figure 2.4 and is available for download at github.com/computative/resample.

As we explained, much theoretical work has gone into making the mathematics of the various types of bootstrap rigorous. However, some of the most important results were published relatively soon after. Only two years after Efron (1979),

Philosophy of the independent bootstrap

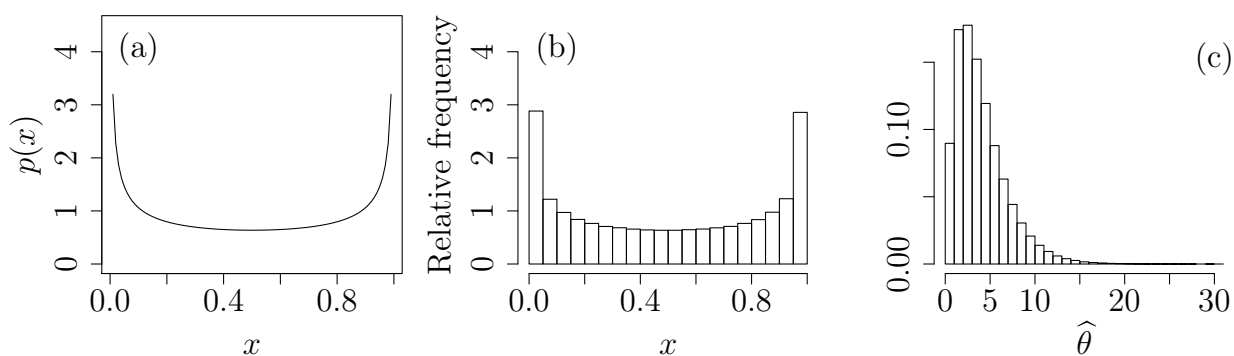


Figure 2.5: Suppose you wanted to estimate the probability distribution $p(t)$ of the estimator $\hat{\theta}$. The obvious way to do this is to compute many replicas of some $\hat{\theta} = \hat{\theta}(\mathbf{X})$ by drawing lots of numbers from some pdf $p(x)$, such as the one of fig 2.5(a). Then by plotting the histogram of the replicas $\hat{\theta}$, you obtain the estimate of the pdf $p(t)$, namely (b). Bradley Efron asked the question: What happens if we replace the exact distribution $p(x)$ by an estimate, namely the histogram of the relative frequency (b)? It turns out that in an asymptotical sense, we still obtain the same estimate, (c). This is the philosophy of the independent bootstrap. After the estimate (c) has been obtained we can estimate any statistic thereof, for example $V(\hat{\theta})$ using \widehat{S}^2 .

Flow chart of the independent bootstrap

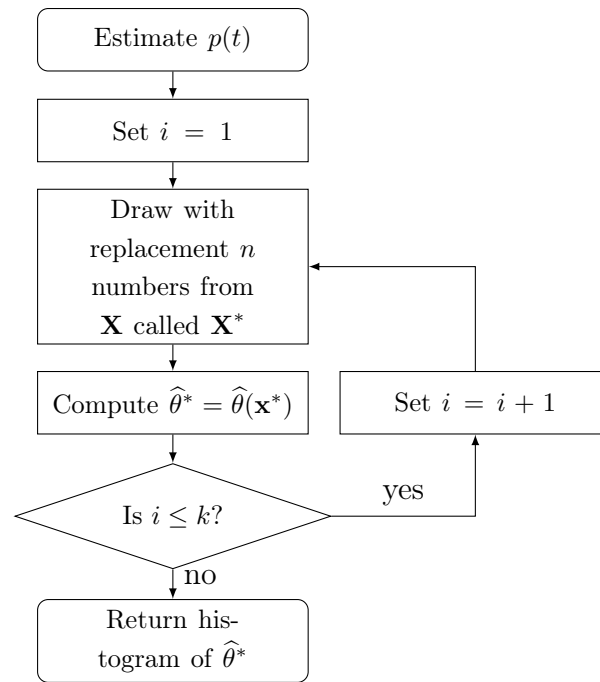


Figure 2.6: The independent bootstrap works like this: (1) Draw with replacement n numbers for the observed variables $\mathbf{x} = (x_1, x_2, \dots, x_n)$. (2) Define a vector \mathbf{x}^* containing the values which were drawn from \mathbf{x} . (3) Using the vector \mathbf{x}^* compute $\hat{\theta}^*$ by evaluating $\hat{\theta}$ under the observations \mathbf{x}^* . Repeat this process k times. When you are done, you can draw a histogram of the relative frequency of $\hat{\theta}^*$. This is your estimate of the probability distribution $p(t)$. Using this probability distribution you can estimate any statistic thereof. In principle you never draw the histogram of the relative frequency of $\hat{\theta}^*$. Instead you use the estimators corresponding to the statistic of interest. For example, if you are interested in estimating the variance of $\hat{\theta}$, apply the estimator $\hat{\sigma}^2$ to the values $\hat{\theta}^*$.

proof in the case that $(\theta, \hat{\theta}) = (\langle X \rangle, \bar{X})$ came from Bickel and Freedman (1981). Despite the importance of these quantities, we will not be fully satisfied with this result. This is because bootstrapping is most useful when $\hat{\theta}$ is not the sample mean, as we discussed. Before formulating the theorem, some convenient notation is introduced. In ordinary real analysis, we are often interested in convergence of sequences of real numbers $\{a_n\}_{n=1}^{\infty}$ with limit $a \in \mathbb{R}$. In the case that the sequence $\{A_n\}_{n=1}^{\infty}$ is comprised of random variables, it is clearly nonsense to say that the limit is a real number. One could argue that it makes more sense to say that the limit is a random variable. However, due to amount of additional structure that comes with measure theory, there are many interesting ways of defining such limits. And van der Vaart (1998) covers some of the most useful modes of convergence:

- If F_{A_n} is the cdf of A_n and F_A is the cdf of A , and F_{A_n} converges pointwise to F_A then we say that there is **convergence in distribution** and denoted by $A_n \xrightarrow{d} A$.
- If for every $\varepsilon > 0$, the sequence of real numbers $a_n = P(\|A_n - A\| > \varepsilon)$ converge to zero, then we say that there is **convergence in probability**, denoted by $A_n \xrightarrow{P} A$.
- If for every $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that for all $n \geq N$ we have $\|A_n - A\| < \varepsilon$ with probability 1, then we say that there is **almost sure convergence** denoted by $A_n \xrightarrow{\text{a.s.}} A$.

More generally, we will say that an event B happens **almost surely** if $P(B) = 1$ (Øksendal 2014; McDonald and Weiss 2012). On that note we are ready for the theorem due to Bickel and Freedman (1981):

THEOREM 15 (Bickel-Freedman theorem). *Assume X_1, X_2, \dots are independent identically distributed random variables with variance σ^2 , and assume $X \sim N(0, \sigma^2)$, then given X_1, X_2, \dots, X_n ,*

- $n^{1/2}(\bar{X}^* - \bar{X}) \xrightarrow{d} X$ as $n \rightarrow \infty$ almost surely.
- $\hat{\sigma}^* \xrightarrow{P} \sigma$ as $n \rightarrow \infty$ almost surely.

For more a general $\hat{\theta}$ the problem is harder because the assumptions are weaker. Parr (1985) established the result under relaxed conditions on $\hat{\theta}$. We will consider the case that $\hat{\theta} = \hat{\theta}$ is one-dimensional and require the strong type of differentiability which was defined in section 2.1; namely Fréchet differentiability. All estimators encountered thus far were expressed in terms of \mathbf{X} . However, Huber and Ronchetti (2009) explains that many estimators with practical use can also be expressed in terms of their cumulative distribution function. This area of

estimation contains ***M-estimation***, and is not economical to discuss in full generality. However, using our experience with real analysis, it is possible to tackle the theory relevant to maximum likelihood estimation on compact intervals $[a, b]$. Going back to the way justification was given for maximum likelihood estimators in section 2.3.1 it is clear that the maximum likelihood estimator is

$$\hat{\theta} = \operatorname{argmin}_{\theta} \sum_{i=1}^n -\log(f(x_i; \theta)) = \sum_{i=1}^n -\log \left(\frac{\partial F}{\partial x}(x_i; \theta) \right) = \hat{\theta}(F). \quad (2.40)$$

This shows that the Maximum likelihood estimators are M -estimators. In fact M estimation was motivated by maximum likelihood and is a generalization thereof (Huber and Ronchetti 2009). These estimators suffices for this thesis, since the maximum likelihood estimators are MVUE according to section 2.3.1 and is the best we can do with our present tools. We introduce the Fréchet derivative for this space of estimators. Let $H([a, b])$ be the largest open set of cdfs on $[a, b]$. It is well known that $[a, b]$ is compact (Munkres 2000). Therefore, example 8 shows that $H([a, b])$ is contained in the Banach space $C([a, b], \mathbb{R})$ under the sup-norm, $\|\cdot\|_{\infty}$ from example 6. Example 8 also shows that by letting ψ be a continuous function on $X \equiv [a, b]$ such that (1) $\int_X \psi \, dF = 0$, (2) $\int_X \psi^2 \, dF < \infty$ and (3) $F \in H$ be the cdf of X_i with pdf f , then

$$(A(F))(\theta) = \left(\int_X \psi \, dF \right)(\theta) = \int_X \psi(x; \theta) f(x; \theta) \, dx$$

is a bounded linear functional on H . So if T is any $H([a, b]) \rightarrow \mathbb{R}$ function, we say that T is Fréchet differentiable at F if

$$\lim_{C \rightarrow 0} \frac{|T(F) - T(F + C) - (A(C))(T)|}{\|C\|_{\infty}} = 0 \quad (\text{Parr 1985}).$$

You may wonder what the function ψ is. According to Huber and Ronchetti (2009), it defines the type of estimation equation used. In the case of maximum likelihood estimation with F is twice continuously differentiable,

$$\psi(x; T) = -\frac{\partial \log f}{\partial \theta}(x; T) = -\left(\frac{1}{\partial F / \partial x} \frac{\partial^2 F}{\partial x \partial \theta} \right)(x; T) \quad (\text{Huber and Ronchetti 2009}).$$

Using this definition, we are finally ready for a theorem due to Parr (1985) which explains consistency of the bootstrap estimator more generally. It says that if the Fréchet derivative of $\hat{\theta}$ exists at $F \in H$, then the bootstrap estimator is consistent:

THEOREM 16 (Parr theorem). *Assume X_1, X_2, \dots, X_n are independent identically distributed with support $X = [a, b]$, $\hat{\theta}$ is Fréchet differentiable at F , $\sigma^2 = V(n^{1/2}[\hat{\theta}^* - \hat{\theta}])$ and $Z \sim N(0, 1)$, then given X_1, X_2, \dots, X_n we have $n^{1/2}(\hat{\theta}^* - \hat{\theta}) \xrightarrow{P} \sigma Z$ as $n \rightarrow \infty$.*

The dependent bootstrap

In the case that the variables X_1, X_2, \dots, X_n are dependent, the above procedure breaks down. At least one of the problems are: Independent bootstrapping assumes that X_1, X_2, \dots, X_n all come from the same marginal probability distribution, $p(x) = p(x_i)$ for all $1 \leq i \leq n$. This is clear because variables are independent, so it follows from the definitions of independence and conditional probability that $p(x_i) = p(x_i|x_j \neq x_i)$. Also the second definition given of independence says that the joint probability distribution of all the X_i is $p(x_1, x_2, \dots, x_n) = p(x)^n$ by the product rule. Since this does not carry over in the case that the variables are dependent, there are at least two problems:

1. Since there is dependence between the observation, observing x_j reveal information about x_i for some $1 \leq i \leq n$. The extra information supplied by observing x_j means that $p(x_i) \neq p(x_i|x_j \neq x_i)$ as explained above. That means if we treated the variables as independent (i.e. let $p(x_i) = p(x_i|x_j \neq x_i)$), then all estimators which are sensitive to the difference between $p(x_i)$ and $p(x_i|x_j \neq x_i)$ estimate wrong systematically. Consider for example the case that $\{X_i\}$, is a time series where we wrongly assumed that the X_i were independent, then the autocovariance estimator $\hat{\gamma}(1) = \widehat{\text{Cov}}(X_i, X_{i+1}) = \hat{0} = 0$ would estimate zero. In independent bootstrapping is induced by pick single observations with replacement as we explained. We explained above that this is equivalent to constructing the histogram for X_i and drawing observations from the histogram. But by hypothesis, that is precisely the estimate of the marginal distribution $p(x)$.
2. If we do not assume that $p(x_i) = p(x_i|x_j \neq x_i)$, but instead assume that $p(x_i) \neq p(x_i|x_j \neq x_i)$, then the natural way to proceed is to treat the whole set $\{X_i\}$ as one observation $\mathbf{X} = (X_1, X_2, \dots, x_n)$ and give it a multivariate probability distribution with covariance Σ which encodes the dependence. But there is at least one problem, in this case we only have one observation, namely \mathbf{X} ! This problem is however fixable in the case that the dependent data are a stationary time series, as explained by Politis and Romano (1994) and adapted here.

That is to assume that the dependence in the data set can be related to the linear dependence, which is the covariance. In the case that the data are a stationary time series and the autocovariance $\gamma(h) \rightarrow 0$ as $h \rightarrow 0$. Since the autocovariance measures linear dependence, we assume that when the linear dependence is zero, the variables can be treated as independent. If $H \in \mathbb{N}$ a number such that $\gamma(h) \approx 0$ for all $h \geq H$, then we treat X_i and X_{i+h} as independent for all $h \geq H$ and all $1 \leq i \leq n - h$. If we split the observation and make the following

definitions:

$$\mathbf{X} = (\underbrace{X_1, X_2, \dots, X_H}_{\equiv \mathbf{X}'_1}, X_{H+1}, \dots, X_{2H}, \underbrace{X_{2H+1}, \dots, X_{3H}}_{\equiv \mathbf{X}'_3}, X_{3H+1}, \dots, X_n).$$

Then the components of \mathbf{X}'_1 and \mathbf{X}'_3 are almost independent, moreover section 2.4 explains that a stationary time series is identically distributed. So to these vectors can be treated as independent and identically distributed, so we can use the usual machinery of independent bootstrapping. But since the size of each vector \mathbf{X}'_i is larger than one, we require to draw less than n such vectors with replacement to compute $\hat{\theta}^*$. Instead we concatenate the drawn observations \mathbf{X}'_i into one long vector \mathbf{X}^* of length n , and discard any observations left over. This procedure has the advantage that the vector \mathbf{X}^* has the same autocovariance as \mathbf{X} , so it can be used to estimate γ and consequently $V(\hat{\theta})$ according to chapter 2.4. See figure 2.7 for sample code which follows the description given above. The code is also available for download from github.com/computative/resample.

2.5.2 Manual blocking method

The manual blocking method was made popular by Flyvbjerg and Pedersen (1989) and has become one of the standard ways to estimate $V(\hat{\theta})$ for exactly one $\hat{\theta}$, namely $\hat{\theta} = \bar{X}$. Their paper has become a citation classic (cited more than 1000 times according to Google scholar). But the proof given by Flyvbjerg and Pedersen (1989) is not rigorous. I corresponded with Associate professor Flyvbjerg in the summer of 2017 and it became clear to me that a paper on the blocking method with rigorous modern mathematics and numerics could be useful. Thus the missing proof of the manual blocking method is one of the main results of this thesis, and contained in the results, see theorem 17. But this begs the question: What can then be said about the blocking method in the methods of this thesis? I think it is appropriate to give the idea of the blocking method, analogous to the treatment given for bootstrapping, as well as an overview of the mathematics of Flyvbjerg and Petersen (1989).

Assume $n = 2^d$ for some integer $d > 1$ and X_1, X_2, \dots, X_n is a stationary time series to begin with. This guarantees that $\gamma(h)$ exists according to section 2.4. Moreover, assume that the time series is asymptotically uncorrelated. We switch to vector notation by arranging X_1, X_2, \dots, X_n in an n -tuple. Define:

$$\mathbf{X} = (X_1, X_2, \dots, X_n).$$

The strength of the blocking method is evident when the number of observations, n is large. For large n , the complexity of dependent bootstrapping scales poorly, but the blocking method does not, moreover, it becomes more accurate the larger

```

# tsboot.py

def tsboot(data, statistic, R, l):
    t = zeros(R); n = len(data); k = ceil(float(n)/l);
    inds = arange(n); t0 = time()

    # time series bootstrap
    for i in range(R):
        # construct bootstrap sample from
        # k chunks of data. The chunksize is l
        _data = concatenate([data[j:j+l] for j in
                             randint(0, n-l, k)]) [0:n];
        t[i] = statistic(_data)

    return t

# define a function which returns your chosen
# estimator theta-hat
def stat(data):
    theta-hat = mean(data)
    return theta-hat

t = tsboot(X, stat, 2**12, 2**10)

```

Figure 2.7: The code follows the algorithm outlined in the text. Consider first the function `tsboot()`. In the `for`-loop, this function repeatedly estimates the function called `statistic()` under the resampled data by concatenating chunks of the data which are uncorrelated by `concatenate([data[j:j+l] for j in randint(0, n-l, k)]) [0:n]`. The function `statistic()` is passed as an argument to `tsboot()`. The array `t` is eventually returned, which contains all the estimates $\hat{\theta}$, and can be plotted or analysed in other ways, such as by calling `std(t)` from `numpy` to estimate the standard error of $\hat{\theta}$. The function `std(t)` is just the estimator $\hat{\sigma}^2$.

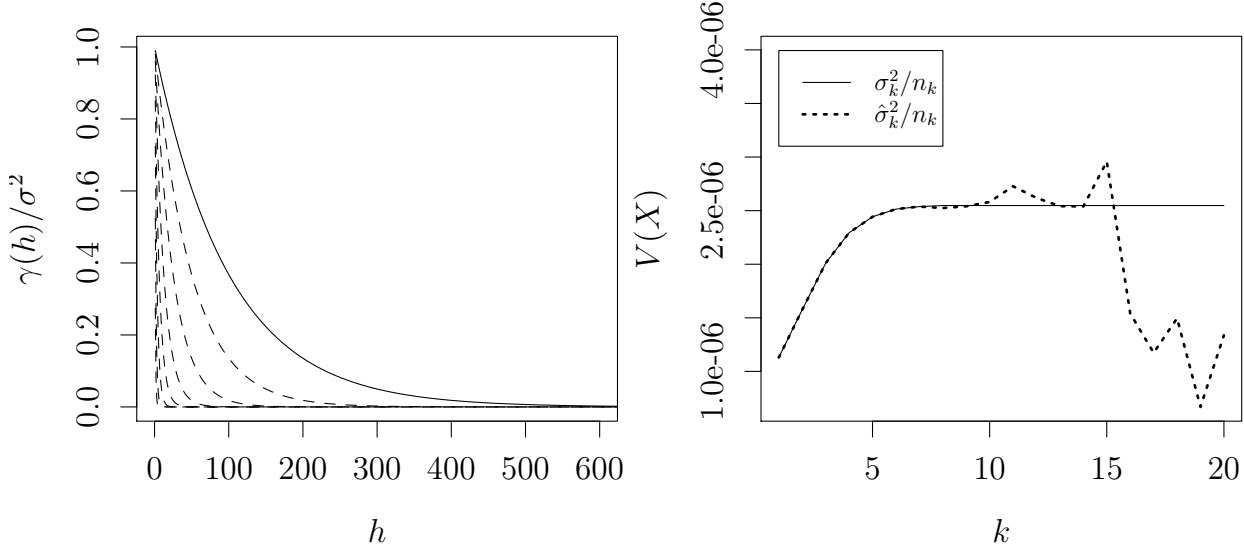


Figure 2.8: In the left panel: A typical autocovariance function $\gamma(h)$ plotted against h . The dashed lines illustrate how $\gamma_k(h) \leq \gamma_{k-1}(h)$ for all $1 \leq k \leq d-1$ in the case of applying blocking transformation. In fact, the results show that γ_k converges to the zero-function of \mathbb{N} .

In the right panel: It is a consequence of the behavior $\gamma_k(h) \leq \gamma_{k-1}(h)$ that for all $1 \leq k \leq d-1$ we have $\sigma_k^2/n_k \geq \sigma_{k-1}^2/n_{k-1}$ for all $1 \leq k \leq d-1$. The solid line is the "unestimated estimate", σ_k^2/n_k . That is, the estimate of $V(\bar{X})$ where the only source of error is the truncation error e_k . It is clear that this estimate initially is too optimistic (too small), and then as we apply blocking transformations, the estimate rises up to the correct value. By equation (2.48), it is clear that this means that σ_k^2/n_k becomes constant, as we clearly see in the plot.

The dashed line is the estimator $\hat{\sigma}_k^2/n_k$ which contains an extra error since σ_k^2 itself is unknown and has to be estimated. As $k \rightarrow d$ (here $d = 20$), we have $n_k \rightarrow 1$, and so the *standard error* of $V(\hat{\sigma}_k^2/n_k)$ becomes very large according to equation (2.47). We see this in the figure, because the estimate starts to depart from the value σ_k^2/n_k . This is the reason it is important to stop the algorithm at the right time. This can be done with a plot, like the one above, or using the automated scheme proposed here which takes care of everything for you.

In this case it is relatively clear from the plot of $\hat{\sigma}_k^2/n_k$ for which k we have that σ_k^2/n_k becomes constant. However if the amount of data is smaller, i.e. n_k is small, this becomes more difficult to determine, because then the estimator $\hat{\sigma}_k^2/n_k$ diverges sooner. Perhaps even before the graph of σ_k^2/n_k becomes constant.

n is, as the results will show. As such the method is relatively ad-hoc. We now define blocking transformations. The idea is to take the mean of subsequent pair of elements from \mathbf{X} and form a new vector \mathbf{X}_1 . Continuing in the same way by taking the mean of subsequent pairs of elements of \mathbf{X}_1 we obtain \mathbf{X}_2 , and so on. In accordance with Flyvbjerg and Petersen (1989) define \mathbf{X}_i recursively by:

$$\begin{aligned} (\mathbf{X}_0)_k &\equiv (\mathbf{X})_k \\ (\mathbf{X}_{i+1})_k &\equiv \frac{1}{2} \left((\mathbf{X}_i)_{2k-1} + (\mathbf{X}_i)_{2k} \right) \quad \text{for all} \quad 1 \leq i \leq d-1 \end{aligned} \quad (2.41)$$

In this way, we say that \mathbf{X}_k is subject to k **blocking transformations**. We now have d vectors $\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_{d-1}$ containing the subsequent averages of observations. It turns out that if the components of \mathbf{X} is a stationary time series, then the components of \mathbf{X}_i is a stationary time series for all $0 \leq i \leq d-1$ (see lemma 3 or Flyvbjerg and Petersen (1989)). And so we can compute the autocovariance, the variance, sample mean, and number of observations for each i . Let $\gamma_i, \sigma_i^2, \bar{X}_i$ denote the autocovariance, variance and average of the elements of \mathbf{X}_i and let n_i be the number of elements of \mathbf{X}_i . It follows by induction that $n_i = n/2^i$. Using the definition of the blocking transformation and the distributive property of the covariance, it is clear that since $h = |i - j|$

$$\begin{aligned} \gamma_{k+1}(h) &= \text{Cov}((X_{k+1})_i, (X_{k+1})_j) \\ &= \frac{1}{4} \text{Cov}((X_k)_{2i-1} + (X_k)_{2i}, (X_k)_{2j-1} + (X_k)_{2j}) \\ &= \begin{cases} \frac{1}{2} \gamma_k(2h) + \frac{1}{2} \gamma_k(2h+1) & \text{if } h = 0 \\ \frac{1}{4} \gamma_k(2h-1) + \frac{1}{2} \gamma_k(2h) + \frac{1}{4} \gamma_k(2h+1) & \text{else} \end{cases} \end{aligned} \quad (2.42)$$

According to lemma 3, since \mathbf{X} is asymptotic uncorrelated by assumption, \mathbf{X}_k is also asymptotic uncorrelated. Let's turn our attention to the variance of the sample mean $V(\bar{X})$. According to equation (2.43) we have

$$V(\bar{X}_k) = \frac{\sigma_k^2}{n_k} + \underbrace{\frac{2}{n_k} \sum_{h=1}^{n_k-1} \left(1 - \frac{h}{n_k}\right) \gamma_k(h)}_{\equiv e_k} = \frac{\sigma_k^2}{n_k} + e_k \quad \text{if } \gamma_k(0) = \sigma_k^2. \quad (2.43)$$

The term e_k is called the **truncation error**:

$$e_k = \frac{2}{n_k} \sum_{h=1}^{n_k-1} \left(1 - \frac{h}{n_k}\right) \gamma_k(h). \quad (2.44)$$

We show that $V(\overline{X}_i) = V(\overline{X}_j)$ for all $0 \leq i \leq d-1$ and $0 \leq j \leq d-1$. This follows by induction. For the induction step write:

$$\begin{aligned} n_{j+1}\overline{X}_{j+1} &= \sum_{i=1}^{n_{j+1}} (\mathbf{X}_{j+1})_i \stackrel{(2.41)}{=} \frac{1}{2} \sum_{i=1}^{n_j/2} (\mathbf{X}_j)_{2i-1} + (\mathbf{X}_j)_{2i} \\ &= \frac{1}{2} [(\mathbf{X}_j)_1 + (\mathbf{X}_j)_2 + \cdots + (\mathbf{X}_j)_{n_j}] = \underbrace{\frac{n_j}{2}}_{=n_{j+1}} \overline{X}_j = n_{j+1}\overline{X}_j. \end{aligned} \quad (2.45)$$

And so by repeated use of this equation we get $V(\overline{X}_i) = V(\overline{X}_0) = V(\overline{X})$ for all $0 \leq i \leq d-1$. This has the consequence that

$$V(\overline{X}) = \frac{\sigma_k^2}{n_k} + e_k \quad \text{for all} \quad 0 \leq k \leq d-1. \quad (2.46)$$

Flyvbjerg and Petersen (1989) claims that the sequence $\{e_k\}_{k=0}^{d-1}$ is decreasing, and conjecture that the term e_k can be made as small as we would like by making k (and hence d) sufficiently large. The sequence is decreasing because it is possible to show (as we will in proposition 6) that γ_k converges uniformly to the zero-function on \mathbb{N} . That means we can apply blocking transformations until e_k is sufficiently small, and then estimate $V(\overline{X})$ by $\hat{\sigma}_k^2/n_k$. Moreover, since γ_k converges uniformly to the zero function on \mathbb{N} , see figure 2.8 for an illustration.

