Asymptotic Properties of the Plug-in Estimator of the Discrete Entropy Under Dependence

Raffaello Seri and Mario Martinoli

***Abstract*— We consider the estimation of the entropy of a discretely-supported time series through a plug-in estimator. We provide a correction of the bias and we study the asymptotic properties of the estimator. We show that the widely-used correction proposed by Roulston (1999) is incorrect as it does not remove the $O\left(N^{-1}\right)$ part of the bias while ours does. We provide the asymptotic distribution and we show that it differs when the values taken by the marginal distribution of the process are equiprobable (a situation that we call *degeneracy*) and when they are not. We introduce estimators of the bias, the variance and the distribution under degeneracy and we study the estimation error. Finally, we propose a goodness-of-fit test based on entropy and give two motivations for it. The theoretical results are supported by specific numerical examples.**

***Index Terms*— Discrete entropy, time series, plug-in estimator, bias correction, degenarate distribution, goodness-of-fit test.**

I. Introduction

SINCE the seminal works of [107] and [67], the entropy plays a central role in information and communication theory. The Shannon entropy and the Kullback–Leibler divergence can be observed from several perspectives but, overall, they can be classified as uncertainty measures.

The Shannon entropy has been applied to many fields of information theory, including the estimation of the entropy rate of information sources (see [59], [87], [88]), the estimation of functionals of probability distributions (see [6], [55]), the analysis of texts and symbol sequences (see [10], [65], [103]) and machine learning research (see, e.g., [111]). Other important branches of application of Shannon entropy are psychology (see [72], [73]), physics (see [2], [18], [38], [121]), and economics and finance (see [68], [76], [126]).

It is therefore natural that many efforts have focused on its estimation. Many papers have been devoted to the estimation of the entropy for observations with a continuous distribution function or with a discrete one, as well as for independent and identically distributed (iid) observations or for dependent data.

Manuscript received November 26, 2019; revised April 5, 2021; accepted August 5, 2021. Date of publication September 1, 2021; date of current version November 22, 2021. The work of Mario Martinoli was supported by the PRIN Grant 2017 "How Good Is Your Model? Empirical Evaluation and Validation of Quantitative Models in Economics." *(Corresponding author: Mario Martinoli.)*

Raffaello Seri is with DiECO, Università degli Studi dell'Insubria, 21100 Varese, Italy.

Mario Martinoli is with the Institute of Economics and EMbeDS, Sant'Anna School of Advanced Studies, 56127 Pisa, Italy (e-mail: m.martinoli@santannapisa.it).

Communicated by E. Gassiat, Associate Editor for Probability and Statistics.
Digital Object Identifier 10.1109/TIT.2021.3109307

The main contributions related to the case of data with continuous distributions exploit nonparametric estimation methods, such as kernel or nearest-neighbor estimators. Among these, we quote several papers dealing with iid data (see [1], [15], [45], [71]), some of which investigate the behavior of the bias, and a few works that tackle the case of time series (see, for instance, [37] and [52]). The most important results achieved in the literature are reviewed by [13].

However, the case that attracted most attention, and that we will consider in this paper, is the one of the entropy of data coming from discretely-supported distributions, a situation that applies both to genuinely discrete data and to discretized (also called symbolized) ones. In the iid case, the most natural estimator is the so-called maximum likelihood or plug-in estimator, obtained replacing the discrete probability with its maximum likelihood estimator. Two facts about it were early recognized in a 1954 unpublished report by Miller and Madow. As witnessed by a summary of this paper in [72, p. 45], the authors showed that the asymptotic behavior of the estimated entropy depends on whether all values assumed by the discrete process have the same probability or not. In the second case, the statistic is asymptotically normally distributed, while in the first case it is asymptotically distributed as a chi-square. The second discovery (see also [20], [77], [101]) is that the estimator is biased in finite samples.

The case of iid discrete observations has further been explored by several authors (see [6], [7], [40], [47], [87], [89], [124], [125]). This has led to the availability of a large number of alternatives to the plug-in estimator, like the Grassberger ([38], [40]), the best upper bound ([87]), the unseen ([116]) as well as several polynomial approximation estimators ([54], [55], [119]). Many of these papers provided methods to overcome the bias. We refer to [54], [55] for a complete review of the most recent contributions in this field, and to [54, Sec. V] for an extensive simulation study comparing the performance of several estimators.

The extension of the iid case to the one of dependent data has proceeded along two directions, both of them associated with different estimation strategies. To clarify what we mean, we consider a stationary stochastic process $\{\dots, x_{t-1}, x_t, x_{t+1}, \dots\}$ and we use the informal notation $p(\cdot)$ $(p(\cdot\,|\,\cdot))$ to denote the (conditional) probability mass or density associated with its argument. The symbol $\mathbb{E}$ denotes expectation.

The first direction is associated with the *entropy rate* appearing in the Shannon–McMillan–Breiman theorem. This

result states that, when $N \to \infty$, minus the normalized logarithm of the density of the time series $\{x_1, \ldots, x_N\}$, i.e. $-\frac{1}{N} \ln p(x_1, \ldots, x_N)$, converges to the entropy rate $-\mathbb{E} \ln p(x_t | \ldots, x_{t-1})$, defined as minus the expected value of the logarithm of the conditional probability of x_t given its past $\{\ldots, x_{t-1}\}$ (see, e.g., [3]). The relevance of this result to information theory lies in the *asymptotic equipartition property* (see [3] for some references). Estimation in this direction has been thoroughly investigated in the literature, see [32], [39], [58], [64], [65]. In dynamical systems, a similar quantity to the entropy rate appears in the *metric entropy* defined by Kolmogorov and Sinai (see, e.g., [24]), whose estimation has been considered in [41], [103], [110].

The second direction is connected with quantities appearing in several measures or tests of statistical dependence, both in time series and dynamical systems, like in [29], [30], [57], [82]–[85], [94], [113], [118]. In this case, time series data are used to estimate the same formula of the iid case, $-\mathbb{E} \ln p(x_t)$, where the probability p is defined by the marginal distribution of the process (see, e.g., [102]). This second approach can be extended to compute the *block entropy* $-\mathbb{E} \ln p(x_{t-k+1}, \ldots, x_{t-1}, x_t)$, i.e. the entropy of the block $\{x_{t-k}, \ldots, x_{t-1}, x_t\}$. By using the properties of conditional probabilities, one can then compute the *conditional* or *differential entropy* $-\mathbb{E} \ln p(x_t | x_{t-k}, \ldots, x_{t-1})$, the entropy of x_t conditionally on its recent past $\{x_{t-k}, \ldots, x_{t-1}\}$, as a difference of block entropies. This fact seems to draw together the two directions as, letting k diverge, one recovers the entropy rate of the Shannon–McMillan–Breiman theorem. However, the two directions are generally associated with different estimation strategies and problems. Indeed, estimating the differential entropy $-\mathbb{E} \ln p(x_t | x_{t-k}, \ldots, x_{t-1})$ as an approximation of the entropy rate $-\mathbb{E} \ln p(x_t | \ldots, x_{t-1})$ may introduce a severe bias (see [103, p. 416] and [32, p. 75]), unless the process is a Markov chain (see [54, pp. 2859-2860]). However, even bounding ourselves to the block entropy with small k or $k = 1$, the estimation efforts in this direction have been limited. Indeed, most articles study the plug-in estimator or variants thereof and apply to the dependent case formulas derived for iid observations without modification (see [48, p. 102], [103, p. 416] and [102]). Among these, we highlight the paper of [102], who evaluates the bias and the asymptotic distribution of the entropy. However, as we will show below, the bias formulas proposed by [102], and thus his bias correction, are not correct, as are his asymptotic variance formulas.

In this paper we consider the entropy appearing in the second direction outlined above. We analyze in detail the plug-in entropy estimator H_N obtained replacing the probabilities of each value assumed by the process with their natural estimators based on a sequence of dependent observations of length N. Our aim is to fill some gaps in the literature, mainly concerning its consistency and asymptotic distribution, and to correct some incorrect results.

First of all, we show that, under stationarity, the observed entropy H_N converges almost surely to a limit H_∞ which is a random variable. Under stationary ergodicity, this limit H_∞ becomes a fixed value. We characterize the bias of H_N showing that it disappears asymptotically and, if the process is a fortiori α-mixing with $\sum_{n=1}^{\infty} \alpha(n) < \infty$, H_N has bias $O\left(N^{-1}\right)$. We then propose a bias correction and we compare it with the one proposed by [102]. The evidence shows that the correction in [102] does not remove the $O\left(N^{-1}\right)$ part of the bias while ours does. Despite the wrong correction proposed by [102], during the last twenty years many authors have considered his formulas, fostering the propagation of the error in information theory (see [44], [66], [86], [90]–[92], [122]), neurosciences (see [19], [51], [63], [70], [95]), physiology (see [123]), engineering (see [56]) and organizational research (see [17]).