It is natural then to think the best estimate we could then make is $\hat{\sigma}_{d-1}^2/n_{d-1}$. One could expect this because if $k = d-1$, then the truncation error $e_{d-1} \leq e_k$ for all $k \geq d-1$. But there is a problem with this, as we shall see next: If $k \rightarrow d-1$, then $V(\hat{\sigma}_{d-1}^2/n_{d-1})$ grows to an appreciable size, and so the standard error of $\hat{\sigma}_{d-1}^2/n_{d-1}$ can become unacceptably large. In that case, $\hat{\sigma}_{d-1}^2/n_{d-1}$ is not very useful (Flyvbjerg and Petersen 1989).

The idea of Flyvbjerg and Petersen (1989) is that if the conditions of the central limit theorem for dependent random variables are satisfied, see theorem 9, then the components of \mathbf{X}_k are asymptotically independent identically normal distributed as k grows. The asymptotic independence follows by theorem 3 because each elements of \mathbf{X}_k is the mean of random variables which are asymptotic uncorrelated, as we explained above. In this case, it is immediate by theorem 5 that

$$\begin{aligned} V\left(\frac{\hat{\sigma}_k^2}{n_k}\right) &= 2\frac{\sigma_k^4}{n_k^4}(n-1) = \underbrace{\left(\frac{\sigma^2}{n_k}\right)^2}_{(V(\overline{X})-e_k)^2} 2\frac{n_k-1}{n_k^2} = \left(V(\overline{X})-e_k\right)^2 \underbrace{\frac{2}{n_k}\left(1-\frac{1}{n_k}\right)}_{\geq 1/n_k} \\ &\geq \left(V(\overline{X})-e_k\right)^2 \frac{2}{n_k^2}, \end{aligned} \quad (2.47)$$

since the truncation error e_k is decreasing toward zero, this shows that the standard error of $\hat{\sigma}_k^2/n_k$ grows as n_k decreases. This may seem as bad news, because then the estimate of $V(\bar{X})$ has a relatively large error due to the standard error of $\hat{\sigma}_k^2/n_k$ even though the truncation error e_k is small. The question then is, how can we find the ideal k such that essentially $e_k = 0$, but simultaneously ensure that the standard error of $\hat{\sigma}_k^2/n_k$ is as small as possible? Assume $j \geq k$, and $e_k = 0$, then also e_j must be zero and so it is possible to interpret from Flyvbjerg and Petersen (1989) that

$$\begin{aligned} 0 = |0| &= |V(\bar{X}) - V(\bar{X})| = |V(\bar{X}_0) - V(\bar{X}_0)| \stackrel{(2.45)}{=} |V(\bar{X}_j) - V(\bar{X}_k)| \\ &\stackrel{(2.43)}{=} \left| \frac{\sigma_j^2}{n_j} + \underbrace{e_k}_{=0} - \frac{\sigma_k^2}{n_k} - e_k \right| = \left| \frac{\sigma_j^2}{n_j} - \frac{\sigma_k^2}{n_k} \right|. \end{aligned} \quad (2.48)$$

That means that there could be some point on the graph of σ_k^2/n_k where σ_k^2/n_k become constant. See figure 2.8 for a demonstration of what this looks like in practice. Manual blocking is exactly this, to plot the graph of σ_k^2/n_k against k , and hope to find a stationary point where σ_k^2/n_k becomes constant. The stationary point on this graph is where you stop increasing k .

Despite using a few results from the present thesis to justify the claims, the above is my interpretation of Flyvbjerg and Petersen (1989). As you can see, it is not at all mathematically precise, but the results of this thesis will fix this and proposes an automated way to estimate $V(\bar{X})$.

2.6 Physical models

2.6.1 Ising model and Metropolis algorithm

The Ising model is a classic in simulations of ferromagnetism (Ising 1925), and has been used much in computational physics and statistical mechanics (Upton and Cook 2014). So much that Newman and Barkema (1999) calls it the most thoroughly researched model in the whole of statistical physics. It is common to visualize the model as a grid, here represented by an $L \times L$ matrix M , of boolean variables with values $+1$ and -1 . So letting m_{ij} denote the elements of M , then $m_{ij} \in \{1, -1\}$ for all $1 \leq i, j \leq L$. The idea is that the boolean values represent the direction of the magnetic dipole moment at each site. As figure 2.9 shows, they are arranged in a grid such that there can be interaction with their neighbouring sites (Upton and Cook 2014). Let J be a constant that defines the interaction strength and define $[M]$ to be all pairs of 2-tuples of neighbouring elements from M . Then the energy is defined to be the sum of all products $-Jm_{ij}m_{kl}$ for $(i, j), (k, l) \in [M]$ (Newman and Barkema 1999). Consider $m_{ij}m_{i(j+1)}$ for example, then the product equals 1 if and only if m_{ij} and $m_{i(j+1)}$ has the same sign,

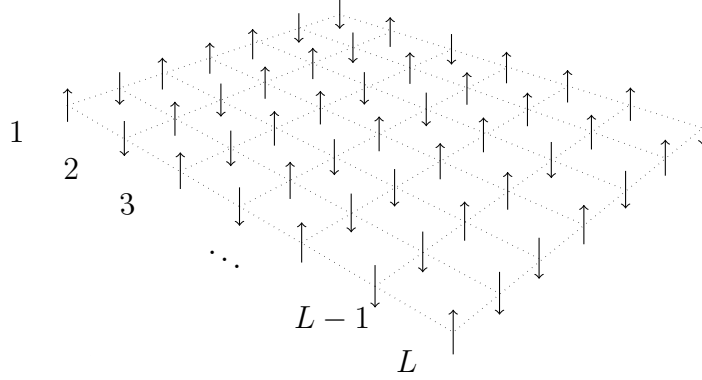


Figure 2.9: The Ising model is a textbook example of ferromagnetic simulation in statistical mechanics. It works by assuming that the system being studied is a grid of boolean variables representing the direction of the magnetic dipole moment. We assume that the magnetic moments interact with one and another. This is manifested in the way energy calculations are performed. Since there is interaction between the magnetic dipole moments and the grid is finite $L \times L$, there is a problem at the boundary and consequently, it is common to assume periodic boundary conditions.

and this gives the interaction. If $J = 0$, then clearly there is no interaction. If $J > 0$ then the interaction is assumed ferromagnetic, and antiferromagnetic if $J < 0$. It is clear that since M is an $L \times L$ matrix of boolean variables, it can be in exactly $n = 2^{L^2}$ configurations due to the product rule for L^2 -tuples (Devore and Berk 2012). So if S denotes the state space of M and let $i \in S$ denote any configuration of M , then S contains exactly n elements. Since the energy depends on the configuration of M , we let E_i denote the energy of such a configuration. The probability that M is in any configuration i is given by the **Boltzmann distribution**:

$$p(i; \beta) = \frac{1}{Z} \exp(-\beta E_i) \quad \text{where } \beta = (k_B T)^{-1}.$$

Here k_B is the Boltzmann constant, and T is the temperature (Schroeder 1999; Newman and Barkema 1999). The partition function Z is unknown, and normalizes the Boltzmann distribution. Assume that M is in configuration $i \in S$; whilst doing numerics, periodic boundary conditions on the lattice described by M are assumed since the grid is finite. Thus if $a \bmod b$ denotes the modulo operation (Fraleigh 2002), $S_L = \{1, 2, \dots, L\}$ is a section of the natural numbers then $S_L^2 = S_L \times S_L$ and

$$E_i = -J \sum_{(i,j) \in S_L^2} (m_{ij} m_{i(j+1 \bmod L)} + m_{ij} m_{(i+1 \bmod L)j}). \quad (2.49)$$

Despite Z being unknown, it is possible to generate a Markov chain that has the Boltzmann distribution as its stationary distribution by applying theorem 14 with

$C\pi_i = (1/Z) \exp(-\beta E_i)$ for all temperatures T . According to the discussion on page 50, the Markov chain should be generated as follows: Assume X_0 is given, then generate X_n by drawing j with probability q_{ij} from S given that $X_{n-1} = i$ and, afterwards making a transition with probability α_{ij} . According to Newman and Barkema (1999), it is common to let q_{ij} be a uniform distribution, so $q_{ij} = c$. The constant c must normalize q_{ij} , so it is determined using arithmetic series:

$$1 = \sum_{j=1}^{2^{L^2}} q_{ij} = \sum_{j=1}^{2^{L^2}} c = c \frac{1}{2} 2^{L^2} (2^{L^2} + 1) = c 2^{L^2-1} (2^{L^2} + 1)$$

$$\text{if and only if} \quad q_{ij} = c = \frac{2^{1-L^2}}{2^{L^2} + 1}.$$

The function α_{ij} is determined by inserting the expressions obtained, so

$$\alpha_{ij} = \min \left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}, 1 \right) = \min \left(\frac{(1/Z) \exp(-\beta E_j) c}{(1/Z) \exp(-\beta E_i) c}, 1 \right) = \min (e^{\beta(E_i - E_j)}, 1). \quad (2.50)$$

The last problem to overcome is to determine how to get a mechanism that ensures the transition happens with probability α_{ij} . Clearly, the number $\alpha_{ij} \in [0, 1]$, so consider the following lemma

LEMMA 2. *Suppose $X \sim \text{unif}(0, 1)$ is a uniform random variable on $[0, 1]$ and there is a real number $\alpha \in [0, 1]$. Then the probability that $X \leq \alpha$ is precisely α .*

PROOF. Suppose the hypothesis is true. Let $p(x)$ be the pdf of X . Since $X \sim \text{unif}(0, 1)$, the pdf of X is

$$p(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{else} \end{cases}.$$

Finally compute the probability that $X \leq \alpha$ using the definition of cumulative probability:

$$P(X \leq \alpha) = \int_{-\infty}^{\alpha} p(x) \, dx = \int_0^{\alpha} 1 \, dx = x \Big|_0^{\alpha} = \alpha,$$

as required. ■

After the state j has been chosen, α_{ij} is computed and compared to X . If $X \leq \alpha_{ij}$, then set $X_{n+1} = j$. If not, then according to the discussion of page 50, set $X_{n+1} = i$. The final algorithm is then:

1. Pick $X_0 \in S$

```

double E = S(A,m,n);
double M = sum(A,m,n);
for (int k = 1; k <= N; k++) {
    for (int i = 0; i < m; i++) {
        for (int j = 0; j < n; j++) {
            int u = rand_m(gen);
            int v = rand_n(gen);
            double dE = 2*A[u][v]*(A[u][mod(v+1,n)] +
                                   A[mod(u+1, m)][v] +
                                   A[u][mod(v-1, n)] +
                                   A[mod(u-1, m)][v]);
            if (exp(-beta*dE) >= randouble(gen) ) {
                A[u][v] = - A[u][v];
                E += dE;
            }
        }
    }
}

```

Figure 2.10: One implementation of the Ising model with Metropolis algorithm in C++. Here the matrix M is called A . The function $S()$ just computes the energy E_i for the initial state X_0 in accordance with equation (2.49). It is time efficient to not evaluate the sum of E_i at each iteration of the Ising model, instead just compute the difference between the current energy and the proposed energy, in the program, this variable is called dE , and is the difference $E_j - E_i$ from equation (2.50). Since the stationary distribution of $\{X_n\}$ is the Boltzmann distribution, the expected energy is estimated by computing the mean. For this reason any implementation of the algorithm requires that the energy is sampled at each iteration after an optional run-in period to ensure that the Markov chain is in the stationary distribution.

2. Pick $j \in S$ according to q_{ij}
3. Compute α_{ij}
4. Check if $X \leq \alpha_{ij}$. If yes: Set $X_{n+1} = j$. If no: Set $X_{n+1} = i$.
5. Set $n = n + 1$ and $i = j$
6. Go back to step (2) and repeat as many times as desired.

The discussion on page 50 showed that this ensures that $\{X_n\}$ is a time reversible Markov chain that has stationary probability $\pi_i = (1/Z)e^{-\beta E_i}$, according to theorem 14. See figure 2.10 for sample code of one C++-implementation.

2.6.2 Two dimensional N -electron quantum dot and Metropolis-Hasting algorithm

The applicability of theorem 14 to variational quantum methods is vast. For if a guess at the state function for an Hamiltonian is proposed, it is possible to generate samples of the observables of interest. This is done by sampling the observables under configurations $i \in S$, of the quantum system.

The guess at the state function is denoted by ψ_T and is called the **trail state function**. According to theorem 14, it is possible to let $C\pi = \psi_T$, since this ensures that the generated Markov chain $\{X_n\}$ takes values in the state space S and that $\{X_n\}$ will follow the stationary distribution ψ_T . The sampling of the observables is performed by evaluating the them under each X_k for all k . Once a random sample of the observables is available, the inference machinery take over and can produce estimates of the expected value and standard error of a whole host of functions thereof.

Specifically, for this thesis, the ground state energy of a quantum dot comprised

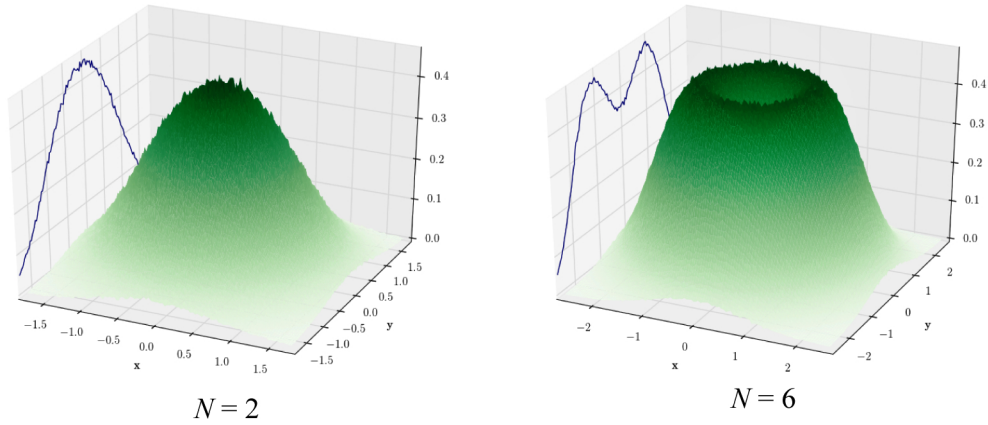


Figure 2.11: Reduced density functions (Szabo and Ostlund 1996) for two dimensional N -electron quantum dot at frequency $\omega = 1$ for a theory of a harmonic oscillator potential with Coulomb repulsion. It is due to Høgberget (2013).

of N electrons was sampled to be used for testing in the results. Assume that $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between electron i and j , then the theory is

$$E_T = \sum_{i=1}^N \left[-\frac{1}{2} \frac{\nabla_i^2 \psi_T}{\psi_T} + \frac{1}{2} \omega^2 r_i^2 + \sum_{j=i+1}^N \frac{1}{r_{ij}} \right], \quad (2.51)$$

$$\psi_T = C\psi_0\psi_C, \quad \psi_0(\mathbf{r}_1, \mathbf{r}_2) = \exp\left(-\frac{1}{2}\alpha\omega(r_1^2 + r_2^2)\right), \quad (2.52)$$

$$\psi_C(\mathbf{r}_1, \mathbf{r}_2) = \exp\left(\frac{r_{12}}{\beta r_{12} + 1}\right). \quad (2.53)$$

The parameters α and β are the *variational parameters*, and were assumed given for each ω of interest. In general, these can be estimated by the use of an iterative optimization algorithm such as the method of gradient descent. However, for the results of this thesis, the technicalities of how they are obtained is not important. According to theorem 14, the entire Monte Carlo scheme is specified once π_i and q_{ij} are specified. Suppose $\mathbf{r}_k^{(i)}$ denotes the coordinates of electron k in state number $i \in S$. That means $C\pi_i$ is equal to $\psi_T(\mathbf{r}_1^{(i)}, \mathbf{r}_2^{(i)}, \dots, \mathbf{r}_n^{(i)})$ for all $i \in S$. The trail state function is assumed to be a Slater-Jastrow product (Hammond, Lester Jr., and Reynolds 1994). According to Hammond, Lester Jr., and Reynolds (1994), this is desirable since it exhibits the correct cusp behavior for electrons and it means that the variational parameters has simple physical interpretation. Specifically, suppose $i, j \in \{0, 1, \dots, \frac{n}{2} - 1\}$ and define

$$\psi_T = \det(D_\uparrow) \det(D_\downarrow) \psi_C, \quad \psi_C(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n; \beta) = \exp\left(\sum_{i=1}^n \sum_{j=i+1}^n \frac{\gamma_{ij} r_{ij}}{1 + \beta r_{ij}}\right) \quad (2.54)$$

such that $(D_\uparrow)_{ij} = \phi_{2j}(\mathbf{r}_{2i}; \alpha)$, $(D_\downarrow)_{ij} = \phi_{2j+1}(\mathbf{r}_{2i+1}; \alpha)$

$$\text{and} \quad \phi_j(\mathbf{r}_i; \alpha) = H_{n_{xj}}((\alpha\omega)^{1/2}x_i) H_{n_{yj}}((\alpha\omega)^{1/2}y_i) \exp\left(-\frac{1}{2}\alpha\omega r_i^2\right) \quad (2.55)$$

j	0	1	2	3	4	5	6	7	8	9	10	11	...	
n_{xj}	0	0	1	1	0	0	2	2	1	1	0	0	...	for all
n_{yj}	0	0	0	0	1	1	0	0	1	1	2	2	...	
Spin	\uparrow	\downarrow	\uparrow	\downarrow	\uparrow	\downarrow	\uparrow	\downarrow	\uparrow	\downarrow	\uparrow	\downarrow	...	
$i \in \{0, 1, 2, \dots, n-1\}.$														

In the previous section, the Metropolis algorithm was used in conjunction with the Ising model, there the transition probability q_{ij} was assumed uniform. In contrast, the use of a non-uniform transition probability matrix is called *importance sampling*. Here, the transitional probability q_{ij} will come from a stationary isotropic diffusion process. According to Hammond, Lester Jr., and Reynolds (1994), it is immediate from the Fokker-Planck equation (Kampen 2007) that if $\mathbf{r}_k^{(i)}, \mathbf{r}_k^{(j)}$ represent two positions of particle number k and $\Delta t > 0$ is a parameter, then for all $k \in \{1, 2, \dots, n\}$, such a process must satisfy

$$q_{ij} = G(\mathbf{r}_k^{(i)}, \mathbf{r}_k^{(j)}; \Delta t) = \frac{1}{2\pi\Delta t} \exp\left(-\frac{\|\mathbf{r}_k^{(i)} - \mathbf{r}_k^{(j)} - \Delta t(\nabla\psi_T/\psi_T)\|^2}{2\Delta t}\right). \quad (2.56)$$

According to section 2.2.2, this is the binormal probability distribution with mean $\boldsymbol{\mu} = \Delta t \nabla \psi_T / \psi_T$ and covariance matrix $\Sigma = \Delta t I_2$. Since the covariance matrix $\Sigma = \Delta t I_2$ is symmetric, positive semi-definite (according to proposition 2), it has a unique square root $(\Delta t I_2)^{1/2}$ (Rynne and Youngson 2007). Moreover, since $I_n = I_n^{1/2}$, that means if $\boldsymbol{\xi} \sim N(\mathbf{0}, I_2)$ is some standard binormal random variable then

$$\mathbf{r}_k^{(i)} - \mathbf{r}_k^{(j)} = \Sigma^{1/2} \boldsymbol{\xi} + \boldsymbol{\mu} = (\Delta t)^{1/2} \boldsymbol{\xi} + \Delta t \nabla \psi_T / \psi_T. \quad (2.57)$$

This equation can be used to quickly generate new positions drawn from q_{ij} simply by generating standard binormal random variables and scaling them accordingly. For computing the transition probability α_{ij} from theorem 14, use equation (2.56). The final required ingredient is an expression for the observable evaluated for the chosen value of N . For the purpose of the results, it suffices to assume $N = 2$. And since analytical expressions for ψ_T and E_T are available from equations (2.53) and (2.51), it is straight forward to get an expression for the statistic of interest for $N = 2$. The result is:

$$E_T = \frac{1}{r_{12}} + \frac{1}{2} \omega^2 (1 - \alpha^2) (r_1^2 + r_2^2) + 2\alpha\omega + \frac{\alpha\omega r_{12}}{(1 + \beta r_{12})^2} - \frac{(1 + r_{12} - \beta^2 r_{12}^2)}{r_{12}(1 + r_{12}\beta)^4}.$$

These are all the expressions that are required. To summarize, the idea is then the following: First pick $X_0 \in S$, in other words, choose some starting position for the electrons. Suppose this state is $X_0 = i$. Next pick a state j from q_{ij} , meaning that we draw a standard bivariate normal random variable $\boldsymbol{\xi}$ using equation (2.57). After the state j has been chosen, compute α_{ij} and compare it to a uniformly distributed random variable $X \sim \text{unif}(0, 1)$, this is just the normal use of lemma 2. In order to compute α_{ij} , it is necessary to evaluate ψ_T under the new position. If $X \leq \alpha_{ij}$, then set $X_{n+1} = j$. If not, then according to the discussion of page 50, set $X_{n+1} = i$. Repeat the steps $m \in \mathbb{N}$ times. The final algorithm is then:

1. Pick $X_0 \in S$ and sample E_T under X_0 .
2. Pick $j \in S$ according to q_{ij}
3. Compute α_{ij}
4. Check if $X \leq \alpha_{ij}$. If yes: Set $X_{n+1} = j$. If no: Set $X_{n+1} = i$.
5. Sample E_T under X_{n+1} .
6. Set $n = n + 1$ and $i = j$
7. Go back to step (2) and repeat m times.

By letting `m` denote the number of Monte-Carlo cycles, `r` and `rpp` denote the new and old position matrices for the two particles then a simple program using importance sampling for two particles is quickly built using the `Armadillo` library. See figure 2.12 for the final implementation.

```

for (int i = 0; i < m; i++) {
    // selecting particle to move from uniform distribution
    int k = rand_particle(gen);
    int not_k = (k+1) % 2;

    // computing qij and qji (the importance sampling)
    rijpp = norm(rpp.col(0)-rpp.col(1));
    vec Fpp = -2*a*w*rpp.col(k) + 2*c*(rpp.col(k) -
        rpp.col(not_k))/( (1 + b*rijpp)*(1 + b*rijpp)*rijpp );
    r.col(k) = rpp.col(k) + D*Fpp*dt + randn<vec>(2)*sqrt(dt);
    rij = norm(r.col(0)-r.col(1));
    F = -2*a*w*r.col(k) + 2*c*(r.col(k) - r.col(not_k))/(
        (1 + b*rij)*(1 + b*rij)*rij );
    vec p = rpp.col(k) - r.col(k) - D*dt*F;
    vec q = r.col(k) - rpp.col(k) - D*dt*Fpp;
    double qji = exp(- dot(p,p)/(4*D*dt));
    double qij = exp(- dot(q,q)/(4*D*dt));

    // computing new state function
    wf = exp(-0.5*a*w*( dot(r.col(0),r.col(0)) +
        dot(r.col(1),r.col(1)) ) + r12/(1 + b*r12) );

    // one iteration of Hastings-Metropolis theorem
    if ( wf*wf*qji/(wfpp*wfpp*qij) > rand_double(gen) ) {
        rpp = r; wfpp = wf; rij = norm(r.col(0) - r.col(1) );
    }

    // sample energy
    double e = 1/rij + 0.5*w*w*(1-a*a)*(
        dot(rpp.col(0),rpp.col(0)) +
        dot(rpp.col(1),rpp.col(1)) )
        + 2*a*w + a*w*c*rij/pow(1 + rij*b,2) -
        c*(1+rij*c-b*b*rij*rij)/( rij*pow(1 + rij*b,4)
        );
    E += e/iterations;
}

```

Figure 2.12: Program for Hastings-Metropolis-style implementation of a quantum dot of two electrons. The variational parameters were $\alpha = 0.988664$ and $\beta = 0.397451$ and denoted by **a** and **b** in the program. The variables **r**, **rpp** contain the positions at sequential steps, **F**, **Fpp** contain the quantity $\nabla\psi_T/\psi_T$ for sequential steps and **rij**, **rijpp** contain the distance between electrons at sequential steps. The program returns the energy $E_T = 3.00052 \pm 10^{-5}$ on rejection ratio $r < 10^{-4}$ for the importance sampling and $2^{20} \approx 10^6$ samples. For comparison, the exact answer is known and equals 3.

Chapter 3

Results: Deriving an automatic blocking method

3.1 The plan and preliminaries

We have all the tools required to prove the blocking method, which will serve as an alternative to dependent bootstrapping, since we have discussed that bootstrapping stops working for sufficiently large time series. Recall some of the things we now know.

We know about functions on metric spaces and multivariate probability theory. With regards to time series, we know of weak stationarity, independence, measure of dependence. We know about blocking transformations, and have intuition of the behavior of the method. Using this, theorem 17 will follow.

We also know univariate probability theory, statistical inference, hypothesis testing and how to compute the covariance of quadratic forms. We know about the multinormal- and chi-square distribution. Moreover, we know how to form a chi-square random variable from multinormal random variables and maximum likelihood estimation. We are also familiar with Ibragimov's theorem. And last but not least, we now have theorem 17. This is all we will need to prove theorem 18.

We have some knowledge of autoregressive models, Markov chains, the Hastings-Metropolis' theorem, and physics applications thereof, namely the Ising model and two-dimensional N -electron quantum dots. These will be used to understand the appropriate method validation and extensions to ultra large time series which also follows.

That means nearly all of the things that have been introduced will now come

together to achieve our goals. This also shows that chapter 2 fulfilled our goal of being self-contained yet theoretically economical. Our knowledge of Bayesian- and shrinkage estimation and resampling methods will be useful in the discussion, after which we are nearly done.

3.2 Proof of the blocking method

Section 3.3 states the idea of the announced algorithm. However, in order to understand why the algorithm works, some 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 that will be frequently used in this context, it also justifies that γ_k exists for all integers $k \geq 0$:

LEMMA 3. *Let X_1, X_2, \dots be a stationary time series and \mathbf{X} be the vector of the first $n = 2^d$ sequential observations from X_1, X_2, \dots . Suppose \mathbf{X}_k are the n_k first observations of the time series $(X_k)_1, (X_k)_2, \dots$. Then both $(X_k)_1, (X_k)_2, \dots$ and \mathbf{X}_k are stationary. Moreover, if \bar{X}_k is the sample mean of \mathbf{X}_k , then $\bar{X} = \bar{X}_k$ for all $0 \leq k \leq d - 1$.*

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

$$\langle (X_{k+1})_i \rangle \stackrel{(1.2)}{=} \frac{1}{2} \langle (X_k)_{2i-1} + (X_k)_{2i} \rangle = \frac{1}{2} (\mu + \mu) = \mu.$$

For the covariance; the elements of $\{X_i\}$ are stationary, and therefore $\text{Cov}(X_i, X_j)$ 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 equation (2.42), it is true for $k + 1$, since equation (2.42) 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 $k \geq 0$. The proof works for any smaller time series $\{(X_k)_i\}_{i=a}^b$ for $a \geq 1$. By taking $b = n_k$ this proves \mathbf{X}_k is stationary.

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

$$n_{k+1} \bar{X}_{k+1} = \sum_{i=1}^{n_{k+1}} (X_{k+1})_i \stackrel{(1.2)}{=} \frac{1}{2} \sum_{i=1}^{n_k/2} [(X_k)_{2i-1} + (X_k)_{2i}] = \frac{1}{2} n_k \bar{X}_k = n_{k+1} \bar{X}_k,$$

which provides the induction step. ■

Using lemma 3 and equation (2.43) it is clear that any estimate of $V(\overline{X})$ using σ_k^2/n_k has *truncation error* given by

$$e_k \equiv \frac{2}{n_k} \sum_{h=1}^{n_k-1} \left(1 - \frac{h}{n_k}\right) \gamma_k(h). \quad (3.1)$$

The next proposition is crucial, it says that if $\gamma_0(1), \gamma_1(1), \dots, \gamma_{d-1}(1)$ are known, then the behavior of the truncation error e_k is known.

PROPOSITION 5. *Suppose $2^d \geq 2$ is the number of observations, and σ_k^2 is finite for all $k \in \{0, 1, \dots, d-1\}$. Then the rate of change of the truncation error e_k is*

$$e_k - e_{k+1} = \frac{\gamma_k(1)}{n_k} \quad \text{for all} \quad 0 \leq k < d-1. \quad (3.2)$$

To prove the proposition, we sum each side of equation (2.42) to get

$$\sum_{h=1}^{n_{k+1}-1} \gamma_{k+1}(h) = \frac{1}{2} \sum_{h=1}^{n_k-1} \gamma_k(h) - \frac{1}{4} [\gamma_k(1) + \gamma_k(n_k-1)]. \quad (3.3)$$

Similarly, summing equation (2.42), we arrive at

$$\sum_{h=1}^{n_{k+1}-1} h \gamma_{k+1}(h) = \frac{1}{4} \sum_{h=1}^{n_k-1} h \gamma_k(h) - \frac{n_k}{8} \gamma_k(n_k-1). \quad (3.4)$$

Plugging these equations into the definition of e_k given in (3.1) and using $n_{k+1} = n_k/2$, it is immediate that

$$\begin{aligned} e_{k+1} &= \frac{2}{n_{k+1}} \sum_{h=1}^{n_{k+1}-1} \left(1 - \frac{h}{n_{k+1}}\right) \gamma_{k+1}(h) \\ &= \frac{4}{n_k} \sum_{h=1}^{n_{k+1}-1} \gamma_{k+1}(h) - \frac{8}{n_k^2} \sum_{h=1}^{n_{k+1}-1} h \gamma_{k+1}(h) \stackrel{(3.3)(3.4)}{=} e_k - \frac{\gamma_k(1)}{n_k}, \end{aligned}$$

The following corollary is the most important take away message. 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 & Petersen (1989) blocking method. Explanation of this is provided in the discussion.

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

1. if there exists $k \in \mathbb{N}$ such that for all $i \leq k \leq j$ either: $\gamma_k(1) > 0$ or $\gamma_k(1) \geq 0$ or $\gamma_k(1) = 0$, then the sequence of errors e_k is strictly decreasing or decreasing or constant on $i \leq k \leq j$, respectively.
2. if there exists some $k \in \{0, 1, \dots, d-1\}$ 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 $j \geq k$.

PROOF. Suppose the hypothesis is true and first let $\gamma_k(1) > 0$. That means proposition 5 is true and there exist $k \in \mathbb{N}$ such that $\gamma_k(1) > 0$ for all $i \leq k < j$. If $u, v \in \{i, i+1, \dots, j+1\}$ are distinct natural numbers, we 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

$$\begin{aligned}
 0 &< \sum_{k=u}^{v-1} \frac{\gamma_k(1)}{n_k} = \frac{\gamma_u(1)}{n_u} + \frac{\gamma_{u+1}(1)}{n_{u+1}} + \dots + \frac{\gamma_{v-1}(1)}{n_{v-1}} \\
 &\stackrel{(3.2)}{=} (e_u - e_{u+1}) + (e_{u+1} - e_{u+2}) + \dots + (e_{v-1} - e_v) = e_u - e_v.
 \end{aligned}$$

Now, by adding e_v to each side of the inequality, the first part is proved. To obtain the result in the case $\gamma_k(1) \geq 0$, replace $<$ with \leq 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, we 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. The induction step follows for $k+1$ since equation (2.42) says that $\gamma_{k+1}(i) = \delta_{i0}\sigma_k^2/2$. This proves $\gamma_j(1)$ is zero for all $j \geq k$, so the error is constant by what was proved above. ■

These results will be useful in the automation of the blocking method later in the thesis. And as stated, the corollary proves the behavior of the blocking method. But attentive readers will spot a problem: The sequence e_k may be decreasing and eventually constant if the elements of \mathbf{X}_k become uncorrelated as k increases. But there is no guarantee that the variables become uncorrelated. However, prior users of the 1989 blocking method know the variables do indeed become uncorrelated (and the constant from part 2 of the corollary is zero). So far this is not guaranteed. Although our present results are promising and hint at the conclusions to come, more work is required. We start by a lemma and some interesting consequences of "blocking". In doing so, it is best to introduce an interesting sequence of functions $\{f_k\}$. Fix some integer $k \geq 1$ and define:

$$f_k(i) = \begin{cases} i & \text{if } 0 \leq i \leq 2^k \\ 2^{k+1} - i & \text{if } 2^k \leq i \leq 2^{k+1} \\ 0 & \text{else} \end{cases} \quad (3.5)$$

The sequence $\{f_n\}$ has some nice properties:

LEMMA 4. *The sequence $\{f_k\}$ satisfy the following properties:*

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

PROOF. See the appendix. ■

LEMMA 5. *Suppose X_1, X_2, \dots is a stationary time series and h and k are positive natural numbers, then*

$$\gamma_k(h) = 2^{-2k} \sum_{i=1}^{2^{k+1}-1} f_k(i) \gamma(2^k(h-1) + i). \quad (3.6)$$

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

$$\gamma_1(h) \stackrel{(1.2)}{=} 2^{-2} \left(\gamma_0(2h-1) + 2\gamma_0(2h) + \gamma_0(2h+1) \right) = 2^{-2k} \sum_{i=1}^{2^{k+1}-1} f_k(i) \gamma(2^k(h-1) + i).$$

Assume now that equation (3.6) is true for some $k \geq 1$ and write

$$\begin{aligned} \frac{\gamma_{k+1}(h)}{2^{2(k+1)}} &\stackrel{(1.2)}{=} \frac{2^{-2}}{2^{2(k+1)}} [\gamma_k(2h-1) + 2\gamma_k(2h) + \gamma_k(2h+1)] \\ &\stackrel{(3.6)}{=} \sum_{i=1}^{2^{k+1}-1} f_k(i) \gamma(2^k(2h-2) + i) + 2 \sum_{i=1}^{2^{k+1}-1} f_k(i) \gamma(2^k(2h-1) + i) \\ &\quad + \sum_{i=1}^{2^{k+1}-1} f_k(i) \gamma(2^k(2h) + i) \\ &\stackrel{(3.5)}{=} \sum_{i=1}^{2^{k+1+1}-1} \gamma(2^{k+1}(h-1) + i) [f_k(i) + 2f_k(i - 2^k) + f_k(i - 2^{k+1})] \\ &\stackrel{\text{lemma 4}}{=} \sum_{i=1}^{2^{k+1+1}-1} f_{k+1}(i) \gamma(2^{k+1}(h-1) + i). \end{aligned}$$

In the third equality, the summation limits were shifted by 0, 2^k and 2^{k+1} respectively, in addition we was used that $f_n(i - 2^j) = 0$ whenever $i \leq 2^j$ or $2^{k+1} + 2^j \leq i$ from equation (3.5). This allowed to factor out the term $\gamma(2^{k+1}(h-1) + i)$. ■

Proposition 5 shows that $\gamma_k(1)$ is of special interest to us. Consider the following corollary

COROLLARY 2. Suppose X_1, X_2, \dots is a stationary time series and k is a positive natural number, then

$$2^{2k}\gamma_k(1) = \gamma(1) + 2\gamma(2) + \dots + 2^k\gamma(2^k) + (2^k - 1)\gamma(2^k + 1) + \dots + \gamma(2^{k+1} - 1). \quad (3.7)$$

PROOF. Use the previous lemma with $h = 1$. ■

Using these results, everything is now set for a technical proposition. The statement reveals that a theorem is close.

PROPOSITION 6. Assume that the stationary time series X_1, X_2, \dots has autocovariance $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$. Then $\{\gamma_k\}_{k=1}^\infty$ converges uniformly to the zero-function on \mathbb{N} .

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

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

since \log_2 is a monotonous function. Thus by lemma 5 and the triangular inequality

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

where lemma 4 was used in the third step. By construction, it is possible to assume $k \geq K \geq (1/2)\log_2(2S/\varepsilon)$, so $2^{-2k}S \leq \varepsilon/2$. Assume now that $h = 1$. Then by lemmas 4 and 5 and the triangular inequality:

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

which was required. ■

THEOREM 17 (The blocking method). *Assume that the stationary time series X_1, X_2, \dots has autocovariance $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$. Then for every $\varepsilon > 0$ there exists a natural number K such that $e_k < \varepsilon$ if $K \leq k \leq d$ for the time series X_1, X_2, \dots, 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 6, there exists $K \in \mathbb{N}$ such that if $k \geq K$ then $\gamma_k < \varepsilon/2$ on \mathbb{N} . Moreover, if $d \geq k$, then by the triangular inequality

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

which is the theorem. ■

3.3 Automating calculations

According to the variant of the central limit theorem from the methods, the elements of \mathbf{X}_j are asymptotically multivariate normal if $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ (in addition to some technical assumptions). This is because the elements of \mathbf{X}_k are means of the elements of \mathbf{X} . In that case, the methods say that, $\hat{\gamma}_j(1)$ is the maximum likelihood estimator of $\gamma_j(1)$. Hence if $\hat{\gamma}_j \equiv (\hat{\gamma}_j(1), \dots, \hat{\gamma}_{d-1}(1))$, then $\hat{\gamma}_j \sim N(\boldsymbol{\mu}, \Sigma)$ is asymptotically multivariate normal according to the methods. The idea is to find the first index j such that $\gamma_j(1) = 0$, because by corollary 1, the error e_j becomes constant and there is no reason to expect that σ_k/n_k is a better estimate than σ_j/n_j for any $k > j$. To test this, define

$$M_j = (\hat{\gamma}_j - \boldsymbol{\mu})^\top \Sigma_j^{-1} (\hat{\gamma}_j - \boldsymbol{\mu}) \sim \chi_{d-j}^2. \quad (3.8)$$

Hence $(\hat{\gamma}_j - \boldsymbol{\mu})^\top \Sigma_j^{-1} (\hat{\gamma}_j - \boldsymbol{\mu})$ has a known distribution (according to the methods), which means that it is a test statistic for the hypothesis test

$$\begin{aligned} H_0 : \gamma_j(1) &= 0 \text{ for all } j \geq k, \\ H_a : \text{There exists } k \geq j \text{ such that } \gamma_k(1) &\neq 0. \end{aligned} \quad (3.9)$$

The idea is to pick the smallest j such that the hypothesis test finds no evidence for H_a and take $V(\overline{X}) = \sigma_j^2/n_j$. This works according to the methods: Whenever H_0 is true, the distribution of M_j is known (chi square with $d - j$ degrees of freedom (dof) by equation 3.8), 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, i.e. that the error becomes constant, and iterating further does not improve the estimate. See

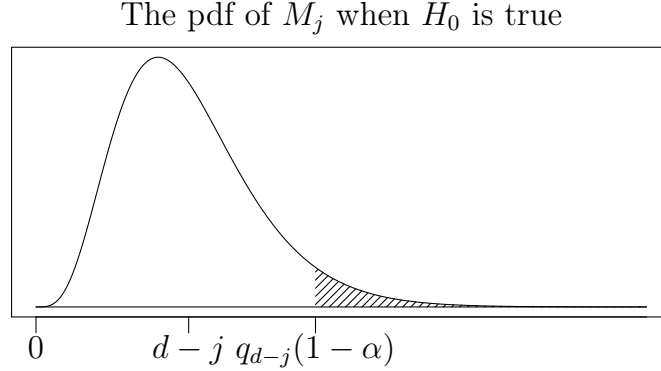


Figure 3.1: Whenever H_0 is true, the pdf of M_j is known and plotted above. The test concludes that H_0 is false if the observed value of M_j is sufficiently unlikely. That is, if the observed value of M_j is larger than 100(1- α)-percentile for a suitable α ; the shaded area represents a probability of α . The value $d-j = \langle M_j \rangle$ is the expected value of M_j whenever H_0 is true since then $M_j \sim \chi_{d-j}^2$ is chi square distributed.

figure 3.1 for illustration. However, an expression of Σ_j has to be determined. In this thesis, the following approximation will be used

$$\Sigma_j = \text{diag}(\sigma_j^4/n_j, \dots, \sigma_{d-j}^4/n_{d-j}). \quad (3.10)$$

The benefit is that inversion of Σ_j is easy. Corollary 3 explains why it is a reasonable approximation. But before proving this, more work is needed.

3.4 Covariance matrix of $\hat{\gamma}_i(\mathbf{1})$ and the matrix Σ_j

Computing the covariance matrix directly is impractical. However, developing linear algebra for the task provides a fruitful alternative. Two lemmas and two propositions are required. We start by considering the lemma that contains all the information required about probability distributions:

LEMMA 6. Assume $\mathbf{1}$ denotes the vector of ones, $\mathbf{X} \sim N(m\mathbf{1}, \sigma^2 I_n)$ and $\mathbf{Y} = \mathbf{X} - \bar{X}\mathbf{1}$. 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 $\Sigma_{\mathbf{Y}} = QQ^T$ and

$$\Sigma_{\mathbf{Y}} = \frac{\sigma^2}{n}(nI_n - \mathbf{1}\mathbf{1}^T). \quad (3.11)$$

PROOF. First note that Y_i is a linear combination of elements of \mathbf{X} , because \mathbf{X} is multivariate normal, that means 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 the methods. The expected value of \mathbf{Y} is $\mathbf{0}$ since $\langle Y_i \rangle = \langle X_i - \bar{X} \rangle = m - m = 0$.

To get equation (3.11), notice that the covariance matrix of \mathbf{X} is diagonal, that 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

$$\begin{aligned} (\Sigma_Y)_{ij} &= \text{Cov}(Y_i, Y_j) = \text{Cov}(X_i - \bar{X}, X_j - \bar{X}) = \\ &= \sigma^2 \delta_{ij} - \frac{1}{n} \sum_{k=1}^n \text{Cov}(X_i, X_k) - \frac{1}{n} \sum_{k=1}^n \text{Cov}(X_j, X_k) + V\bar{X}, \\ &= \sigma^2 \delta_{ij} - \frac{1}{n} \sum_{k=1}^n \sigma^2 \delta_{ik} - \frac{1}{n} \sum_{k=1}^n \sigma^2 \delta_{jk} + \frac{\sigma^2}{n} = \sigma^2 \delta_{ij} - \frac{\sigma^2}{n} \end{aligned} \quad (3.12)$$

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

$$\Sigma_Y \mathbf{q}_k \stackrel{(3.12)}{=} \frac{\sigma^2}{n} (n\mathbf{q}_k - \mathbf{1}(\underbrace{\mathbf{1}^\top \mathbf{e}_k - \mathbf{1}^\top \mathbf{e}_n}_{=1-1=0})) = \sigma^2 \mathbf{q}_k,$$

which proves that σ^2 is an eigenvalue of Σ_Y with multiplicity $n-1$. And since Σ_Y is symmetric, it has a spectral decomposition (Lay 2012):

$$\Sigma_Y = \sigma^2 \sum_{k=1}^{n-1} \mathbf{q}_k \mathbf{q}_k^\top = \sigma^2 [\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_{n-1}] [\mathbf{q}_1^\top \ \mathbf{q}_2^\top \ \dots \ \mathbf{q}_{n-1}^\top]^\top.$$

So if $Q = \sigma [\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_{n-1}]$ then Q is an $n \times (n-1)$ -matrix and $\Sigma_Y = QQ^\top$. Moreover, according to the spectral theorem (Lay 2012), the dimension of $\text{Span}\{\mathbf{q}_1, \dots, \mathbf{q}_{n-1}\}$ equals the multiplicity of σ^2 . Which means that the columns of Q are $n-1$ linearly independent vectors, which also equals its rank. ■

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

$$S_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & \dots & 0 \end{bmatrix} \quad T_i = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \quad (3.13)$$

According to equation (1.2), the matrices T_i generate the observations \mathbf{X}_k subject to i blocking transformations by

$$\mathbf{X}_i = T_i T_{i-1} \dots T_1 \mathbf{X}. \quad (3.14)$$

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

$$\Gamma_i = \frac{1}{2} \frac{1}{n_i} T_1^\top T_2^\top \cdots T_i^\top (S_i + S_i^\top) T_i \cdots T_1. \quad (3.15)$$

These matrices are interesting because they generate the estimator $\hat{\gamma}_i(1)$ from the vector \mathbf{Y} whose probability distribution is multivariate normal:

PROPOSITION 7. *The matrices $\{\Gamma_i\}$ are symmetric. Hence if $\mathbf{Y} = \mathbf{X} - \bar{X}\mathbf{1}$, then $\mathbf{Y}^\top \Gamma_i \mathbf{Y}$ is a quadratic form and $\mathbf{Y}^\top \Gamma_i \mathbf{Y} = \hat{\gamma}_i(1)$.*

PROOF. Fix $0 \leq i \leq d-1$. It's clear that Γ_i is symmetric by construction;

$$\begin{aligned} \Gamma_i^\top &= \frac{1}{2} \frac{1}{n_i} (T_1^\top T_2^\top \cdots T_i^\top (S_i + S_i^\top) T_i \cdots T_1)^\top \\ &= \frac{1}{2} \frac{1}{n_i} (T_i \cdots T_1)^\top (S_i + S_i^\top)^\top (T_1^\top T_2^\top \cdots T_i^\top)^\top = \Gamma_i. \end{aligned}$$

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

$$\begin{aligned} n_i \mathbf{Y}^\top \Gamma_i \mathbf{Y} &\stackrel{(3.15)}{=} \frac{1}{2} \mathbf{Y}^\top T_1^\top \cdots T_i^\top S_i T_i \cdots T_1 \mathbf{Y} + \frac{1}{2} [\mathbf{Y}^\top T_1^\top \cdots T_i^\top S_i T_i \cdots T_1 \mathbf{Y}]^\top \\ &= [T_i \cdots T_1 \mathbf{Y}]^\top S_i T_i \cdots T_1 \mathbf{Y} \stackrel{(3.14)}{=} \mathbf{Y}_i^\top S_i \mathbf{Y}_i. \end{aligned} \quad (3.16)$$

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

$$\begin{aligned} (\mathbf{Y}_{i+1})_k &\stackrel{(1.2)}{=} \frac{1}{2} [(\mathbf{Y}_i)_{2k-1} + (\mathbf{Y}_i)_{2k}] = \frac{1}{2} [(\mathbf{X}_i)_{2k-1} - \bar{X}_i + (\mathbf{X}_i)_{2k} - \bar{X}_i] \\ &\stackrel{(1.2)}{=} (\mathbf{X}_{i+1})_k - 2\frac{1}{2}\bar{X}_i \stackrel{\text{lemma 3}}{=} (\mathbf{X}_{i+1})_k - \bar{X}_{i+1}, \end{aligned} \quad (3.17)$$

where lemma 3 was used twice to get $\bar{X}_i = \bar{X} = \bar{X}_{i+1}$. Third, using the definition of the matrices S_i from equation (3.13) we notice that S_i shifts the indices of vectors by one:

$$\begin{aligned} n_i \mathbf{Y}^\top \Gamma_i \mathbf{Y} &\stackrel{(3.16)}{=} \mathbf{Y}_i^\top S_i \mathbf{Y}_i = \mathbf{Y}_i^\top (S_i \mathbf{Y}_i) = \left((\mathbf{Y}_i)_1, \dots, (\mathbf{Y}_i)_{n_i} \right)^\top \left((\mathbf{Y}_i)_2, \dots, (\mathbf{Y}_i)_{n_i}, 0 \right) \\ &= \sum_{h=1}^{n_i-1} (\mathbf{Y}_i)_{h+1} (\mathbf{Y}_i)_h \stackrel{(2.34)(3.17)}{=} n_i \hat{\gamma}_i(1), \end{aligned} \quad (3.18)$$

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

Attentive readers see that the above result is easily generalized to $\hat{\gamma}_i(k)$ for any

$k \geq 0$ by considering the operators, S_i^k , by raising to a power $k \in \mathbb{Z}$. But according to proposition 5 it suffices to consider $k = 1$. In this case, the following three quantities determine the expression for the covariance matrix Σ . Consider the following lemma

LEMMA 7. *If $\mathbf{1}$ denotes the vector of ones and $i \geq j$, then Γ_i and Γ_j constitute the following*

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

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 equation (3.13),

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

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

$$\begin{aligned} T_i \sum_{k=1}^{n_i-1} \mathbf{e}_k &= T_i \mathbf{e}_1 + T_i \mathbf{e}_2 + T_i \mathbf{e}_3 + \cdots + T_i \mathbf{e}_{n_i-1} \\ &\stackrel{(3.13)}{=} \frac{1}{2} \mathbf{e}_1 + \frac{1}{2} \mathbf{e}_1 + \frac{1}{2} \mathbf{e}_2 + \cdots + \frac{1}{2} \mathbf{e}_{n_i} = \sum_{k=1}^{n_i} \mathbf{e}_k. \end{aligned} \quad (3.19)$$