Subsequently, we provide asymptotic distributional results under α-mixing. We show that in general the statistic, when centered and scaled by $\sqrt{N}$, has a normal asymptotic distribution but, under a condition that we call *degeneracy*, it must be scaled by N and it converges in distribution to a weighted sum of chi-square random variables. The name "degeneracy" is due both to the fact that the variance of the asymptotic normal distribution is null (or degenerate) and to the fact that the entropy behaves like a degenerate V-statistic (see [105, Chapters 5 and 6]). We then propose some estimators of the bias of the entropy. One of them exploits an autocorrelation-consistent covariance matrix estimator (see [80] and [5]). The second one applies when the process is a Markov chain and features the fundamental matrix of the chain (see, for instance, [61] and [104]). Finally, we give a result on the average error induced by the estimation of bias. Our outcomes demonstrate that the Markov bias correction is more precise than the estimator based on the autocorrelation-consistent covariance matrix estimator, and the bias correction slightly increases the variance of the estimator, but the mean squared error is generally improved by the corrections. In the non-degenerate case, we also address estimation of the variance of the entropy. Under degeneracy, the asymptotic distribution depends on some weights that can be estimated. However, this impacts directly on the significance level of tests. Indeed, we show that the Kolmogorov distance between the exact asymptotic distribution and the estimated one is $O_{\mathbb{P}}\left(N^{-1/2}\right)$.

At last, we provide an application of the entropy to a goodness-of-fit test for the marginal distribution of the process and we report the results of a simulation study showing the finite-sample properties of the test.

Throughout the paper, we apply our results to two different examples: a dichotomized first-order autoregressive process and the Gilbert–Shannon–Reeds model (see, e.g., [12]).

The article is organized as follows. Section II introduces some notations that will be used throughout the paper. Section III investigates the limiting behavior of the entropy and proposes formulas for its bias. Section IV introduces the estimators of bias, variance and distribution under degeneracy, and provides results on the errors in the estimation. Section V propose a test of goodness-of-fit based on the entropy. Section VI wraps up the main conclusions. Section VII contains the proofs of the results. The Appendix contains an application to stationary non-ergodic data.

II. Notation

We introduce some notation.

We write $\mathbb{N}$ for the positive integers, $\mathbb{N}_0$ for the non-negative integers and $\mathbb{R}$ for the real numbers. We follow the convention that $0 \ln 0 = 0$.

For sequences, when $n \to \infty$, we use $a_n \simeq b_n$ when $a_n = b_n \cdot (1 + o(1))$, $a_n \asymp b_n$ when $b_n/C \leq a_n \leq C b_n$ for $\infty > C > 0$ and n large enough, $a_n \ll b_n$ when $a_n = o(b_n)$, $a_n \lesssim b_n$ when $a_n \leq C b_n$ (with a_n and b_n non-negative) for $\infty > C > 0$ and n large enough. We use the same notation when the limit is with respect to a continuous variable.

We use capital bold letters, such as $\mathbf{A}$, to denote matrices and lowercase bold letters, such as $\mathbf{a}$, to denote vectors. Let $\boldsymbol{\iota}$ be a vector of ones, $\mathbf{U}$ a square matrix of ones, $\mathbf{I}$ the identity matrix, $\mathbf{0}$ a matrix or a vector of zeros. If a confusion is possible, the dimension will be indicated through an index, as in $\boldsymbol{\iota}_N$. For a vector $\mathbf{a}$, let $\bar{\mathbf{a}}$ be the vector containing the reciprocals of the elements of $\mathbf{a}$. Let $\mathrm{dg}(\mathbf{a})$ be a diagonal matrix having $\mathbf{a}$ on its diagonal. Let $\mathrm{tr}(\mathbf{A})$ be the trace of $\mathbf{A}$, i.e. the sum of the diagonal elements of a square matrix $\mathbf{A}$. For a suitable matrix $\mathbf{A}$, $\mathbf{A}'$ is its transpose, $\mathbf{A}^\star$ its conjugate transpose, $\mathbf{A}^{-1}$ its inverse and $\mathbf{A}^+$ its Moore–Penrose pseudoinverse. The element-wise power of a vector or a matrix is denoted by $\mathbf{A}^{\odot b}$ (so that $\bar{\mathbf{a}} = \mathbf{a}^{\odot(-1)}$), while $\mathbf{A}^b$ is the usual power obtained multiplying $\mathbf{A}$ by itself b times. The element of $\mathbf{A}$ in position (i,j) is denoted as $\mathbf{A}_{ij}$ or $[\mathbf{A}]_{ij}$; the matrix with generic element a_{ij} is denoted $[a_{ij}]$.