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

$$\begin{aligned} n_i \mathbf{1}^\top \Gamma_i \mathbf{1} &= \frac{1}{2} \sum_{u=1}^n \mathbf{e}_u^\top T_1^\top T_2^\top \cdots T_i^\top (S_i + S_i^\top) T_i \cdots T_1 \sum_{v=1}^n \mathbf{e}_v \stackrel{(3.19)}{=} \frac{1}{2} \sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \mathbf{e}_u^\top (S_i + S_i^\top) \mathbf{e}_v \\ &= \frac{1}{2} \sum_{u=1}^{n_i} \sum_{v=2}^{n_i} \mathbf{e}_u^\top \mathbf{e}_{v-1} + \frac{1}{2} \sum_{u=1}^{n_i} \sum_{v=1}^{n_i-1} \mathbf{e}_u^\top \mathbf{e}_{v+1} = n_i - 1, \end{aligned}$$

where orthonormality of $\{\mathbf{e}_k\}$ was used in the final step. Next it is necessary to show $T_i T_i^\top = (1/2) I_{n_i}$ for all $1 \leq i \leq d-1$. 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 (Hazewinkel 1993):

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

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

$$2nn_in_j\mathbf{1}^\top\Gamma_i\Gamma_j\mathbf{1} = 2n_in_j - n_i - 2n_j. \quad (3.20)$$

We prove now two more properties of $\{\mathbf{e}_k\}$: First, see that if M is any $n \times n$, then a diagonal element $m_{kk} = \mathbf{e}_k^\top M \mathbf{e}_k$, so $\text{Tr}(M) = \sum_{k=1}^n \mathbf{e}_k^\top M \mathbf{e}_k$. 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

$$\sum_{k=1}^{n_j-1} \mathbf{e}_k^\top T_{j+1}^\top \cdots T_{j+h}^\top M T_{j+h} \cdots T_{j+1} \mathbf{e}_{k+1} = 2^{-2h} \sum_{k=1}^{n_{j+h}-1} \mathbf{e}_k^\top M \mathbf{e}_{k+1} + K \sum_{k=1}^{n_{j+h}} \mathbf{e}_k^\top M \mathbf{e}_k. \quad (3.21)$$

We prove this by induction. If $h = 1$ then

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

which proves the base case. To get the induction step, assume the hypothesis is true for h then we define the matrix $N = T_{j+h+1}^\top M T_{j+h+1}$. 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, we use again that $T_j T_j^\top = 2^{-1} I_{n_j}$, as well as cyclic permutation of the factors and write $\text{Tr}[\Gamma_i \Gamma_j]$ in the following way:

$$4n_in_j \text{Tr}[\Gamma_i \Gamma_j] \stackrel{(3.15)}{=} 2^{-2j} \text{Tr} \left[T_{j+1}^\top \cdots T_i^\top (S_i + S_i^\top) T_i \cdots T_{j+1} (S_j + S_j^\top) \right] \quad (3.22)$$

Now we distribute the terms in the trace. One of the terms is $\text{Tr}[T_{j+1}^\top \cdots T_i^\top S_i T_i \cdots T_{j+1} S_j^\top]$. To evaluate it, we use what was just proved and write

$$\begin{aligned} \sum_{k=1}^{n_j} \mathbf{e}_k^\top T_{j+1}^\top \cdots T_i^\top S_i T_i \cdots T_{j+1} S_j^\top \mathbf{e}_k &= \sum_{k=1}^{n_j-1} \mathbf{e}_k^\top T_{j+1}^\top \cdots T_i^\top S_i T_i \cdots T_{j+1} \mathbf{e}_{k+1} \\ &\stackrel{(3.21)}{=} 4^{-(i-j)} \sum_{k=1}^{n_i-1} \mathbf{e}_k^\top S_i \mathbf{e}_{k+1} + K \sum_{k=1}^{n_i} \underbrace{\mathbf{e}_k^\top S_i \mathbf{e}_k}_{\mathbf{e}_k^\top \mathbf{e}_{k-1}=0}. \end{aligned}$$

This term equals $4^{-(i-j)}(n_i - 1)$ since $\mathbf{e}_k^\top S_i \mathbf{e}_{k+1} = \mathbf{e}_k^\top \mathbf{e}_k = 1$. We make the replacement $S_i \mapsto S_i^\top$ throughout the above equation, in which case the term

will evaluate to zero since $\mathbf{e}_k^\top S_i^\top \mathbf{e}_{k+1} = \mathbf{e}_k^\top \mathbf{e}_{k+2} = 0$. The third and fourth terms from equation (3.22) are evaluated in a similar way. The sum of all four terms is $2 \cdot 4^{-(i-j)}(n_i - 1)$, hence

$$n^2 n_i n_j \operatorname{Tr}[\Gamma_i \Gamma_j] \stackrel{(3.22)}{=} 2 \frac{1}{4} 2^{-2j} n^2 4^{-(i-j)} (n_i - 1) = \frac{1}{2} n_i^2 (n_i - 1),$$

which is the final part of the lemma. \blacksquare

PROPOSITION 8. *If there is some $m \in \mathbb{R}$ such that the vector $\mathbf{X} \sim N(m\mathbf{1}, \sigma^2 I_n)$, then the expected value of $\hat{\gamma}_i(\mathbf{1})$ is $-\sigma_i^2(n_i - 1)/n_i^2$. Furthermore, the covariance matrix of $(\hat{\gamma}_0(\mathbf{1}), \dots, \hat{\gamma}_{d-1}(\mathbf{1}))^\top$ has elements*

$$\operatorname{Cov}(\hat{\gamma}_i(\mathbf{1}), \hat{\gamma}_j(\mathbf{1})) = 2 \left(\frac{\sigma_i \sigma_j}{n_i n_j} \right)^2 \left[1 + (n_i - 1) \left(\frac{1}{2} n_i^2 - n_j \right) \right],$$

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

PROOF. Assume $0 \leq j \leq i \leq d - 1$. To obtain the expected value of $\hat{\gamma}_i(\mathbf{1})$, use the defining equation (2.34) and notice that the elements of \mathbf{X}_i are independent by hypothesis, so

$$n_i \langle \hat{\gamma}_i(\mathbf{1}) \rangle = \sum_{j=1}^{n_i-1} \underbrace{\langle (\mathbf{X}_i)_j (\mathbf{X}_i)_{j+1} \rangle}_{\gamma_i(\mathbf{1}) + m^2 = 0 + m^2} + \underbrace{\langle \bar{X}^2 \rangle}_{\frac{\sigma_i^2}{n_i} + m^2} - \underbrace{\langle ((\mathbf{X}_i)_j + (\mathbf{X}_i)_{j+1}) \bar{X} \rangle}_{2(m^2 + \sigma_i^2/n_i)},$$

and the first part is proved. Assume now that $\mathbf{Y} = \mathbf{X} - \bar{X}\mathbf{1}$. To get the covariance matrix, note that by lemma 6, \mathbf{Y} is multivariate normal with expected value $\boldsymbol{\mu}$ and there exist a $n \times (n - 1)$ -matrix Q of rank $n - 1$ such that the covariance matrix $\Sigma_{\mathbf{Y}} = QQ^\top$. According to proposition 7, Γ_i and Γ_j are symmetric, so according to the methods

$$\operatorname{Cov}(\hat{\gamma}_i(\mathbf{1}), \hat{\gamma}_j(\mathbf{1})) = \operatorname{Cov}(\mathbf{Y}^\top \Gamma_i \mathbf{Y}, \mathbf{Y}^\top \Gamma_j \mathbf{Y}) \stackrel{\text{Theorem 6}}{=} 2 \operatorname{Tr}(\Sigma_{\mathbf{Y}} \Gamma_i \Sigma_{\mathbf{Y}} \Gamma_j). \quad (3.23)$$

Recall that it is possible to cyclically permute the elements of a trace and that $\Gamma_i = \Gamma_j^\top$, and $\operatorname{Tr}(M) = \operatorname{Tr}(M^\top)$ and $\operatorname{Tr}(\mathbf{1}^\top M \mathbf{1}) = \mathbf{1}^\top M \mathbf{1}$ (since it is a real number) for all square matrices M . Using this and lemma 6, we write $(n^2/\sigma^4) \operatorname{Tr}(\Sigma_{\mathbf{Y}} \Gamma_i \Sigma_{\mathbf{Y}} \Gamma_j)$ in the following way

$$\frac{n^2}{\sigma^4} \operatorname{Tr}(\Sigma_{\mathbf{Y}} \Gamma_i \Sigma_{\mathbf{Y}} \Gamma_j) = n^2 \operatorname{Tr}(\Gamma_i \Gamma_j) + \mathbf{1}^\top \Gamma_i \mathbf{1} \mathbf{1}^\top \Gamma_j \mathbf{1} - 2n \mathbf{1}^\top \Gamma_i \Gamma_j \mathbf{1}.$$

To complete the proof, we use lemma 7, the definition of $n_i = n/2^i$. Furthermore, since the elements of \mathbf{X} are independent, $\sigma_j^2 = \sigma^2/2^j$ by corollary 1. Thus

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

We then multiply each side of the equation by 2 and recall equation (3.23) above, which proves the proposition true. \blacksquare

COROLLARY 3. *Assume there is some $m \in \mathbb{R}$ such that the vector $\mathbf{X} \sim N(m\mathbf{1}, \sigma^2 I_n)$. Then the covariance matrix of γ_j is*

$$\Sigma_j = \text{diag}(\sigma_j^4/n_j, \dots, \sigma_{d-j}^4/n_{d-j})$$

to leading order in $1/n_k$ and the variance of $\hat{\gamma}_j(k)$ is exactly

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

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

3.5 Algorithm

Summarizing the results from the past two subsections yields the following theorem

THEOREM 18. *If X_1, X_2, \dots is a strictly stationary time series such that $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ with $\lim_{n \rightarrow \infty} V(\sum_{i=1}^n X_i) = \infty$ and $\langle |X_i|^{2+a} \rangle < \infty$, for some $a > 0$ then M_j is a test statistic which is asymptotically χ_{d-j}^2 -distributed under the hypothesis $\gamma_k(1) = 0$ for all $k \geq j$. The rejection region are all values M_j larger than $q_{d-j}(1 - \alpha)$ for all $1 \leq j \leq d - 1$.*

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

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

We pick some significance level α ; it is common 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 $M_k \leq q_{d-k}(1 - \alpha)$. Using this k , make the final estimate for the variance $V(\bar{X}) = \hat{\sigma}_k^2/n_k$.

The method has built in safety features (see figure 3.5). This is necessary because the method shall 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 18): (i) the time series is strictly stationary, (ii) $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ and (iii) the χ^2 -approximation works. Therefore

if the method does not conclude that H_0 is true for any $k \geq d - 1$, 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 is little data available. See figure 3.5 for sample code of one implementation, or use flow chart of 3.2 for an overview.

Flow chart of algorithm

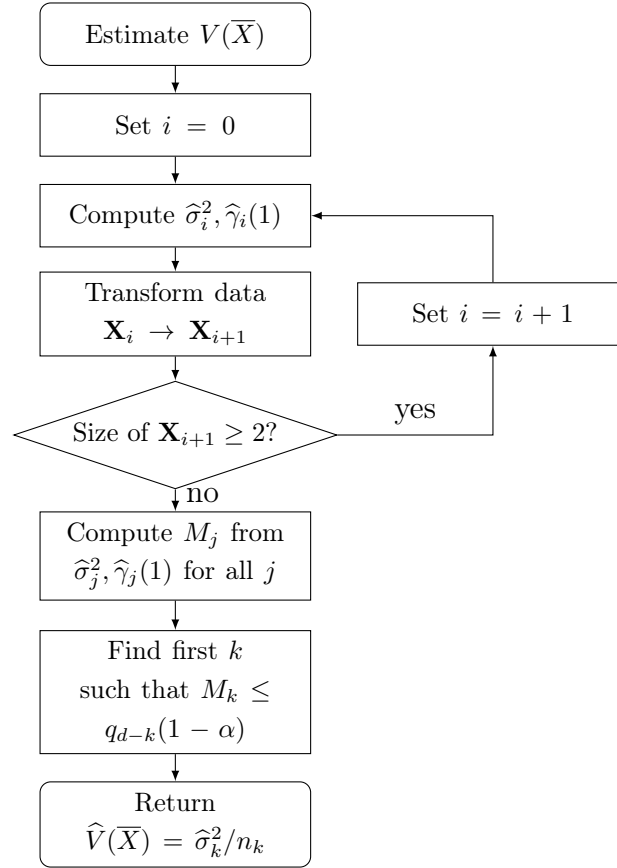


Figure 3.2: Flow chart of algorithm. The idea is to return the estimate of $V(\bar{X})$, as $\hat{\sigma}_k^2/n_k$ for the smallest value of k such that there is no evidence that $\gamma_j(1) \neq 0$ for $j > k$. This is sensible because then, according to corollary 1, there is no reason to believe that the error e_k is reduced by further iterations of the method.

An upper bound of the complexity of the method is $12n + O(\log_2 n)$. Consider the sample code in figure 3.5. The only contributions at order n are from the while loop. At iteration number i , the while loop can be computed using exactly $6n_i + 4$ floating point operations. Using geometric series, the total floating point

operations are

$$\begin{aligned} \text{cost} &= \sum_{j=0}^{d-1} (4 + 6n_j) = 4d + 6 \cdot 2^d \sum_{j=0}^{d-1} 2^{-j} \\ &= 4d + 12(n-1) \leq 12n + O(\log_2 n) \quad \text{as } n \rightarrow \infty \end{aligned} \quad (3.24)$$

For time consuming computations which requires 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 section 3.7, but first consider these test results.

3.6 Test results

Two tests of the algorithm are presented. First, 6080 causal random AR(1) and AR(2) processes were generated. According to the methods, that means the exact value of $V(\bar{X})$ 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 $\gamma(\tau) \leq \gamma(0)e^{-1}$). Gamma regression is suitable because the observations of ϵ^2 are independent, identically gamma-distributed, and the model is $\log(\epsilon^2) = \beta_0 + \beta_1 \log(n/\tau)$. The expected relative error squared is

$$\epsilon^2 = e^{\beta_0} \left(\frac{n}{\tau} \right)^{\beta_1}. \quad (3.25)$$

Maximum likelihood-estimates of β_j and standard errors are given for the causal AR(1) and AR(2) processes in table 3.1. The table shows that if there is 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. Amongst the processes, the type of autocovariance was the main differentiator of the AR(p)-processes. For $p = 1$, the autocorrelation is of the form $\gamma(h)/\sigma^2 = \phi^{-h}$, whilst 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 ($p \geq 0.34$, t-test), and therefore the difference in ϵ^2 between the AR(1) and AR(2) experiments are attributed to the $\gamma(h)$ according to the methods. A plot of the regression analysis is given in figure 3.3 and the regression summaries, see table 3.1.

Second, two standard textbook physics applications were studied. The variance of the mean energy was estimated using the Flyvbjerg & Petersen (1989) blocking method and the automated blocking method. The estimates were compared with dependent bootstrapping using a geometric simulation type (Politis and Romano 1994). In either application, the autocorrelation functions bore resemblance of

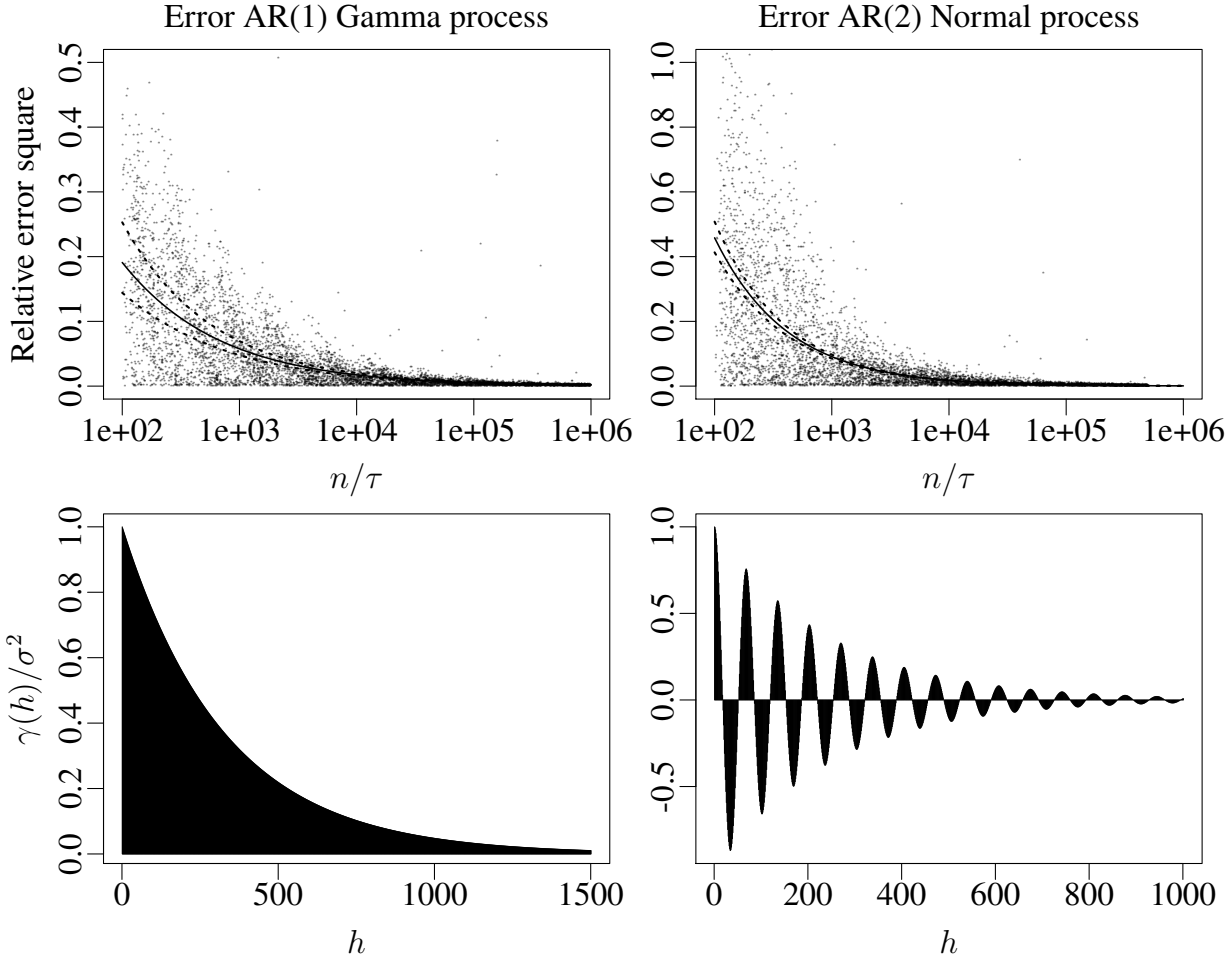


Figure 3.3: Relative error squared of two autoregressive models versus observations per time autocorrelation time-constant. It is clear that there is exponential convergence rate for two common correlation structures in natural sciences. In the left panel: AR(1) autocorrelation is positive with exponential decay, typical of Metropolis-type Markov chains, where it is expected that the observations correlate positively. is $X_k = \phi X_{k-1} + e_k$ with e_k iid gamma distributed. You can view the gamma distribution is a generalization of the chi squared distribution. In the right panel: causal AR(2)-processes have exponential decay, but may be oscillatory as here. The process is given by $X_k = \phi_1 X_{k-1} + \phi_2 X_{k-2} + e_k$ with e_k iid normal distributed. It was found that the method was insensitive to the distribution of the observations ($p \geq 0.34$). 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 modelled 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 $n \gtrsim 25000\tau$.

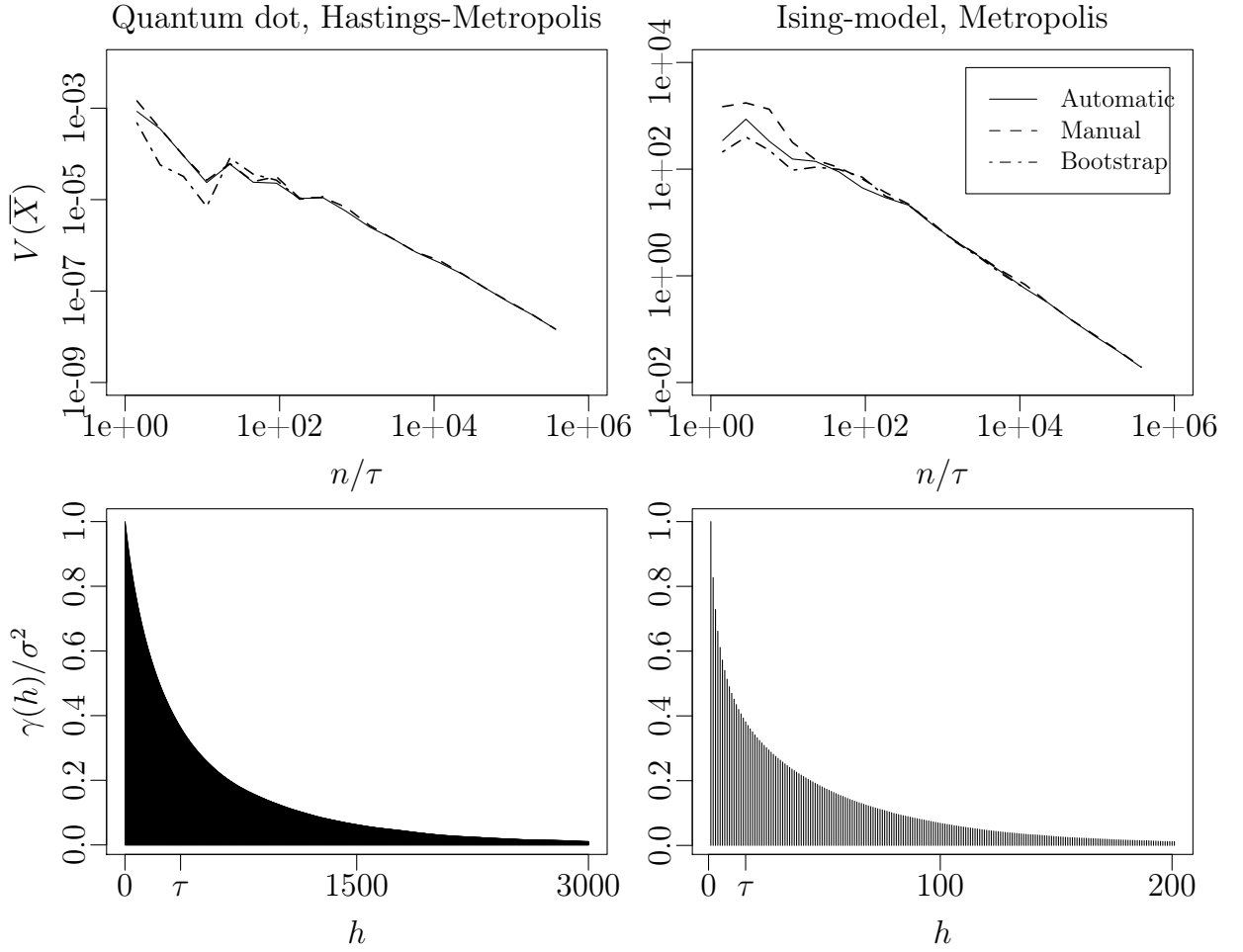


Figure 3.4: Case study of two textbook physics applications for mean variance estimation. The variance of the mean energy was estimated using manual- and automatic Blocking methods and compared with dependent bootstrapping using geometric simulation. The estimates are plotted on top, whilst the autocorrelation depicted on the bottom. Left: 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 $V(\bar{X}) = 1.46 \cdot 10^{-8}$ at 10^8 samples according to all three models, and by the above discussion the error of the variance estimate is expected to correct to near 2nd digit. Right: Ising model implementation of 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 5% 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 $V(\bar{X}) = 1.49 \cdot 10^{-1}$ according to all three methods, and there is reason to believe that the first two digits after the decimal point are correct.

Table 3.1: 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	$< 10^{-16}$

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-likelihood (Gardiner 1985). 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 $T = 2.4$ (Plischke and Bergersen 2006). The energy was sampled from a Boltzman distribution at significance level $\alpha = 0.05$ according to a χ^2 goodness-of-fit test (Devore and Berk 2012) using a stationary, time-reversible Markov chain constructed using the Metropolis algorithm (Metropolis et al. 1953). The autocorrelation time constant was measured to be $\tau = 16$. A plot of the results are contained in figure 3.4.

3.7 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. This 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

Python implementation of algorithm

```

# 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
n = 2**(d-k)
answer = s[k]/n

```

Figure 3.5: 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 from github.com/computative/block

memory which 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 are preserved. Now on the chunk, apply blocking transformations $D - d$ times and form $\mathbf{X}_{D-d} = T_{D-d}T_{D-d-1} \cdots T_1 \mathbf{X}$ where T_i is defined in equation (3.13). 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 amount of transformed data is $2^k \cdot 2^{d-k} = 2^d$, as required. On the parent node, or personal computer doing the final estimate, write the data to memory by concatenating 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 figure 3.2.

The computational cost of this is low. Performing the transformations T_i as it is done in the code of figure 3.5 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:

$$\sum_{j=1}^{D-d} n_{j-1} = \sum_{j=0}^{D-d-1} 2^{D-k-j} = \frac{2^D - 2^d}{2^{k-1}}. \quad (3.26)$$

According to equation (3.24), it is necessary to add the $12(2^d - 1) + 4d$ floating point operations which the parent node must compute at the end, so the total cost is bounded above by

$$\text{cost} \stackrel{(3.26)}{\leq} \underbrace{\frac{2^D}{2^{k-1}}}_{=n} + 2^d \underbrace{\left(12 - \frac{1}{2^{k-1}}\right)}_{\leq 1} \leq n + O(1) \quad \text{as } n \rightarrow \infty.$$

The reason it is possible to rejoin the time series by putting it end-to-end is the same reason why 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 (Politis and Romano 1994; Politis and White 2006).

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

$$\bar{X} = \frac{1}{2^D} \sum_{i=1}^{2^D} X_i = \sum_{j=1}^{2^k} \frac{1}{2^D} \underbrace{\sum_{i=(j-1)2^{D-k}}^{j2^{D-k}} X_i}_{2^{D-k} \hat{\mu}_j} = \frac{1}{2^k} \sum_{j=1}^k \hat{\mu}_j. \quad (3.27)$$

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 does not have to be read back into memory for post-processing later.

Assume the time series is split in 2^k chunks of length 2^{D-k} :

$$X_1, X_2, \dots, X_{2^{D-k}}, \underbrace{X_{2^{D-k}+1}, \dots, X_{2^{D-k}+1}, \dots, X_{2^D}}_{\text{chunk number } i \equiv \mathbf{X}}$$

↙ (Step 1) Transform chunk i by:
 $\mathbf{X}_{D-d} = T_{D-d} \cdots T_2 T_1 \mathbf{X}$

$\mathbf{Y}_i \equiv \mathbf{X}_{D-d}$ (Step 2) Repeat for all i .

↘ (Step 3) Reassemble transformed chunks
 $(\mathbf{Y}_1, \mathbf{Y}_2, \dots, \underbrace{\mathbf{Y}_i}_{\text{Transformed chunk } i}, \dots, \mathbf{Y}_{2^k})^\top$

(Step 4) Apply algorithm, fig. 3.2, to this vector ←

Figure 3.6: 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 $T_{D-d}T_{D-d-1} \cdots T_2T_1$. Define \mathbf{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 $(\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_{2^k})^\top$. 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 figure 3.2. See section 3.7 for more details including the definitions of the numbers D, d and k .

Chapter 4

Discussion

The present study provides the following four new contributions: (a) Rigorous proof of the Flyvbjerg & Petersen (1989) blocking method conforming to the standards of modern mathematics. The results give prospects for new research with relevance to any blocking method. (b) A new 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 new contributions include proof of the behavior of the method. Furthermore, proposition 5 outlines a new approach to (i) estimate the standard error more efficiently, and (ii) provides economical error estimates, both in terms of computational 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 & Petersen (1989) 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 (Flyvbjerg and Petersen 1989) providing motivational mathematics to explain the idea of blocking transformations for standard error estimation. They claim that there exists 'an obvious fixed point', but give no proof thereof (Flyvbjerg and Petersen 1989). For mathematically interested readers, this can present a distraction since fixed points of any function $T_j : A \rightarrow B$ are defined when $A = B$ (Hazewinkel 1989; Borowski and Borwein 1989; McDonald and Weiss 2012). In this context $A \neq B$ since $A = \mathbb{R}^{n_i}$ and $B = \mathbb{R}^{n_i/2}$ (see section 1). Thus, from a mathematical point of view, 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 truncation error e_k from equation 3.1 as a function of $\gamma_k(h)$. Here, γ_k is the autocovariance of the time series \mathbf{X} after

it is subject to k blocking transformations. From this and the transformation properties of $\gamma_k(h)$, it follows that the behavior of the truncation error is given by the quantity $\{\gamma_k(1)\}$, which is the autocovariance of sequential observations, as proposition 5 explains. 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 (1989) appear unaware of this, because they state that the blocking method converges if the autocovariance $\gamma(h) \propto 1/h$ (Flyvbjerg and Petersen 1989), which does not capture the essence, as was proven above. In fact, their blocking method works whenever X_1, X_2, \dots is asymptotic uncorrelated, as theorem 17 makes precise.

Proposition 5 proves that the blocking method is applicable under more general conditions than assumed by Flyvbjerg & Petersen (1989). First, because $\gamma(h)$ needs not be proportional to $1/h$ (note that proposition 5 places no restriction on $\gamma(h)$, although finite variance is required). Therefore, $\gamma(h)$ can have any shape. Second, Flyvbjerg & Petersen (1989) constrain $\gamma(h)$, exactly n degrees of freedom (since γ is a function $\gamma : \{0, 1, \dots, n_1\} \rightarrow \mathbb{R}$), whilst the new results only constrain $\{\gamma_k(1)\}$, exactly $\log_2(n) = d$ degrees of freedom. Theorem 17 may also have theoretical interest in statistical mathematics. The sum $s = \sum_{h=1}^{n-1} (1 - h/n) \gamma(h)$, appears frequently in the study of time series (Shumway and Stoffer 2017), but together proposition 5 and lemma 3 imply that $(2/n)s$ is determined up to a constant¹ by the set $\{\gamma_k(1)\}$. Moreover, according to theorem 17, $\gamma_k(h) \rightarrow 0$ uniformly on \mathbb{N} under blocking transformations. In this way, blocking transformations are intimately linked to s . 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 $e_k = (2/n)s$). 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 (Stein 1956) and applications such as those by Schäfer and Strimmer (Schäfer and Strimmer 2005). Other proposed applications of proposition 5 is the proof of corollary 1, which explains the behavior of the 1989 blocking method, and why the automatic blocking method works. The corollary shows that the estimates, $V(\bar{X})$, 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 $k \geq j$. In case (i) the truncation error $e_k = 0$ since if the components of \mathbf{X}_k are uncorrelated, then

¹That constant is σ_0^2 , as can be proved by iteratively using proposition 5

$\text{Cov}((\mathbf{X}_k)_i, (\mathbf{X}_k)_j) = \sigma_k^2 \delta_{ij}$ by definition, so $\gamma_k(h) = 0$ for $h \geq 1$ 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 6 strengthens this statement to include case (ii) because it shows that γ_k converges uniformly and identically to zero on \mathbb{N} whenever $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$.

The algorithm for the automation is new, although the concept of automatic computation of $V(\bar{X})$ is not new. In physics, the most well known automated method for standard error estimation is perhaps dependent bootstrapping (Politis and Romano 1994). Dependent bootstrapping is useful when n is small or if n is large and the required precision is small. According to Politis and Romano (1994), the method has asymptotically valid procedures even for multivariate parameter spaces. However, high precision estimates for large data sets are often needed. Flyvbjerg and Petersen (1989) 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 $\hat{\sigma}^2$. Typically, this method works well since $\gamma(h) \propto 1/h$, which is commonly used in physics. However, this is not always 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 new automation works for causal AR(p) processes for any p , and places no assumption on the shape of $\gamma(h)$. In addition, the present thesis provides updates that makes it convenient to use the method for multithread computing. Another alternative method is the Gamma method proposed by Wolff (2004). The Gamma method works well with correct set up, with errors claimed to be lower than those of the automatic blocking method proposed here. Wolff (2004) 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 (2004) 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 the recommendation of Wolff (2004), R. Lee et al. (2011) are proponents of a method, which Wolff (2004) has called binning (which is essentially a blocking method). R. Lee et al. (2011) proposed inequalities that can be used to automate calculations. However, this approach requires estimates that may or may not be available. The new automated blocking method has none of the complications mentioned above.

The automation uses one approximation in computing M_j . It was assumed that

Σ_j is diagonal even though it is only diagonal to leading order in $1/n$. This approximation can be avoided by using lemma 8. The benefit of the approximation is that the inversion of Σ_j can be simplified (since if $\Sigma_j = \text{diag}(r_j, \dots, r_{d-j})$, then $\Sigma_j^{-1} = \text{diag}(1/r_j, \dots, 1/r_{d-j})$ is inverted using only $d - j$ floating point operations (Lay 2012)).

The method validation of $\text{AR}(p)$ processes is new (and 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. $\text{AR}(1)$ and $\text{AR}(2)$ processes were chosen because the Wold decomposition says that the random component of any time series can be expressed as an autoregressive model (Shumway and Stoffer 2017). $\text{AR}(1)$ and $\text{AR}(2)$ correlation functions are the two most common ones encountered in modeling of time series. The two text book cases, quantum dots and the Ising model, show that their correlation structures were similar to the $\text{AR}(1)$ -processes. Using equation 3.25, 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 approximately $n = 10^4$ and 10^5 observations for the Ising model and quantum dot, respectively. Whilst it is expected that the second digit is also correct if n approximately 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 (Wolff 2004). However, unlike the estimates due to Wolff (2004), table I 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 applications show that the estimates are similar to those of dependent bootstrapping and the Flyvbjerg & Petersen (1989) blocking method. This is regardless of n (see figure 3.4), is fully automatic and works in $O(n)$ -time.

Chapter 5

Conclusion and perspectives

As stated in the introduction, we have been aiming for a reliable method to estimate standard errors of \bar{X} for large time series. This is important because such time series appear everywhere in Markov chain Monte Carlo methods in computational sciences. Its importance was reinforced since rivaling methods available to physicists are dependent bootstrapping, which stops working when the number of observations n is sufficiently large. Yet, physicists must estimate $V(\bar{X})$, otherwise there is no reason to believe that \bar{X} estimates the expected value of our observations. The popular resampling methods available to physicists has sound mathematical foundation, and now the same is true for the blocking method.

A rigorous proof of the blocking method (Flyvbjerg and Petersen 1989) is a main result of the present study. That method has become one of the most used and cited methods for estimating standard errors $V(\bar{X})^{1/2}$. Second, the proof gives an automated implementation that eliminates the need for human intervention, moreover guarantees that the estimates are rigorously sound, since applying blocking transformations manually requires some skill to ensure that the estimates are not divergent. The method uses Fisherian inference to propose a hypothesis test that can be used to determine the estimate of the standard error. The new method has complexity $O(n)$, and works for all common covariance structures in natural sciences. This should first and foremost appeal to researchers in statistical data analysis, 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. All source code included in the thesis is available for download from github.com/computative/block.

In this thesis, several paths for future research are proposed. Proposition 5 from section 3.2 on page 76 shows that the behavior of any blocking method is determined by the set $\{\gamma_k(1)\}_{k=0}^{d-1}$. Recall that $\gamma_k(1)$ is the autocovariances of

sequential observations in \mathbf{X}_k and k is the number of blocking transformations which have been applied to the time series X_1, X_2, \dots . However, more advanced estimation is needed to use the result for efficient estimation of $V(\bar{X})$. The problem is that for large k , the data available to estimate $\gamma_k(1)$ is 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, mathematically rigorous 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 18 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 science (see for example Jones (2004)), such that it is made explicit for which methods theorem 18 continues to hold.

Appendix A

Lemma. (†) If a particle is measured at $\mu = a/2$ such that $\Psi(0, x) = \delta(x - \mu)$ in an infinite square well at $t = 0$, then $\langle X \rangle = \mu$ for all $t \geq 0$. **PROOF.** For an infinite square well, let's write the wave function for the given initial conditions in the following way:

$$\Psi(x, t) = \frac{1}{\mu} \sum_{n=1}^{\infty} \sin\left(n \frac{\pi}{2}\right) \sin\left(n \frac{\pi x}{2\mu}\right) e^{-iE_n t}; \quad \mu = \frac{a}{2}, \quad \text{for all } x \in [0, a].$$

Fix $t \in [0, \infty)$ and define the function $f_t : [-\mu, \mu] \rightarrow \mathbb{R}$ given by

$$f_t(x) = x |\Psi(\mu + x, t)|^2.$$

We first show that f_t is antisymmetric on $[-\mu, \mu]$. Note that since: $\sin(\theta + \varphi) = \cos \theta \sin \varphi + \sin \theta \cos \varphi$ we have $\sin(\theta - n\pi) = (-1)^n \sin \theta$ (perhaps you see this directly), so we have

$$\sin\left(n \frac{\pi}{2} \frac{1}{\mu} (\mu + x)\right) = -\sin\left(n \frac{\pi}{2} \frac{1}{\mu} (\mu - x - \mu - \mu)\right) = (-1)^{n+1} \sin\left(n \frac{\pi}{2} \frac{1}{\mu} (\mu - x)\right). \quad (\text{A.1})$$

Pick $n \in \mathbb{N}$. If n is odd, then $n + 1$ is even, so there exists $k \in \mathbb{Z}$ such that $n + 1 = 2k$, so we can write

$$\begin{aligned} \sin n \frac{\pi}{2} \sin\left(n \frac{\pi}{2} \frac{1}{\mu} (\mu + x)\right) &\stackrel{(\text{A.1})}{=} \sin\left(n \frac{\pi}{2}\right) \underbrace{(-1)^{2k}}_{=1^{k=1}} \sin\left(n \frac{\pi}{2} \frac{1}{\mu} (\mu - x)\right) \\ &= \sin n \frac{\pi}{2} \sin\left(n \frac{\pi}{2} \frac{1}{\mu} (\mu - x)\right). \end{aligned} \quad (\text{A.2})$$

On the other hand, if n is even, then $\sin(n\pi/2) = 0$ so the above equality is automatic. That means (A.2) is true for all $n \in \mathbb{N}$ (*) so

$$\begin{aligned} f_t(x) &= x|\Psi(\mu + x, t)|^2 = x \left| \frac{1}{\mu} \sum_{n=1}^{\infty} \sin\left(n\frac{\pi}{2}\right) \sin\left(n\frac{\pi}{2}\frac{1}{\mu}(\mu + x)\right) e^{-iE_n t} \right|^2 \\ &\stackrel{(*)}{=} x \left| \frac{1}{\mu} \sum_{n=1}^{\infty} \sin\left(n\frac{\pi}{2}\right) \sin\left(n\frac{\pi}{2}\frac{1}{\mu}(\mu - x)\right) e^{-iE_n t} \right|^2 \\ &= -(-x)|\Psi(\mu - x, t)|^2 = -f_t(-x). \end{aligned}$$

This proves that f_t is antisymmetric, so its integral on $[-\mu, \mu]$ is zero and therefore:

$$\begin{aligned} \langle X \rangle(t) &= \int_0^a \Psi(x, t)^* x \Psi(x, t) dx = \int_0^a (\mu - \mu + x) |\Psi(x, t)|^2 dx \\ &= \underbrace{\mu \int_0^a |\Psi(x, t)|^2 dx}_{=1} + \int_0^a (x - \mu) |\Psi(x, t)|^2 dx \\ &= \mu + \underbrace{\int_{-\mu}^{\mu} u |\Psi(u + \mu, t)|^2 du}_{= \int f_t(u) du = 0} = \mu, \end{aligned}$$

Where we made the substitution $u = x - \mu$ in the penultimate step. ■

LEMMA 8. *The sequence $\{f_k\}$ satisfy the following properties:*

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

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

$$\begin{aligned} \sum_{i=1}^{2^{k+1}-1} f_k(i) &= 1 + 2 + \cdots + 2^k + (2^k - 1) + \cdots + 2 + 1 \\ &= 2^k + 2 \sum_{i=1}^{2^k-1} i = 2^k + 2 \frac{2^k - 1}{2} 2^k = 2^{2k}. \end{aligned}$$

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

$$\begin{aligned} 1 &= f_1(1) = f_0(1) + 2f_0(1 - 1) + f_0(1 - 2) = 1 + 0 + 0 \\ 2 &= f_1(2) = f_0(2) + 2f_1(2 - 1) + f_0(2 - 2) = 0 + 2 + 0 \\ 1 &= f_1(3) = f_0(3) + 2f_2(3 - 1) + f_0(3 - 2) = 0 + 0 + 1. \end{aligned}$$

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

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

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. ■

Table A.1: Chi-square 90-,95- and 99-percentiles: Percentiles in the chi squared distribution for $1 \leq d - k \leq 48$, which suffices for any error estimation with $\leq 10^{14}$ observations, at the three significance levels that performed best, $1 - \alpha = 0.99, 0.95$ and 0.90 .

$d - k$	$q_{d-k}(0.99)$	$q_{d-k}(0.95)$	$q_{d-k}(0.9)$
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

Cited literature

- Agresti, Alan (2015). *Foundations of Linear and Generalized Linear Models*. 1st ed. New Jersey: Wiley & Sons, Inc.
- Apostol, Tom M. (1961). *Calculus, Volume 1: Introduction, with Vectors and Analytic Geometry*. 1st ed. New York: Blaisdell publishing company.
- Bickel, Peter J. and David A. Freedman (1981). “Some Asymptotic Theory for the Bootstrap”. In: *Ann. Statist.* 9.6, pp. 1196–1217.
- Boas, Mary L. (2005). *Mathematical Methods in the Physical Sciences*. 3rd ed. New Jersey: Wiley. ISBN: 0-471-19826-9.
- Borowski, Ephraim J. and Joathan M. Borwein (1989). *Fixed Point*. London.
- Bradley, Richard C. (1987). “The Central Limit Question under ρ -Mixing”. In: *Rocky Mountain Journal of Mathematics* 17.1, pp. 95–114.
- Brockwell, Peter J. and Richard A. Davis (2016). *Introduction to Time Series and Forecasting (Springer Texts in Statistics)*. 3rd ed. Switzerland: Springer.
- DeGroot, Morris H. and Mark J. Schervish (2012). *Probability and Statistics*. 4th ed. Pearson Education, Inc.
- Denby, B. (2004). “Trends in Physics Data Analysis Algorithms”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 534.1, pp. 343–347.
- Devore, Jay L. and Kenneth L. Berk (2012). *Modern Mathematical Statistics with Applications*. 2nd ed. London: Springer.
- Dyson, Sir Frank Watson, Sir Arthur Eddington, and Charles Davidson (1920). “IX. A Determination of the Deflection of Light by the Sun’s Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919”. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 220.571-581, pp. 291–333.
- Efron, Bradley (1979). “Bootstrap Methods: Another Look at the Jackknife”. In: *The Annals of Statistics* 7.1, pp. 1–26.
- (1986). “Why Isn’t Everyone a Bayesian?” In: *The American Statistician* 40, pp. 1–5.
- (1987). *The Jackknife, the Bootstrap, and Other Resampling Plans (CBMS-NSF Regional Conference Series in Applied Mathematics)*. Philadelphia: Society for Industrial and Applied Mathematics.

- Efron, Bradley and Carl Morris (1977). “Stein’s Paradox in Statistics”. In: *Scientific american* 236.5, pp. 119–127.
- Flyvbjerg, H. and H.G Petersen (1989). “Error Estimates on Averages of Correlated Data”. In: *The Journal of Chemical Physics* 91, pp. 461–466.
- Fraleigh, John B. (2002). *A First Course in Abstract Algebra, 7th Edition*. 7th ed. New York: Pearson.
- Gardiner, Crispin W. (1985). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. 2nd ed. Berlin: Springer-Verlag.
- Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2014). *Bayesian Data Analysis*. 3rd ed. Boca Raton: CRC Press.
- Grimmett, Geoffrey and Dominic Welsh (2014). *Probability: An Introduction*. 2nd ed. Oxford: Oxford University Press.
- Hammond, B.L., William A. Lester Jr., and Peter J. Reynolds (1994). *Monte Carlo Methods In Ab Initio Quantum Chemistry (World Scientific Lecture and Course Notes in Chemistry ; Vol. 1)*. Singapore: Wspc.
- Hazewinkel, Michiel (1989). “Fixed Point”. In: *Encyclopedia of mathematics* 4.
- (1993). “Tensor Product”. In: *Encyclopedia of mathematics* 9.
- Høgberget, Jørgen (2013). *Quantum Monte-Carlo Studies of Generalized Many-Body Systems*. Master Thesis. Oslo: University of Oslo.
- Huber, Peter J. and Elvezio M. Ronchetti (2009). *Robust Statistics*. 2nd ed. New Jersey: Wiley.
- Ibragimov, Il’dar Abdullovich (1975). “A Note on the Central Limit Theorem for Dependent Random Variables”. In: *Teor. Veroyatnost. i Primenen.* 20.1, pp. 135–141.
- Ising, Ernst (1925). “Beitrag Zur Theorie Des Ferromagnetismus”. In: *Zeitschrift für Physik* 31.1, pp. 253–258.
- James, W. and Charles Stein (1961). “Estimation with Quadratic Loss”. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 361–379.
- Jones, Galin L. (2004). “On the Markov Chain Central Limit Theorem”. In: *Probability Surveys* 1, pp. 299–320.
- Kampen, N.G. Van (2007). *Stochastic Processes in Physics and Chemistry, Third Edition (North-Holland Personal Library)*. 3rd ed. Amsterdam: North Holland.
- Lay, David C. (2012). *Linear Algebra and Its Applications*. 4th ed. Boston: Addison-Wesley.
- Lee, John M. (2011). *Introduction to Topological Manifolds*. 2nd ed. New York: Springer.
- (2013). *Introduction to Smooth Manifolds*. 2nd ed. New York: Springer.

- Lee, R.M., G.J. Conduit, N. Nemec, P. López Ríos, and N.D. Drummond (2011). “Strategies for Improving the Efficiency of Quantum Monte Carlo Calculations”. In: *Physical review E* 83.6, p. 066706.
- Lindstrøm, Tom (2006). *Kalkulus*. 3rd ed. Oslo: Universitetsforlaget.
- (2017). *Spaces: An Introduction to Real Analysis*. Providence, Rhode Island: American Mathematical Society.
- Mathai, A.M. and Serge B. Provost (1992). *Quadratic Forms in Random Variables*. New York: Marcel Dekker.
- McDonald, John N. and Neil A. Weiss (2012). *A Course in Real Analysis*. 2nd ed. Oxford: Academic press.
- Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.
- Munkres, James (2000). *Topology*. 2nd ed. Upper Saddle River, New Jersey: Prentice Hall, Inc.
- Newman, M. E. J. and G. T. Barkema (1999). *Monte Carlo Methods in Statistical Physics*. Oxford: Clarendon Press.
- Øksendal, Bernt (2014). *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. 6th ed. Berlin: Springer. ISBN: 3-540-04758-1.
- Parr, William C. (1985). “The Bootstrap: Some Large Sample Theory and Connections with Robustness”. In: *Statistics & Probability Letters* 3.2, pp. 97–100.
- Peskin, Michael and Daniel V Schroeder (1995). *An Introduction to Quantum Field Theory*. New York: Westview Press.
- Plischke, Michael and Birger Bergersen (2006). *Equilibrium Statistical Physics*. 3rd ed. New Jersey: World scientific.
- Politis, Dimitris N. and Joseph P. Romano (1994). “The Stationary Bootstrap”. In: *Journal of the American Statistical Association* 89.428, pp. 1303–1313.
- Politis, Dimitris N. and Halbert White (2006). “Automatic Block-Length Selection for the Dependent Bootstrap”. In: *Econometric Reviews* 23.1, pp. 53–70.
- Riley, K. F. and M. P. Hobson (2013). *Essential Mathematical Methods for the Physical Sciences*. Cambridge: Cambridge university press.
- Ross, Sheldon M. (2014). *Introduction to Probability Models, Eleventh Edition*. 11th ed. Oxford: Academic Press. ISBN: 0-12-407948-2.
- Rynne, Bryan and M.A. Youngson (2007). *Linear Functional Analysis (Springer Undergraduate Mathematics Series)*. 2nd ed. London: Springer.
- Schäfer, Juliane and Korbinian Strimmer (2005). “A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics”. In: *Statistical Applications in Genetics and Molecular Biology* 4.1, pp. 1–30.

- Schroeder, Daniel V (1999). *Thermal Physics*. 1st ed. San Francisco: Addison-Wesley.
- Shumway, Robert H. and David S. Stoffer (2017). *Time Series Analysis and Its Applications with R Examples*. 4th ed. New York: Springer.
- Stein, Charles (1956). “Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Contributions to the Theory of Statistics* 1, pp. 197–206.
- Sweeting, T. J. (1980). “Uniform Asymptotic Normality of the Maximum Likelihood Estimator”. In: *The Annals of Statistics* 8.6, pp. 1375–1381.
- Szabo, Attila and Neil S. Ostlund (1996). *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mineola: Dover Publications.
- Tukey, John W. (1958). “Abstracts of Papers”. In: *The Annals of Mathematical Statistics* 29.2, pp. 614–623.
- Upton, Graham and Ian Cook (2014). *A Dictionary of Statistics 3e (Oxford Quick Reference)*. 3rd ed. Oxford: Oxford University Press.
- van der Vaart, A. W. (1998). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. 1st ed. Cambridge: Cambridge University Press.
- Viviani, Vincenzo (1717). *Racconto Istorico Della Vita Di Galileo Galilei*. Firenze.
- Wolff, Ulli (2004). “Monte Carlo Errors with Less Errors”. In: *Computer Physics Communications* 156.

Index of notation

- $\mathbf{1}$ vector of ones, 82
 α significance value, 13
 $A \subseteq B$: A is subset or equal B , 9
 $A \subset B$: A is subset of B , 9
 $\text{AR}(p)$ autoregressive model of order p , 45
 $A \cup B$ union of sets, 10
 B Bayes solution, 42
 $B(x; r)$ open ball at x with radius r , 7
 β_j regression coefficient, 90
 $\text{Bias}(\hat{A}; A)$, 14
 BR Bayes risk, 42
 $\text{Cov}(A, B)$ covariance of A, B , 17
 $\text{Cov } \mathbf{A}$ covariance matrix of \mathbf{A} , 18
 $C(X, \mathbb{R})$ set of continuous $X \rightarrow \mathbb{R}$ functions, 11
 $\text{diag}(x_1, \dots, x_n)$ diagonal matrix, 23
 δ_{ij} Kronecker delta, 78
 $d(x, y)$ metric, 7
 $\langle E \rangle$ expected value of E , 14
 $\langle \mathbf{E} \rangle$ vector of elements $\langle E_i \rangle$, 18
 \equiv definition, 7
 $\xrightarrow{\text{a.s.}}$ almost sure convergence, 58
 \xrightarrow{d} convergence in distribution, 58
 \xrightarrow{P} convergence in probability, 58
 $E \sim \text{N}$: E has the distribution N , 22
 $\{\mathbf{e}_k\}_{k=1}^n$ standard basis for \mathbb{R}^n , 85
 $E \sim \text{N}$ E has approximate distribution N , 22
 ϵ^2 expected relative error squared, 91
 E_T trail energy, 70
 $\{f_k\}_{k=1}^\infty$ a sequence, 9
 $f \circ g$ composition of f and g , 9
 $\gamma(h)$ autocovariance, 1
 $\hat{\gamma}(h)$ estimator of $\gamma(h)$, 45
 $\hat{\gamma}_k(1)$ estimator of $\gamma_k(1)$, 84
 Γ_k generator of $\hat{\gamma}_k(1)$, 83
 $\gamma_k(h)$ autocovariance of \mathbf{X}_k , 64
 $\gamma_k(1)$: $\gamma_k(h)$ evaluated at $h = 1$, 77
 H_a alternative hypothesis, 33
 H_0 null hypothesis, 33
 $H(X)$ largest open set of cdfs on X , 16
iid independent, identically distributed, 16
 I_n identity matrix of rank n , 24
 $\inf A$ largest lower bound of A , 13
 $\lim_{H \rightarrow B} T(H)$ limit in Banach space, 12
 $|\cdot|$ absolute value, 7
 $\|\cdot\|_X$ norm of X , 11
 $\|\cdot\|_\infty$ sup-norm, 12
 \log natural logarithm, 9
 μ an expected value, 1
 MBR minimum Bayes risk, 42
 $\min\{x_1, x_2, \dots, x_n\}$ smallest of all x_i , 8
 M_j blocking test statistic, 81
 MSE mean squared error, 43
 \mathbb{N} natural numbers $1, 2, \dots$, 7
 $\text{N}(\mu, \sigma^2)$ normal distribution, 22
 n usually size of \mathbf{X} , 64

- n_k size of \mathbf{X}_k , 64
- not: negation, 51
- \otimes Kronecker product, 85
- \emptyset the empty set, 9
- $\boldsymbol{\theta}$ vector of parameters, 38
- $\hat{\boldsymbol{\theta}}$ estimator of $\boldsymbol{\theta}$, 38
- $O(n)$ asymptotic bound, 2
- P measure of probability, 1
- $P(E|K)$ probability of E given K , 22
- π_i stationary distribution, 47
- P_{ij} transition probability matrix, 47
- \mathbb{Q} rational numbers, 7
- $q_j(p)$ chi squared percentile, 88
- R risk function, 42
- \mathbb{R} real numbers, 7
- ρ metric of $C(X, \mathbb{R})$, 11
- σ^2 a variance, 1
- S^2 unbiased estimator of σ^2 , 15
- S_i shift operator, 83
- Σ usually a covariance matrix, 18
- $\hat{\sigma}^2$ estimator of σ^2 , 14
- Σ_j blocking covariance matrix, 82
- σ_k^2 variance of elements in \mathbf{X}_k , 64
- $\Sigma_{\mathbf{Y}}$ covariance matrix of \mathbf{Y} , 83
- $\text{Supp } f$ support of f , 9
- τ time constant for γ , 90
- T_i blocking matrix, 83
- $\text{Tr}(A)$ trace of matrix A , 85
- $\text{unif}(a, b)$ continuous uniform distribution, 68
- $V(X)$ variance of X , 1
- $V(\mathbf{E})$ covariance matrix of \mathbf{E} , 19
- X_i observation number i , 1
- \bar{X} sample mean, 1
- χ_ν^2 chi squared distribution with ν dof, 23
- $(\mathbf{X}_k)_i$ element number i of \mathbf{X}_k , 18
- \bar{X}_k mean of elements of \mathbf{X}_k , 64
- \mathbf{X}_k time series subject to k blocking transformations, 64
- ψ_T trail state function, 70
- end of proof, 8
- end of example, 8

Index

- central limit theorem, 26
- almost surely, 58
- alternative hypothesis, 33
- aperiodic, *see* Markov chain, aperiodic
- autocovariance, 45
- autoregressive model, 45
- Banach space, 11
- Bayes risk, 42
- Bayes solution, 42
- Bayesian statistics, 38
- Bias($\hat{A}; A$), 14
- bijection, 8
- blocking transformation, 64
- Boltzmann distribution, 67
- cdf, *see* cumulative distribution
- central limit theorem
 - for dependent variables, *see* Ibragimov thm.
- Chebyshev's inequality, 1
- chi-square distribution, 23
- class, 47
- CLT, *see* central limit theorem
- communicating states, 47
- complete, *see* metric space, complete
- conditional probability, 22
 - distribution, 20
- confidence
 - interval, 32
- convergence
 - almost surely, 58
 - in distribution, 58
 - in probability, 58
 - pointwise, 9
 - sequence, 10
 - uniform, 10
- countable, 9
- covariance, 17
- cumulative distribution, 12
- decision function, 42
- dependence, 19
- detailed balance, 49
- ergodic, *see* Markov chain, ergodic
- estimate, 14
 - estimator, 14
 - M , 59
 - minimal variance unbiased estimator, 31
 - point estimation, 30
- Fisherian, 28
- Fréchet derivative, 12
- function
 - codomain, 8
 - continuous, 8
 - Lipschitz, 8
 - domain, 8
- H_a , *see* alternative hypothesis
- Hastings-Metropolis
 - algorithm, 50
 - Metropolis algorithm, 51
 - theorem, 50
- H_0 , *see* null hypothesis
- hypothesis testing, 32

- Ibragimov theorem, 46
- identically distributed, 14
- identically equal, 8
- importance sampling, 71
- independence, 15, 19
- inference, *see* statistical inference
- injective, 8
- irreducible, 47

- James-Stein estimator, 43
- joint probability, 17

- likelihood, 30, 39
- limit, 10
- linear process, *see* autoregressive model
- Lipschitz, *see* function, continuous, Lipschitz
- loss function, 42

- marginal distribution, 20
- Markov chain, 47
 - aperiodic, 48
 - ergodic, 48
 - periodic, 48
 - time reversed, 48
 - time reversible, 49
- maximum likelihood, 30
 - estimator, 30
- mean squared error, 43
- metric space, 7
 - metric, 7
 - complete, 10
- minimum Bayes risk, 42
- MSE, *see* mean squared error
- MVUE, *see* estimate, minimal variance unbiased estimator

- normal distribution, 21
 - multivariate normal, 21
 - standard normal, 21
- normed space, 11
- null hypothesis, 33
- open ball, 7

- p -value, 37
- percentile, 13
 - function, *see* quantile, function
- periodic, *see* Markov chain, periodic
- posterior distribution, 39
- prior distribution, 39
- product rule (of probability), 22

- quantile, *see* percentile
 - function, 13

- random vector, 18
- recurrent, 47
 - null, 47
 - positive, 47
- ρ -mixing, *see* uncorrelated, asymptotic
- risk function, 42

- sample covariance, 45
- sample mean, 14
- sample variance, 14
- score function, 30
- sequence
 - Cauchy, 10
 - convergent, 10
 - of functions, 9
- set
 - compact, 10
 - connected, 9
 - open, 7
- shrinkage estimator, 44
- significance, 36
- standard error, 16
- state space, 47
- stationary, 44
 - probability, 47
 - strictly, 44
 - weakly, *see* stationary
- statistic, 37
- statistical inference, 28
- Stein's Paradox, 43
- sup-norm, 12
- support, 9

- surjective, 8
- test statistic, 34
- time series, 44
- trail state function, 70
- transient, 47
- transition probability, 47
 - n -step, 47
- triangular inequality, 7
- truncation error, 64
- unbiased, 14
- uncorrelated, 22
 - asymptotic, 46
- variational monte carlo, 70
 - parameters, 71

Editorial Acknowledgment E [REDACTED] Jonsson

pre@aps.org

Tue 2018-05-15 19:53

To: Marius Jonsson <mariujon@student.matnat.uio.no>;

Re: E [REDACTED]

Standard error estimation by an automated blocking method
by Marius Jonsson

Dear Marius Jonsson,

The editors acknowledge receipt of this manuscript on 15 May 2018
and are considering it as a Regular Article in Physical Review E.
When sending correspondence regarding this manuscript please refer to
the code number [REDACTED]

Physical Review requests ORCID identifiers from corresponding authors.
Please provide your ORCID identifier at <https://authors.aps.org/Profile/>.
If you do not have an ORCID identifier you may obtain one at <https://orcid.org/>.

We understand your submission of this manuscript to certify the
following:

- The paper represents original work of the listed authors.
- The manuscript as presented accurately reflects the scientific results.
- All of the authors made significant contributions to the concept, design, execution, or interpretation of the research study.
- All those who made significant contributions were offered the opportunity to be listed as authors.
- All of the listed authors are aware of and agree to the submission of this manuscript.
- The manuscript has not been published, and is not now and will not be under consideration by another journal while it is considered here.
- The authors have provided information to the editors about relevant unpublished manuscripts, including whether any version of this manuscript was previously considered by an APS journal.
- The authors accept the established procedures for selecting manuscripts for publication.


To obtain current information regarding the status of your manuscript,
you may consult the Author Status Inquiry System at
<http://authors.aps.org/STATUS/>.

Yours sincerely,

Juan-Jose Lieten-Santos
Associate Editor
Physical Review E
Email: pre@aps.org
<http://journals.aps.org/pre/>

Celebrating 125 Years of the Physical Review
<https://journals.aps.org/125years> #PhysRev125

Please verify the following information and notify the Editorial office of any corrections:

Code number: E 
Journal: Physical Review E Regular Article
Received: 15 May 2018
Section: Computational Physics

Preprint number:

(x) Please supply the eprint archive number, if available.

PhySH Concepts:
Statistical methods (Primary)|Data analysis|Monte Carlo
methods|Probability theory|Real & complex analysis|Stochastic
processes & statistics

Title: Standard error estimation by an automated blocking method

Collaboration:

1 Author(s):

Marius Jonsson

Standard error estimation by an automated blocking method

Marius Jonsson¹

¹*Department of Physics, University of Oslo, N-0316 Oslo, Norway*

(Received 15 May 2018)

Background: The sample mean $\bar{X} = \sum_{i=1}^n X_i$ is probably the most popular estimator of the expected value in all sciences, the standard error is given by the square root of the variance $\text{Var}(\bar{X})$.

Purpose: This work aims at providing stringent and modern treatment of the Blocking method (sensu Flyvbjerg & Petersen 1989), a popular way to compute $\text{Var}(\bar{X})$ for correlated data.

Methods: Linear algebra, results from multivariate probability theory, real analysis and Fisherian statistical inference were used. The method validation was performed by Metropolis-Hasting-type sampling and autoregressive models.

Results: Here a new approach to estimation of $\text{Var}(\bar{X})$ for time series data is presented. The method applies to stationary data, such as stationary Markov chains and other stationary time series. The method complexity is bounded by $12n + O(\log_2 n)$ floating point operations, but this can be reduced to $n + O(1)$ in large computations. The convergence in relative error squared is better than $\propto n^{-1/2}$. The method is insensitive to the probability distribution of the observations. It is proven that only a small part of the correlation structure is relevant to the convergence rate of the method. From this, the 1989 Blocking method follows as a corollary. The result is also used to propose a hypothesis test to survey the relevant part of the correlation structure. The method is sufficiently robust to operate without supervision. An algorithm and sample code showing the implementation is available for PYTHON, C++ and R. This code is available for download². Method validation using autoregressive AR(1)- and AR(2)-processes and physics applications are included. This method has an accuracy similar to dependent bootstrapping, but scales in $O(n)$ -time. The method is easily adapted to multithread applications and time series larger than computing cluster memory.

Conclusions: By applying stringent mathematics, the Blocking method was automated, which will be helpful for all users of this method for computing standard errors of means generated from correlated data series.

PACS numbers: 02.50.-r, 02.70.Ns, 02.70.Rr, 02.70.Ss

I. INTRODUCTION

Estimation of the variance of sample means $\bar{X} = (1/n) \sum_{i=1}^n X_i$ is essential in natural sciences[1]. This is because \bar{X} is a typical estimator of the expected value of X_i if the elements of $\{X_i\}_{i=1}^n$ are identically distributed. The variance of the mean is the expected square error of the estimate. Already in 1867, Chebyshev[2] explained this by showing that if the observations have expected value μ , variance σ^2 , and \bar{X} has finite non-zero variance, $\text{Var}(\bar{X})$, then for any real number $k > 0$

$$P(|\bar{X} - \mu| > k[\text{Var}(\bar{X})]^{1/2}) \leq \frac{1}{k^2} \quad (\text{Chebyshev's ineq.})$$

If the observations are independent and identically distributed, the variance of the mean is easily obtained by setting $\text{Var} \bar{X} = \sigma^2/n$ [3], but for correlated data, the computation is more complicated[4]. Here, however, I show that if there is some integer $d > 1$ such that $n = 2^d$, and X_1, \dots, X_n are observations from a stationary time series, then the 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 [5]. This refers to forming a new sample of data by taking the mean of every pair of subsequent observations. To be precise, the *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

$$B_i = \frac{1}{2} (A_{2i-1} + A_{2i}) \quad (1)$$

Such transformations are applied in many areas of probability theory, and in 1989, Flyvbjerg and Petersen [5] made popular a method where blocking transformations reduced the correlations of the data, and proposed a way to estimate the variance of the sample mean. In the present study, the mathematics is developed and automation of the method is given. Thus, I elevate the rigor of the treatment to the level of modern mathematics and provide an automation that is robust enough to operate without supervision. I recycle much of the philosophy of Flyvbjerg and Petersen (1989), but the mathematics is new. 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 equation (1) until the correlation of observations is no longer significantly different from zero. The results show that the behavior of this process is determined by the autocovariance of the observations, $\gamma(1)$. 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. By that, it is possible to speed up calculations of the standard error of the mean by factors of thousands or more, as compared to dependent bootstrapping or other methods of complexity $O(n^2)$ or $O(n \log n)$. I give the algorithm and sample code².

The preliminary background (definitions and theorems) is

² PYTHON, R and C++ code is available: github.com/computative/block

scattered throughout in text blocks where they are required. First, proof of the blocking method is given and second, an automation which 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 is 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, \dots , but also the parts of it which it is possible to sample: That is, the first n observations: X_1, X_2, \dots, 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 $\text{Cov}(X_i, X_j)$ only depends on the difference $h = |i - j|$ for all $1 \leq i, j \leq n[4]$. A time series is *strictly stationary* if the cumulative distribution function (cdf) of all sets of the form $\{X_i, X_{i+1}, \dots, X_{i+k}\}$ equals the cdf of the set $\{X_{i+j}, X_{i+j+1}, \dots, X_{i+j+k}\}$ for all $i, j, k[4]$. A strictly stationary time series with finite variance is stationary [6]. The function $\gamma(h) = \text{Cov}(X_i, X_{i+h})$ is the *autocovariance* of $\{X_j\}_{j=1}^\infty$. ►

$\text{Var}(\bar{X})$ will be estimated by a quantity σ_k^2/n_k , which is subject to an error e_k , for $k \in \{0, 1, 2, \dots\}$. 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 $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$, 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, \dots has autocovariance $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$. Then for every $\varepsilon > 0$ there exists a natural number K such that $e_k < \varepsilon$ if $K \leq k \leq d$ for the time series X_1, X_2, \dots, X_{2d} .

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

Theorem. If X_1, X_2, \dots is a strictly stationary time series such that $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ with $\lim_{n \rightarrow \infty} \text{Var}(\sum_{i=1}^n X_i) = \infty$ and $(|X_i|^{2+a}) < \infty$, for some $a > 0$ then M_j is a test statistic which is asymptotic χ_{d-j}^2 -distributed under the hypothesis $\gamma_k(1) = 0$ for all $k \geq j$. The rejection region are all values M_j larger than $q_{d-j}(1 - \alpha)$ for all $1 \leq j \leq d - 1$.

Using this theorem, calculations are automated in 6 steps, see figure 6. Users primarily interested in the algorithm can jump ahead to section 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 $\text{Cov}(X_i, X_j) = 0$. A time series is *uncorrelated* if $\gamma(h) = 0$ for all $h \neq 0$ and *asymptotic uncorrelated* if the autocovariance $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$. The matrix consisting of elements $\Sigma_{ij} = \gamma(|i - j|)$ is the *covariance matrix* of \mathbf{X} , and $\hat{\gamma}(h)$ is the *sample covariance* and $\hat{\sigma}^2$ is the *sample variance* if they are given by

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

Subscripts are used on the variables to indicate that $(X_k)_1, (X_k)_2, \dots$ and \mathbf{X}_k are *subject to k blocking transformations* if they are related to X_1, X_2, \dots and \mathbf{X} by k repeated transformations of the type (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, \dots is subject to zero transformations, $\{(X_0)_i\} = \{X_i\}$ 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 $\{X_i\}$ 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: $\bar{X}_k, \sigma_k^2, \hat{\sigma}_k^2, \gamma_k(h), \hat{\gamma}_k(h)$. Assuming $h = |i - j|$ and using the definition of the blocking transformation, equation (1), and the distributive property of the covariance, it is clear that

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

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

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n^2} \text{Cov} \left[\sum_{i=1}^n X_i, \sum_{j=1}^n X_j \right] \\ &= \frac{1}{n^2} \left[n\gamma(0) + (n-1)\gamma(1) + \dots + \gamma(n-1) \right. \\ &\quad \left. + (n-1)\gamma(-1) + (n-2)\gamma(-2) + \dots + \gamma(1-n) \right] \\ &= \frac{\sigma^2}{n} + \frac{2}{n} \sum_{h=1}^{n-1} \left(1 - \frac{h}{n} \right) \gamma(h) \quad \text{if } \gamma(0) = \sigma^2 \end{aligned} \quad (4)$$

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. It also justifies that γ_k exists for all integers $k \geq 0$:

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

Proof. We first show that the time series $(X_k)_1, (X_k)_2, \dots$ 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 $i \geq 1$ and so the base case is trivially satisfied. For the induction step, write

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

For the covariance; the elements of $\{X_i\}$ are stationary, and therefore $\text{Cov}(X_i, X_j)$ 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 equation (3), it is true for $k+1$, since equation (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 $k \geq 0$. The proof works for any smaller time series $\{(X_k)_i\}_{i=a}^b$ for $a \geq 1$. By taking $b = n_k$ this proves \mathbf{X}_k is stationary.

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

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

which provides the induction step. \blacksquare

Using lemma 1 and equation (4) it is clear that any estimate of $\text{Var}(\bar{X})$ using σ_k^2/n_k has truncation error given by

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

The next proposition is crucial, it says that if $\gamma_0(1), \gamma_1(1), \dots, \gamma_{d-1}(1)$ are known, then the behavior of the truncation error e_k is known.

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

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

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

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

Similarly, sum equation (3)

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

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

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

The following corollary is the most important take away. 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 & Petersen (1989) blocking method. Explanation of this is provided in the discussion.

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

1. *if there exists $k \in \mathbb{N}$ such that for all $i \leq k \leq j$ either: $\gamma_k(1) > 0$ or $\gamma_k(1) \geq 0$ or $\gamma_k(1) = 0$, then the sequence of errors e_k is strictly decreasing or decreasing or constant on $i \leq k \leq j$, respectively.*
2. *if there exists some $k \in \{0, 1, \dots, d-1\}$ 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 $j \geq k$.*

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 $i \leq k < j$. If $u, v \in \{i, i+1, \dots, j+1\}$ 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

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

Now, by adding e_v to each side of the inequality, the first part is proven. To obtain the result in the case $0 \leq \gamma_k(1)$, replace $<$

with \leq 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 [2]. The induction step follows for $k + 1$ since equation (3) says that $\gamma_{k+1}(i) = \delta_{i0}\sigma_k^2/2$. This proves $\gamma_j(1)$ is zero for all $j \geq k$, so the error is constant by what was proved above. ■

These results will be useful in the automation of the blocking method later in the paper. 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, prior users of the 1989 blocking method 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 hints at the conclusions to come, a bit more work is required. Start by a lemma and some interesting consequences of 'blocking', in doing so, introduction of an interesting sequence of functions $\{f_k\}$ is appropriate. Fix $k \in \mathbb{N}$ and define:

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

$$\begin{aligned} \frac{\gamma_{k+1}(h)}{2^{2(k+1)}} &\stackrel{(1)}{=} \frac{2^{-2}}{2^{2(k+1)}} [\gamma_k(2h-1) + 2\gamma_k(2h) + \gamma_k(2h+1)] \\ &\stackrel{(10)}{=} \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma(2^k(2h-2) + i) + 2 \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma(2^k(2h-1) + i) + \sum_{i=1}^{2^{k+1}-1} f_k(i)\gamma(2^k(2h) + i) \\ &\stackrel{(9)}{=} \sum_{i=1}^{2^{k+1}+1-1} \gamma(2^{k+1}(h-1) + i) [f_k(i) + 2f_k(i-2^k) + f_k(i-2^{k+1})] \stackrel{\text{lemma 2}}{=} \sum_{i=1}^{2^{k+1}+1-1} f_{k+1}(i)\gamma(2^{k+1}(h-1) + i). \end{aligned}$$

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 $i \leq 2^j$ or $2^{k+1} + 2^j \leq i$ from equation (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. Consider the following corollary

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

$$2^{2k}\gamma_k(1) = \gamma(1) + 2\gamma(2) + \dots + 2^k\gamma(2^k) + (2^k - 1)\gamma(2^k + 1) + \dots + \gamma(2^{k+1} - 1). \quad (11)$$

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

Using these results, everything is now set to finalize the investigation of γ under blocking transformations.

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

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

Proof. See the appendix. ■

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

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

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

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

Assume now that equation (10) is true for some $k \geq 1$ and write

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

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

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

since \log_2 is a monotonous function. Thus by lemma 3 and

the triangular inequality

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

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

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

which was required. \blacksquare

The blocking method follows immediately.

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

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 $k \geq K$ then $\gamma_k < \varepsilon/2$ on \mathbb{N} . Moreover, if $d \geq k$, then by the triangular inequality

$$\begin{aligned} e_k = |e_k| &\leq \frac{2}{n_k} \sum_{h=1}^{n_k-1} \left(1 - \frac{h}{n_k}\right) |\gamma_k(h)| \leq \frac{2}{n_k} \sum_{h=1}^{n_k-1} |\gamma_k(h)| \\ &\leq \frac{\varepsilon}{n_k} \sum_{h=1}^{n_k-1} 1 = \varepsilon \frac{n_k - 1}{n_k} < \varepsilon \end{aligned}$$

which is the theorem. \blacksquare

III. AUTOMATING CALCULATIONS

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

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

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

where $|\cdot|$ denotes the determinant [7]. 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 [8]. The components of \mathbf{Y} are independent if and only if there exists $\sigma_1, \dots, \sigma_n > 0$ such that $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ [7]. A random variable $X_i \sim \chi_1^2$ is *chi-square distributed* with 1 degree of freedom (later: df) if it is the square of a *standard normal* random variable [2]. That means there exists a random variable $Z \sim N(0, 1)$ such that $X_i = Z^2$. A sum of n independent chi square random variables with 1 df, $\sum_{i=1}^n X_i$, is *chi-square distributed* with n df 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 $(Y_1, \dots, Y_n)^\top = \mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$, then $(\mathbf{Y} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi_n^2$ (*Proof.* Write $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$. What is the distribution of \mathbf{Z} and $\mathbf{Z}^\top \mathbf{Z}$?). If $X \sim \chi_n^2$, the $100(1-\alpha)$ -percentile is the value $q_v(1-\alpha) \in \mathbb{R}$ such that $P(X > q_v(1-\alpha)) = \alpha$. Chi-square percentiles are tabulated in the appendix. If X_1, X_2, \dots is a strictly stationary time series, which is asymptotic uncorrelated such that $\text{Var}(X_i) < \infty$ and $\lim_{n \rightarrow \infty} \text{Var}(\sum_{i=1}^n X_i) = \infty$, and $\langle |X_i|^{2+k} \rangle < \infty$ for some $k > 0$, then Ibragimov [9] has proved that the central limit theorem holds [10].

The Fisherian approach to inference is the most common type of inference in natural sciences [11] and will be used in the following. Suppose there is a pdf $f(\mathbf{y}; \boldsymbol{\theta})$ for \mathbf{Y} that depends on a parameter $\boldsymbol{\theta}$ and it is necessary to test whether there exists evidence that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ or if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ based on the observations \mathbf{Y} . Let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{Y})$ be an estimator for $\boldsymbol{\theta}$. The hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is called the *null hypothesis*, denoted H_0 . It is common to pick H_0 such that the consequences of an incorrect test conclusion is minimized. The *alternative hypothesis* is denoted by H_a is typically the negation of H_0 . Suppose there exists a $\mathbb{R}^k \rightarrow \mathbb{R}$ -function $G(\hat{\boldsymbol{\theta}})$ such that G has known pdf g whenever H_0 is true, then $G(\hat{\boldsymbol{\theta}})$ is called a *test statistic*. The values of $G(\hat{\boldsymbol{\theta}})$ that are sufficiently unlikely according to g whenever H_0 is true, is called the *rejection region*. And if the estimated value $G(\hat{\boldsymbol{\theta}})$ is in the rejection region, H_0 is rejected in favor of H_a . Prior investigation, $\alpha \in (0, 1)$ is chosen such that

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

Here $P(A|B)$ denotes conditional probability. Since g is known, this determine 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 $G(\hat{\theta})$ 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 favour of H_a . The *likelihood* L of θ is the function $L(\theta) = f(\mathbf{Y}; \theta)$. The estimator $\hat{\theta}$ maximizing L is called the *maximum likelihood estimator*, and is asymptotically multivariate normal distributed [8]. The estimator $\hat{\gamma}_k(1)$ is a maximum likelihood estimator if \mathbf{X} is multivariate normal. ►

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

$$M_j = (\hat{\gamma}_j - \boldsymbol{\mu})^\top \Sigma_j^{-1} (\hat{\gamma}_j - \boldsymbol{\mu}) \sim \chi_{d-j}^2. \quad (12)$$

Hence $(\hat{\gamma}_k - \boldsymbol{\mu})^\top \Sigma_j^{-1} (\hat{\gamma}_k - \boldsymbol{\mu})$ has a known distribution (according to preliminaries 3), which means that it is a test statistic for the hypothesis test

$$\begin{aligned} H_0 : \gamma_j(1) &= 0 \text{ for all } j \geq k, \\ H_a : \text{There exists } k \geq j \text{ such that } \gamma_k(1) &\neq 0 \end{aligned} \quad (13)$$

The idea is to pick the smallest j such that the hypothesis

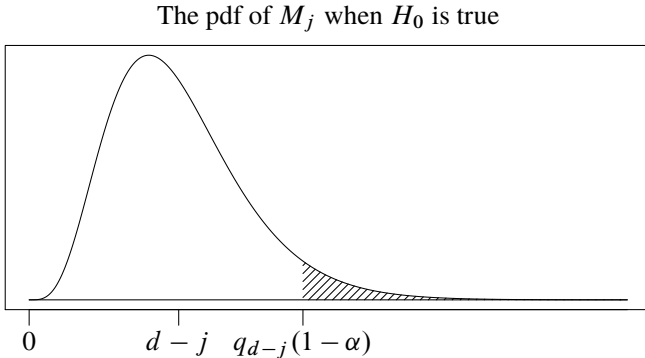


FIG. 1. Whenever H_0 is true, the pdf of M_j is known and plotted above. The test concludes that H_0 is false if the observed value of M_j is sufficiently unlikely. That is, if the observed value of M_j is larger than $100(1 - \alpha)$ -percentile for a suitable α ; the shaded area represent a probability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

test finds no evidence for H_a and take $\text{Var}(\bar{X}) = \sigma_j^2/n_j$. Thus an appropriate estimator is $\widehat{\text{Var}}(\bar{X}) = \hat{\sigma}_j^2/n_j$. This

works according to preliminaries 3: Whenever H_0 is true, the distribution of M_j is known (chi square with $d - j$ df by equation 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, i.e. that the error becomes constant, and iterating further does not improve the estimate. See figure 1 for illustration. However, an expression of Σ_j has to be determined. In this paper, the following approximation will be used

$$\Sigma_j = \text{diag}(\sigma_j^4/n_j, \dots, \sigma_{d-j}^4/n_{d-j}). \quad (14)$$

The benefit is that inversion of Σ_j is easy. Corollary 3 explains why it is a reasonable approximation. But before proving this, more work is needed.

A. Covariance matrix of $\hat{\gamma}_i(1)$ and the matrix Σ_j

Computing the covariance matrix directly is impractical. However, developing linear algebra for the task provide 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. If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ with Σ is singular. Then if A, B are symmetric $n \times n$ -matrices, and there exists some $n \times r$ -matrix Q of rank r such that $\Sigma = QQ^\top$, then

$$\text{Cov}(\mathbf{Y}^\top A \mathbf{Y}, \mathbf{Y}^\top B \mathbf{Y}) = 2\text{Tr}(\Sigma A \Sigma B) + 4\boldsymbol{\mu}^\top A \Sigma B \boldsymbol{\mu}. \quad (15)$$

For proof see [12]. ►

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

Lemma 4. Assume $\mathbf{1}$ denotes the vector of ones, $\mathbf{X} \sim N(m\mathbf{1}, \sigma^2 I_n)$ and $\mathbf{Y} = \mathbf{X} - \bar{X}\mathbf{1}$. 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 $\Sigma_Y = QQ^\top$ and

$$\Sigma_Y = \frac{\sigma^2}{n}(nI_n - \mathbf{1}\mathbf{1}^\top). \quad (16)$$

Proof. First note that Y_i is a linear combination of elements of \mathbf{X} , because \mathbf{X} is multivariate normal, that means 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 $\langle Y_i \rangle = \langle X_i - \bar{X} \rangle = m - m = 0$. To get equation (16), notice that the covariance matrix of \mathbf{X} is diagonal, that 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

$$\begin{aligned} (\Sigma_Y)_{ij} &= \text{Cov}(Y_i, Y_j) = \text{Cov}(X_i - \bar{X}, X_j - \bar{X}) = \\ &= \sigma^2 \delta_{ij} - \frac{1}{n} \sum_{k=1}^n \text{Cov}(X_i, X_k) - \frac{1}{n} \sum_{k=1}^n \text{Cov}(X_j, X_k) + \text{Var } \bar{X}, \\ &= \sigma^2 \delta_{ij} - \frac{1}{n} \sum_{k=1}^n \sigma^2 \delta_{ik} - \frac{1}{n} \sum_{k=1}^n \sigma^2 \delta_{jk} + \frac{\sigma^2}{n} = \sigma^2 \delta_{ij} - \frac{\sigma^2}{n} \end{aligned} \quad (17)$$

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

$$\Sigma_Y \mathbf{q}_k \stackrel{(17)}{=} \frac{\sigma^2}{n} (n\mathbf{q}_k - \underbrace{\mathbf{1}(\mathbf{1}^\top \mathbf{e}_k - \mathbf{1}^\top \mathbf{e}_n)}_{=1-1=0}) = \sigma^2 \mathbf{q}_k.$$

which proves that σ^2 is an eigenvalue of Σ_Y with multiplicity $n-1$. And since Σ_Y is symmetric, it has a spectral decomposition[13]:

$$\Sigma_Y = \sigma^2 \sum_{k=1}^{n-1} \mathbf{q}_k \mathbf{q}_k^\top = \sigma^2 [\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_{n-1}] [\mathbf{q}_1^\top \ \mathbf{q}_2^\top \ \dots \ \mathbf{q}_{n-1}^\top]^\top.$$

So if $Q = \sigma[\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_{n-1}]$ then Q is an $n \times (n-1)$ -matrix and $\Sigma_Y = QQ^\top$. Moreover, according to the spectral theorem[13], the dimension of $\text{Span}\{\mathbf{q}_1, \dots, \mathbf{q}_{n-1}\}$ equals the multiplicity of σ^2 . Which means that the columns of Q are $n-1$ linearly independent vectors, which also equals its rank. ■

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

$$S_i = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & \dots & 0 \end{bmatrix} \quad T_i = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \quad (18)$$

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

$$\mathbf{X}_i = T_i T_{i-1} \dots T_1 \mathbf{X} \quad (19)$$

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

$$\Gamma_i = \frac{1}{2n_i} T_1^\top T_2^\top \dots T_i^\top (S_i + S_i^\top) T_i \dots T_1 \quad (20)$$

These matrices are interesting because they generate the estimator $\hat{\gamma}_i(1)$ from the vector \mathbf{Y} whose probability distribution is multivariate normal:

Proposition 3. *The matrices $\{\Gamma_i\}$ are symmetric. Hence if $\mathbf{Y} = \mathbf{X} - \bar{X}\mathbf{1}$, then $\mathbf{Y}^\top \Gamma_i \mathbf{Y}$ is a quadratic form and $\mathbf{Y}^\top \Gamma_i \mathbf{Y} = \hat{\gamma}_i(1)$.*

Proof. Fix $0 \leq i \leq d-1$. It's clear that Γ_i is symmetric by construction;

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

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

$$\begin{aligned} n_i \mathbf{Y}^\top \Gamma_i \mathbf{Y} &\stackrel{(20)}{=} \frac{1}{2} \mathbf{Y}^\top T_1^\top \dots T_i^\top S_i T_i \dots T_1 \mathbf{Y} \\ &\quad + \frac{1}{2} [\mathbf{Y}^\top T_1^\top \dots T_i^\top S_i T_i \dots T_1 \mathbf{Y}]^\top \\ &= [T_i \dots T_1 \mathbf{Y}]^\top S_i T_i \dots T_1 \mathbf{Y} \stackrel{(19)}{=} \mathbf{Y}_i^\top S_i \mathbf{Y}_i \end{aligned} \quad (21)$$

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

$$\begin{aligned} (\mathbf{Y}_{i+1})_k &\stackrel{(1)}{=} \frac{1}{2} [(\mathbf{Y}_i)_{2k-1} + (\mathbf{Y}_i)_{2k}] \\ &= \frac{1}{2} [(\mathbf{X}_i)_{2k-1} - \bar{X}_i + (\mathbf{X}_i)_{2k} - \bar{X}_i] \\ &\stackrel{(1)}{=} (\mathbf{X}_{i+1})_k - 2\frac{1}{2}\bar{X}_i \stackrel{\text{lemma 1}}{=} (\mathbf{X}_{i+1})_k - \bar{X}_{i+1}, \end{aligned} \quad (22)$$

where lemma 1 was used twice to get $\bar{X}_i = \bar{X} = \bar{X}_{i+1}$. Third, using the definition of the matrices S_i from equation (18) notice that S_i shifts the indices of vectors by one:

$$\begin{aligned} n_i \mathbf{Y}^\top \Gamma_i \mathbf{Y} &\stackrel{(21)}{=} \mathbf{Y}_i^\top S_i \mathbf{Y}_i = \mathbf{Y}_i^\top (S_i \mathbf{Y}_i) \\ &= ((\mathbf{Y}_i)_1, \dots, (\mathbf{Y}_i)_{n_i})^\top ((\mathbf{Y}_i)_2, \dots, (\mathbf{Y}_i)_{n_i}, 0) \\ &= \sum_{h=1}^{n_i-1} (\mathbf{Y}_i)_{h+1} (\mathbf{Y}_i)_h \stackrel{(2)(22)}{=} n_i \hat{\gamma}_i(1), \end{aligned} \quad (23)$$

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

Experts will immediately see that the above result is easily generalized to $\hat{\gamma}_i(k)$ for any $k \geq 0$ by considering the operators, S_i^k , by raising to a power $k \in \mathbb{Z}$. But according to proposition 1 it suffices to consider $k = 1$. In this case, the following three quantities determine the expression for the covariance matrix Σ . Consider the following lemma

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

$$\begin{aligned} n_i \mathbf{1}^\top \Gamma_i \mathbf{1} &= n_i - 1, \\ n^2 n_i n_j \text{Tr}[\Gamma_i \Gamma_j] &= \frac{1}{2} n_i^2 (n_i - 1), \\ 2n n_i n_j \mathbf{1}^\top \Gamma_i \Gamma_j \mathbf{1} &= 2n_j (n_i - 1) - n_i. \end{aligned}$$

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 equation (18),

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

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

$$\begin{aligned} T_i \sum_{k=1}^{n_i-1} \mathbf{e}_k &= T_i \mathbf{e}_1 + T_i \mathbf{e}_2 + T_i \mathbf{e}_3 + \cdots + T_i \mathbf{e}_{n_i-1} \\ &\stackrel{(18)}{=} \frac{1}{2} \mathbf{e}_1 + \frac{1}{2} \mathbf{e}_1 + \frac{1}{2} \mathbf{e}_2 + \cdots + \frac{1}{2} \mathbf{e}_{n_i} = \sum_{k=1}^{n_i} \mathbf{e}_k. \end{aligned} \quad (24)$$

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

$$\begin{aligned} n_i \mathbf{1}^\top \Gamma_i \mathbf{1} &= \frac{1}{2} \sum_{u=1}^n \mathbf{e}_u^\top T_1^\top T_2^\top \cdots T_i^\top (S_i + S_i^\top) T_i \cdots T_1 \sum_{v=1}^n \mathbf{e}_v \\ &\stackrel{(24)}{=} \frac{1}{2} \sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \mathbf{e}_u^\top (S_i + S_i^\top) \mathbf{e}_v \\ &= \frac{1}{2} \sum_{u=1}^{n_i} \sum_{v=2}^{n_i} \mathbf{e}_u^\top \mathbf{e}_{v-1} + \frac{1}{2} \sum_{u=1}^{n_i} \sum_{v=1}^{n_i-1} \mathbf{e}_u^\top \mathbf{e}_{v+1} = n_i - 1, \end{aligned}$$

where orthonormality of $\{\mathbf{e}_k\}$ was used in the final step. Next is necessary to show $T_i T_i^\top = (1/2)I_{n_i}$ for all $1 \leq i \leq d-1$. 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 [14]:

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

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

$$2nn_i n_j \mathbf{1}^\top \Gamma_i \Gamma_j \mathbf{1} = 2n_i n_j - n_i - 2n_j. \quad (25)$$

Prove now two more properties of $\{\mathbf{e}_k\}$: First, see that if M is any $n \times n$, then a diagonal element $m_{kk} = \mathbf{e}_k^\top M \mathbf{e}_k$, so $\text{Tr}(M) = \sum_{k=1}^n \mathbf{e}_k^\top M \mathbf{e}_k$, 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

$$\begin{aligned} \sum_{k=1}^{n_j-1} \mathbf{e}_k^\top T_{j+1}^\top \cdots T_{j+h}^\top M T_{j+h} \cdots T_{j+1} \mathbf{e}_{k+1} \\ = 2^{-2h} \sum_{k=1}^{n_{j+h}-1} \mathbf{e}_k^\top M \mathbf{e}_{k+1} + K \sum_{k=1}^{n_{j+h}} \mathbf{e}_k^\top M \mathbf{e}_k. \end{aligned} \quad (26)$$

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

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

Which proves the base case. To get the induction step, assume the hypothesis true for h then define the matrix $N = T_{j+h+1}^\top M T_{j+h+1}$. 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 $T_j T_j^\top = 2^{-1} I_{n_j}$, as well as cyclic permutation of the factors and write $\text{Tr}[\Gamma_i \Gamma_j]$ in the following way:

$$4n_i n_j \text{Tr}[\Gamma_i \Gamma_j] \stackrel{(20)}{=} 2^{-2j} \text{Tr} \left[T_{j+1}^\top \cdots T_i^\top (S_i + S_i^\top) T_i \cdots T_{j+1} (S_j + S_j^\top) \right] \quad (27)$$

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

$$\begin{aligned} \sum_{k=1}^{n_j} \mathbf{e}_k^\top T_{j+1}^\top \cdots T_i^\top S_i T_i \cdots T_{j+1} S_j^\top \mathbf{e}_k \\ = \sum_{k=1}^{n_j-1} \mathbf{e}_k^\top T_{j+1}^\top \cdots T_i^\top S_i T_i \cdots T_{j+1} \mathbf{e}_{k+1} \\ \stackrel{(26)}{=} 4^{-(i-j)} \sum_{k=1}^{n_i-1} \mathbf{e}_k^\top S_i \mathbf{e}_{k+1} + K \sum_{k=1}^{n_i} \underbrace{\mathbf{e}_k^\top S_i \mathbf{e}_k}_{\mathbf{e}_k^\top \mathbf{e}_{k-1}=0} \end{aligned}$$

This term equals $4^{-(i-j)}(n_i - 1)$ since $\mathbf{e}_k^\top S_i \mathbf{e}_{k+1} = \mathbf{e}_k^\top \mathbf{e}_k = 1$. Make the replacement $S_i \mapsto S_i^\top$ throughout the above equation, in which case the term will evaluate to zero since $\mathbf{e}_k^\top S_i^\top \mathbf{e}_{k+1} = \mathbf{e}_k^\top \mathbf{e}_{k+2} = 0$. The third and fourth terms from equation (27) are evaluated in a similar way. The sum of all four terms is $2 \cdot 4^{-(i-j)}(n_i - 1)$, hence

$$n^2 n_i n_j \text{Tr}[\Gamma_i \Gamma_j] \stackrel{(27)}{=} 2 \frac{1}{4} 2^{-2j} n^2 4^{-(i-j)}(n_i - 1) = \frac{1}{2} n_i^2 (n_i - 1),$$

which is the final part of the lemma. \blacksquare

Proposition 4. *If there is some $m \in \mathbb{R}$ such that the vector $\mathbf{X} \sim \mathcal{N}(m\mathbf{1}, \sigma^2 I_n)$, then the expected value of $\hat{\gamma}_i(1)$ is $-\sigma_i^2(n_i - 1)/n_i^2$. Furthermore, the covariance matrix of $(\hat{\gamma}_0(1), \dots, \hat{\gamma}_{d-1}(1))^\top$ has elements*

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

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

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

$$n_i \langle \hat{\gamma}_i(1) \rangle = \sum_{j=1}^{n_i-1} \underbrace{\langle (\mathbf{X}_i)_j (\mathbf{X}_i)_{j+1} \rangle}_{\gamma_i(1)+m^2=0+m^2} + \underbrace{\langle \bar{X}^2 \rangle}_{\frac{\sigma_i^2}{n_i}+m^2} - \underbrace{\langle ((\mathbf{X}_i)_j + (\mathbf{X}_i)_{j+1}) \bar{X} \rangle}_{2(m^2+\sigma_i^2/n_i)},$$

and the first part is proven. Assume now that $\mathbf{Y} = \mathbf{X} - \bar{X}\mathbf{1}$. To get the covariance matrix, note that by lemma 4, \mathbf{Y} is multivariate normal with expected value $\boldsymbol{\mu}$ and there exist a $n \times (n-1)$ -matrix \mathbf{Q} of rank $n-1$ such that the covariance matrix $\Sigma_{\mathbf{Y}} = \mathbf{Q}\mathbf{Q}^\top$. According to proposition 3, Γ_i and Γ_j are symmetric, so according to preliminaries 4

$$\begin{aligned} \text{Cov}(\hat{\gamma}_i(1), \hat{\gamma}_j(1)) &= \text{Cov}(\mathbf{Y}^\top \Gamma_i \mathbf{Y}, \mathbf{Y}^\top \Gamma_j \mathbf{Y}) \\ &\stackrel{(15)}{=} 2\text{Tr}(\Sigma_{\mathbf{Y}} \Gamma_i \Sigma_{\mathbf{Y}} \Gamma_j), \end{aligned} \quad (28)$$

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

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

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

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

Multiply each side of the equation by 2 and recall equation (28) above, which proves the proposition true. \blacksquare

Corollary 3. Assume there is some $m \in \mathbb{R}$ such that the vector $\mathbf{X} \sim \mathcal{N}(m\mathbf{1}, \sigma^2 I_n)$. Then the covariance matrix of \mathbf{y}_j is

$$\Sigma_j = \text{diag}(\sigma_j^4/n_j, \dots, \sigma_{d-j}^4/n_{d-j})$$

to leading order in $1/n_k$ and the variance of $\hat{\gamma}_j(k)$ is exactly

$$\text{Var}(\hat{\gamma}_j(1)) = \left(\frac{\sigma_j}{n_j} \right)^4 \left[2 + n_j(n_j - 1)(n_j - 2) \right],$$

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

B. Algorithm

Using this, and summarizing the results from the two previous sections it is clear that

Theorem 2. If X_1, X_2, \dots is a strictly stationary time series such that $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ with $\lim_{n \rightarrow \infty} \text{Var}(\sum_{i=1}^n X_i) = \infty$ and $\langle |X_i|^{2+a} \rangle < \infty$, for some $a > 0$ then M_j is a test statistic which is asymptotic χ_{d-j}^2 -distributed under the hypothesis $\gamma_k(1) = 0$ for all $k \geq j$. The rejection region are all values M_j larger than $q_{d-j}(1-\alpha)$ for all $1 \leq j \leq d-1$.

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

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

Pick some significance level α . It's 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 $M_k \leq q_{d-k}(1-\alpha)$. Using this k , make the final estimate for the variance $\text{Var}(\bar{X}) = \hat{\sigma}_k^2/n_k$.

The method has built in safety features (see figure 5). This is necessary because the method shall 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) $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ and (iii) the χ^2 -approximation works. Therefore if the method does not conclude that H_0 is true for any $k \geq d-1$, one of these are false, and the fault is caught with an i.f.-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 is little data available. See figure 5 for sample code of one implementation, or use flow chart of 2 for an overview.

An upper bound of the complexity of the method is $12n + O(\log_2 n)$. Consider the sample code in figure 5. The only contributions at order n are from the while loop. At iteration number i , the while loop can be computed using exactly $6n_i + 4$ floating point operations. Using geometric series, the total floating point operations are

$$\begin{aligned} \text{cost} &= \sum_{j=0}^{d-1} (4 + 6n_i) = 4d + 6 \cdot 2^d \sum_{j=0}^{d-1} 2^{-j} \\ &= 4d + 12(n-1) \leq 12n + O(\log_2 n) \quad \text{as } n \rightarrow \infty \end{aligned} \quad (29)$$

For time consuming computations which requires 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 section III D, but first consider these test results.

Flow chart of algorithm

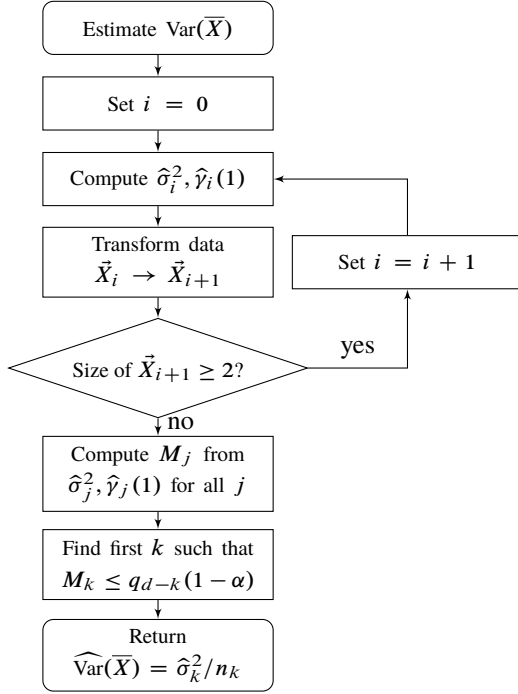


FIG. 2. Flow chart of algorithm. The idea is to return the estimate of $\text{Var}(\bar{X})$, as $\hat{\sigma}_k^2/n_k$ for the smallest value of k such that there is no evidence that $\gamma_j(1) \neq 0$ for $j > k$. This is sensible because then, according to corollary 1, there is no reason to believe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

C. Test results

The method validation uses autoregressive models, because Wold decomposition justifies their use in modelling stationary time series [4]. Moreover, $\text{Var}(\bar{X})$ can be computed exactly for autoregressive models. This makes them ideal for our purpose.

Preliminaries 5. An autoregressive model of order p denoted $\text{AR}(p)$ 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 are independent, identically distributed with zero expected value and constant variance σ^2 for all t . The autoregressive models used are order 1 and 2, they have autocovariance in closed form and are stationary [4]. Since for all stationary processes $\gamma(h) = \gamma(-h)$, $\text{Var}(\bar{X})$ is known for $\text{AR}(1)$ and $\text{AR}(2)$ - processes because ABC Define 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 $|z_i| > 1$. ►

It is known that causal $\text{AR}(1)$ and $\text{AR}(2)$ process are both asymptotic uncorrelated and stationary[4]. Two tests of the algorithm are presented. First, 6080 causal random $\text{AR}(1)$ and

TABLE I. Regression summary for the $\text{AR}(p)$ processes: Regression table for the $\text{AR}(1)$ process (left) and $\text{AR}(2)$ (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 $\text{AR}(1)$ and $\text{AR}(2)$ 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	$< 10^{-16}$

$\text{AR}(2)$ processes were generated. According to preliminaries 5, that means the exact value of $\text{Var}(\bar{X})$ can be computed from the autoregressive coefficients ϕ_i for each of the $\text{AR}(p)$ 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 $\gamma(\tau) \leq \gamma(0)e^{-1}$). Gamma regression is suitable because the observations of ϵ^2 are independent, identically gamma-distributed, and the model is $\log(\epsilon^2) = \beta_0 + \beta_1 \log(n/\tau)$. The expected relative error squared is

$$\epsilon^2 = e^{\beta_0} \left(\frac{n}{\tau}\right)^{\beta_1} \quad (30)$$

Maximum likelihood-estimates of β_j and standard errors are given for the causal $\text{AR}(1)$ and $\text{AR}(2)$ processes in table I. The table shows that if there is 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 $\text{AR}(1)$ - and $\text{AR}(2)$ -processes respectively. It is also evident that the convergence rate of the $\text{AR}(2)$ processes are faster than the $\text{AR}(1)$ processes. Amongst the processes, the type of autocovariance was the main differentiator of the $\text{AR}(p)$ -processes. For $p = 1$, the autocorrelation is of the form $\gamma(h)/\sigma^2 = \phi^{-h}$, whilst 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 ($p \geq 0.34$, t-test), and therefore the difference in ϵ^2 between the $\text{AR}(1)$ and $\text{AR}(2)$ experiments are attributed to the $\gamma(h)$ according to preliminaries 2. A plot of the regression analysis is given in figure 3 and the regression summaries, see table I.

Second, two standard textbook physics applications were studied. The variance of the mean energy was estimated using the Flyvbjerg & Petersen (1989) blocking method and the automated blocking method. The estimates were compared with dependent bootstrapping using geometric simulation type [15]. In either application, the autocorrelation functions bore resemblance of the $\text{AR}(1)$ 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

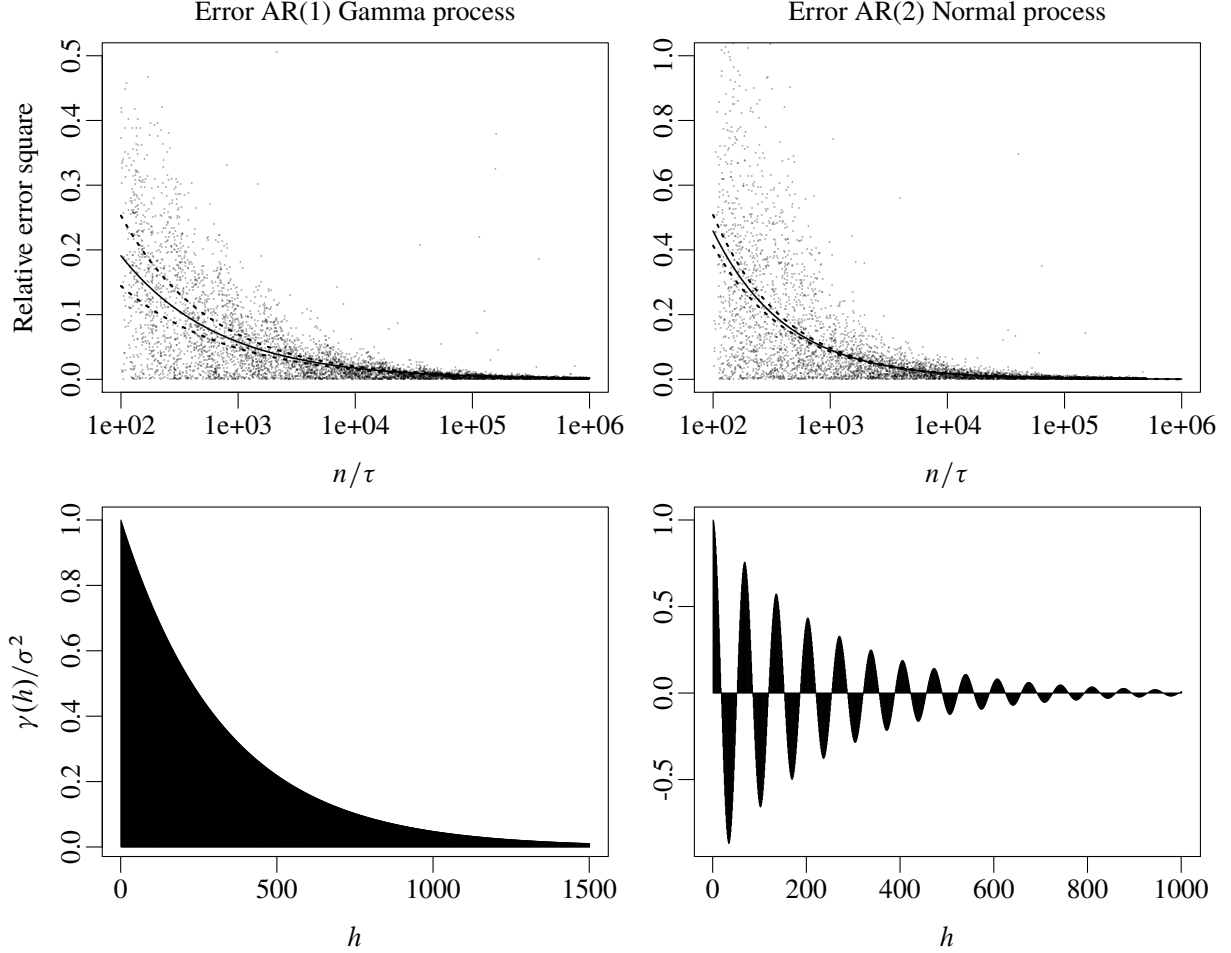


FIG. 3. Relative error squared of two autoregressive models versus observations per time autocorrelation time-constant. There is exponential convergence rate for two common correlation structures in natural sciences. Left: AR(1) autocorrelation is positive with exponential decay, typical of Metropolis-type Markov chains, where it is expected that the observations correlate positively. The process was distributed Gamma(1,1) Right: AR(2)-processes have exponential decay, but may be oscillatory as here, the process was distributed multivariate standard normal. It was found that the method was insensitive to the distribution of the observations ($p \geq 0.34$). 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 modelled 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 $n \gtrsim 25000\tau$.

a Fokker-Plank type-prior [16]. 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 $T = 2.4$ [17]. The energy was sampled from a Boltzman distribution at significance level $\alpha = 0.05$ according to a χ^2 goodness-of-fit test [2] using a stationary, time-reversible Markov chain constructed using the Metropolis algorithm [18]. The autocorrelation time constant was measured to be $\tau = 16$. A plot of the results are contained in figure 4.

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 which can be allocated on this node is 2^d real numbers.

Assume now that the total length of the time series is

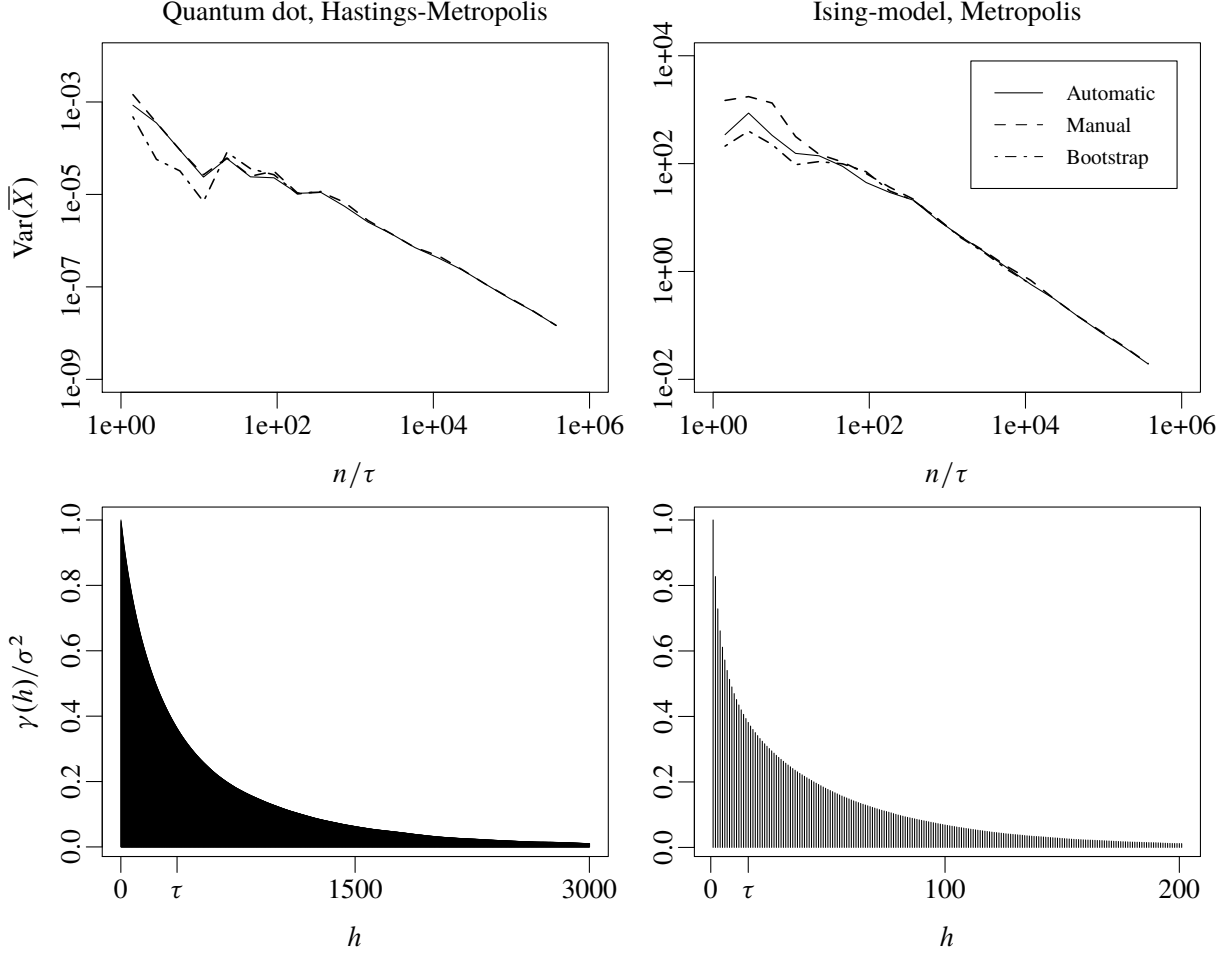


FIG. 4. Case study of two textbook physics applications for mean variance estimation. The variance of the mean energy was estimated using manual- and automatic Blocking methods and compared with dependent bootstrapping using geometric simulation. The estimates are plotted on top, whilst the autocorrelation depicted on the bottom. Left: Two electron quantum dot with trial 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 $\text{Var}(\bar{X}) = 1.46 \cdot 10^{-8}$ at 10^8 samples according to all three models, and by the above discussion the error of the variance estimate is expected to correct to near 2nd digit. Right: Ising model implementation of 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 $\text{Var}(\bar{X}) = 1.49 \cdot 10^{-1}$ according to all three methods, and there is reason to believe that the circa the first two digits are correct.

$n = 2^D$, 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 are preserved. Now on the chunk, apply blocking transformations $D - d$ times and form $\mathbf{X}_{D-d} = T_{D-d} T_{D-d-1} \cdots T_1 \mathbf{X}$ where T_i is defined in equation (18). 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 is of all the transformed chunks is $2^k \cdot 2^{d-k} = 2^d$, as required. On the parent node, or personal computer doing the final estimate, write the data to memory by concatenating 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 figure 2.

The computational cost of this is low. Performing the transformations T_i as it is done in the code of figure 5 require precisely n_{i-1} floating point operations, as you can check. So the total number of floating point operations is computed using geometric series:

$$\sum_{j=1}^{D-d} n_{j-1} = \sum_{j=0}^{D-d-1} 2^{D-k-j} = \frac{2^D - 2^d}{2^{k-1}} \quad (31)$$

According to equation (29), it is necessary to add the $12(2^d - 1) + 4d$ floating point operations which the parent node must

Python implementation of algorithm

```
# 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
n = 2**(d-k)
answer = s[k]/n
```

FIG. 5. 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 from github.com/computative/block

compute at the end, so the total cost is bounded above by

$$\text{cost} \stackrel{(31)}{\leq} \underbrace{2^D}_{=n} \frac{1}{2^{k-1}} + 2^d \left(12 - \underbrace{\frac{1}{2^{k-1}}}_{\leq 1}\right) \leq n + O(1) \quad \text{as } n \rightarrow \infty$$

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 [15, 19]

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 $\hat{\mu}_j = \sum_{i=(j-1)2^{D-k}}^{j2^{D-k}} X_i$ then write

$$\bar{X} = \frac{1}{2^D} \sum_{i=1}^{2^D} X_i = \sum_{j=1}^{2^k} \frac{1}{2^D} \underbrace{\sum_{i=(j-1)2^{D-k}}^{j2^{D-k}} X_i}_{2^{D-k} \hat{\mu}_j} = \frac{1}{2^k} \sum_{j=1}^k \hat{\mu}_j \quad (32)$$

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 is 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 does not have to be read back into memory for post-processing later.

Assume the time series is split in 2^k chunks of length 2^{D-k} :

$$X_1, X_2, \dots, X_{2^{D-k}}, \underbrace{X_{2^{D-k}+1}, \dots, X_{2^{D-k}+1}, \dots, X_{2^D}}_{\text{chunk number } i \equiv \vec{X}}$$

(Step 1) Transform chunk i by:

$$\vec{X}_{D-d} = T_{D-d} \cdots T_2 T_1 \vec{X}$$

$$\vec{Y}_i \equiv \vec{X}_{D-d}$$

(Step 2) Repeat for all i .

(Step 3) Reassemble transformed chunks

$$(\vec{Y}_1, \vec{Y}_2, \dots, \underbrace{\vec{Y}_i, \dots, \vec{Y}_{2^k}}_{\text{Transformed chunk } i})^T$$

(Step 4) Using this vector, apply algorithm (figure 2)

FIG. 6. 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 $T_{D-d} T_{D-d-1} \cdots T_2 T_1$. 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 $(Y_1, Y_2, \dots, Y_{2^k})^T$. 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 figure 2. See section III D for more details including the definitions of the numbers D, d and k .

DISCUSSION

The present study provides the following four new contributions: (a) Rigorous proof of the Flyvbjerg & Petersen (1989) blocking method conforming to the standard of modern mathematics. The results give prospects for new research with relevance to any blocking method. (b) A new 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 new contributions include proof of the behavior of the method. Furthermore, proposition 1 outline a new 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 & Petersen (1989) 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 [5] 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 [5]. For mathematically interested readers, this can present a distraction since fixed points of any function $T_j : A \rightarrow B$ is defined when $A = B$ [20–22]. In this context $A \neq B$ since $A = \mathbb{R}^{n_i}$ and $B = \mathbb{R}^{n_i/2}$ (see the section 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 (1989) appear unaware of this, because they state that the blocking method converges if $\gamma(t) \propto 1/t$ [5], which does not capture the essence, as was proved in theorem 1. In fact, their blocking method works whenever X_1, X_2, \dots 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 & Petersen (1989). First, because $\gamma(h)$ needs not be proportional to $1/h$ (note that proposition 1, which places no restriction on $\gamma(h)$, although finite variance is required). Therefore, $\gamma(h)$ can have any shape. Second, Flyvbjerg & Petersen (1989) constrain $\gamma(h)$, exactly n degrees of freedom (since γ is a function $\gamma : \{0, 1, \dots, n_1\} \rightarrow \mathbb{R}$), whilst the new 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 $s = \sum_{h=1}^{n-1} (1 - h/n)\gamma(h)$, appear frequently in the study of time series [4], together proposition 1 and lemma 1

together imply that $(2/n)s$ is determined up to a constant ¹ by the set $\{\gamma_k(1)\}$. Moreover, according to theorem 1, $\gamma_k(h) \rightarrow 0$ uniformly on \mathbb{N} under blocking transformations. In this way, blocking transformations are intimately linked to s . 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 $e_k = (2/n)s$). 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 [23] and applications such as those by Schäfer and Strimmer [24]. Other proposed applications of proposition 1 is the proof of corollary 1, which explains the behavior of the 1989 blocking method, and why the automatic blocking method works. The corollary shows that the estimates, $\text{Var}(\bar{X})$, 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 $k \geq j$. In case (i) the truncation error $e_k = 0$ since if the components of \mathbf{X}_k are uncorrelated, then $\text{Cov}((\mathbf{X}_k)_i, (\mathbf{X}_k)_j) = \sigma_k^2 \delta_{ij}$ by definition, so $\gamma_k(h) = 0$ for $h \geq 1$ 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 $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$.

The algorithm for the automation is new, although the concept of automatic computation of $\text{Var}(\bar{X})$ is not new. In physics, the most well known automated method for standard error estimation is perhaps dependent bootstrapping [15]. Dependent bootstrapping is useful when n is small or if n is large and the required precision is small. According to Politis and Romano (1994), the method has asymptotically valid procedures even for multivariate parameter spaces. However, high precision estimates for large data sets are often needed. Flyvbjerg & Petersen [5] 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 $\hat{\sigma}^2$. Typically, this method works well since $\gamma(h) \propto 1/h$, which is commonly used in physics. However, this is not always 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 correlation structures, for example oscillatory AR(2)-like processes introduced here: The new automation works for causal AR(p) processes for any p , and places no assumption on the shape of $\gamma(t)$. In addition, the present paper provides updates which makes it convenient

¹ That constant is σ_0^2 , as can be proved by iteratively using proposition 1

to use the method for multithread computing. Another alternative method is the Gamma method proposed by Wolff (2004) [25]. The Gamma method works well with correct set up, with errors claimed to be lower than those of the automatic blocking method proposed here. Wolff (2004) 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 (2004) 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 (2004) recommendation, Lee et al. [26] are proponents of a method which Wolff has called binning (which is essentially a blocking method). Lee et al. (2011) proposed inequalities that can be used to automate calculations. However, this approach requires estimates that may or may not be available. The new automated blocking method has none of the complications mentioned above.

The automation uses one approximation in computing M_j . It was assumed that Σ_j is diagonal even though it is only diagonal to leading order in $1/n$. This approximation can be avoided by using lemma 4. The benefit of the approximation is that the inversion of Σ_j can be simplified (since if $\Sigma_j = \text{diag}(r_j, \dots, r_{d-j})$, then $\Sigma_j^{-1} = \text{diag}(1/r_j, \dots, 1/r_{d-j})$ is inverted using only $d - j$ floating point operations [13]).

The method validation of $\text{AR}(p)$ processes is new (and 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. $\text{AR}(1)$ and $\text{AR}(2)$ processes were chosen because Wold decomposition says that the random component of any time series can be expressed as an autoregressive model [4]. $\text{AR}(1)$ and $\text{AR}(2)$ correlation functions are the two most common ones encountered in modeling of time series. The two text book cases, quantum dot and the Ising model, show that their correlation structures were similar to the $\text{AR}(1)$ -processes. Using equation 30, 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. Whilst 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 [25]. However, unlike the estimates due to Wolff (2004), table I 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 & Petersen (1989) blocking method, regardless of n (see figure 4), is fully automatic and works in $O(n)$ -time.

CONCLUSION AND PERSPECTIVES

A rigorous proof of the blocking method (Flyvbjerg and Petersen 1989) is a main result of the present study. That method has become one of the industry standards for estimating standard errors $\text{Var}(\bar{X})^{1/2}$ 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 new 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 github.com/computative/block.

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 $\text{Var}(\bar{X})$. The problem is that for large k , the data available to estimate $\gamma_k(1)$ is 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 MCMC that are common in computational physics (see for example [27]), such that it is made explicit for which methods theorem 2 continues to hold.

ACKNOWLEDGMENTS

The author is indebted to Professor Morten Hjorth-Jensen (Department of physics, University of Oslo) and Associate professor Henrik Flyvbjerg (Department of Micro- and Nanotechnology, Technical University of Denmark) for valuable support during the development of the results. Thanks go to Professors Ørnulf Borgan and Anders Rygh Swensen (Department of mathematics, University of Oslo) for helpful comments to drafts of the manuscript. Professors Galin L. Jones (School of Statistics, University of Minnesota) and Richard C. Bradley (Department of mathematics, Indiana University) had technical comments which were helpful and the author is grateful to Kenneth Ravn (University of Oslo) for contributing essential pattern-finding skills for proposition 4.

APPENDIX

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

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

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

$$\begin{aligned} \sum_{i=1}^{2^{k+1}-1} f_k(i) &= 1 + 2 + \cdots + 2^k + (2^k - 1) + \cdots + 2 + 1 \\ &= 2^k + 2 \sum_{i=1}^{2^k-1} i = 2^k + 2 \frac{2^k - 1}{2} 2^k = 2^{2k} \end{aligned}$$

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

$$\begin{aligned} 1 &= f_1(1) = f_0(1) + 2f_0(1 - 1) + f_0(1 - 2) = 1 + 0 + 0 \\ 2 &= f_1(2) = f_0(2) + 2f_1(2 - 1) + f_0(2 - 2) = 0 + 2 + 0 \\ 1 &= f_1(3) = f_0(3) + 2f_2(3 - 1) + f_0(3 - 2) = 0 + 0 + 1 \end{aligned}$$

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

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

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, 2012).
 - [2] J. L. Devore and K. L. Berk, *Modern Mathematical Statistics with Applications* (Springer, 2012), 2nd ed.
 - [3] M. H. DeGroot and M. J. Schervish, *Probability and Statistics* (Pearson Education, Inc., 2012), 4th ed.
 - [4] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications with R examples* (Springer, 2017), 4th ed.
 - [5] H. Flyvbjerg and H. Petersen, *The Journal of Chemical Physics* pp. 461–466 (1989).
 - [6] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting (Springer Texts in Statistics)* (Springer, Switzerland, 2016), 3rd ed.
 - [7] A. Agresti, *Foundations of linear and generalized linear models* (Wiley & Sons, Inc., 2015), 1st ed.
 - [8] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis* (CRC Press, 2014), 3rd ed.
 - [9] I. A. Ibragimov, *Teor. Veroyatnost. i Primenen.* **20**, 135 (1975).
 - [10] R. C. Bradley, *Rocky Mountain Journal of Mathematics* **17**, 95 (1987).
 - [11] B. Efron, *The American Statistician* **40**, 1 (1986).
 - [12] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables* (Marcel Dekker, Inc, 1992).
 - [13] D. C. Lay, *Linear algebra and its applications* (Addison-Wesley, 2012), 4th ed.
 - [14] M. Hazewinkel, *Encyclopedia of mathematics* **9** (1993).
 - [15] D. N. Politis and J. P. Romano, *Journal of the American Statistical Association* **89**, 1303 (1994).
 - [16] C. W. Gardiner, *Handbook of stochastic methods for physics, chemistry and the natural sciences* (Springer-Verlag, 1985), 2nd ed.
 - [17] M. Plischke and B. Bergersen, *Equilibrium Statistical Physics* (World scientific, 2006), 3rd ed.
 - [18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *The Journal of Chemical Physics* **21**, 1087 (1953).
 - [19] D. N. Politis and H. White, *Econometric Reviews* **23**, 53 (2006).
 - [20] M. Hazewinkel, *Fixed point* (1989).
 - [21] E. J. Borowski and J. M. Borwein, *Fixed point* (1989).
 - [22] J. N. McDonald and N. A. Weiss, *A course in real analysis* (Academic press, 2012), 2nd ed.
 - [23] C. Stein, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Contributions to the Theory of Statistics* **1**, 197 (1956).
 - [24] J. Schäfer and K. Strimmer, *Statistical Applications in Genetics and Molecular Biology* **4**, 1 (2005).
 - [25] U. Wolff, *Computer Physics Communications* (2004).
 - [26] R. Lee, G. Conduit, N. Nemeć, P. López Ríos, and N. Drummond, *Physical review E* **83**, 066706 (2011).
 - [27] G. L. Jones, *Probability Surveys* **1**, 299 (2004).

TABLE II. Chi-square 90-,95- and 99-percentiles: Percentiles in the chi squared distribution for $1 \leq d - k \leq 48$, which suffices for any error estimation with $\leq 10^{14}$ observations, at the three significance levels that performed best, $1 - \alpha = 0.99, 0.95$ and 0.90 . See for example [2] 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