The notation $\|\cdot\|_p$ indicates the Schatten norm, that is $\|\mathbf{A}\|_p := [\sum_i (s_i(\mathbf{A}))^p]^{\frac{1}{p}}$ where s_i is the i-th singular value of $\mathbf{A}$, i.e. the square root of the i-th non-negative eigenvalue of $\mathbf{A}^\star\mathbf{A}$. We will use mainly the nuclear norm $\|\cdot\|_1$ and the Frobenius norm $\|\cdot\|_2$, also written $\|\cdot\|_F$. When applied to a vector $\mathbf{a}$, the notation $\|\cdot\|_{L^p}$ denotes the vector norm defined as $\|\mathbf{a}\|_{L^p} := (\sum_i |a_i|^p)^{\frac{1}{p}}$; when applied to a matrix $\mathbf{A}$, it denotes the matrix norm induced by the vector norm as $\|\mathbf{A}\|_{L^p} := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_{L^p}}{\|\mathbf{x}\|_{L^p}}$.

We use $\sim$, as in $X \sim \mathcal{N}(\mu, \sigma^2)$, to denote that X is distributed as the random variable on the right-hand side. The notations $\to_{\mathbb{P}}$ and $\to_{\mathcal{D}}$ denote convergence in probability and in distribution respectively. For $\sim$ and $\to_{\mathcal{D}}$ we sometimes write, with a small abuse of notation, that $X_n \to_{\mathcal{D}} X$ where X is a random variable with a given distribution. The symbols $\mathbb{E}$ and $\mathbb{V}$ respectively denote the expectation and the variance of a random variable or vector.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $T : \Omega \to \Omega$ a measurable transformation. T is called *measure-preserving* if $\mathbb{P}(TA) = \mathbb{P}(A)$ for any $A \in \mathcal{A}$. Let us define a trajectory of a stochastic process as $x(\omega) = \{\dots, x_1(\omega), x_2(\omega), \dots\}$ (note that in the following we will systematically neglect the argument ω). We can identify T with the *shift transformation*, i.e. as the function such that $x_t(T\omega) = x_{t+1}(\omega)$, so that $x(T\omega) = \{\dots, x_2(\omega), x_3(\omega), \dots\}$. In this case, a stochastic process $\{\dots, x_1, x_2, \dots\}$ is stationary if the sequences $\{\dots, x_1, x_2, \dots\}$ and $\{\dots, x_{k+1}, x_{k+2}, \dots\}$ have the same distributions, for every $k > 0$. The set A is said to be *invariant* under T if $\mathbb{P}(A \triangle TA) = 0$. The set of invariant sets under T is a σ-algebra denoted $\mathcal{I}$. T is called *ergodic* if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ for any $A \in \mathcal{I}$. A process is ergodic iff:

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K-1} \mathbb{P}\{\{\dots, x_{k+1}, \dots\} \in B, \{\dots, x_1, \dots\} \in A\} = \mathbb{P}\{\{\dots, x_1, \dots\} \in B\}\, \mathbb{P}\{\{\dots, x_1, \dots\} \in A\}$$

for any measurable set A and B. For two sub-σ-fields $\mathcal{G}$ and $\mathcal{H}$ of $\mathcal{A}$, we define the strong and the uniform mixing coefficients as:

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G\in\mathcal{G}, H\in\mathcal{H}} |\mathbb{P}(G \cap H) - \mathbb{P}(G)\,\mathbb{P}(H)|,$$
$$\varphi(\mathcal{G}, \mathcal{H}) = \sup_{G\in\mathcal{G}, H\in\mathcal{H}} |\mathbb{P}(H|G) - \mathbb{P}(H)|.$$

Let us define $\mathcal{F}_{-\infty}^t = \sigma(\dots, x_{t-1}, x_t)$ and $\mathcal{F}_{t+m}^\infty = \sigma(x_{t+m}, x_{t+m+1}, \dots)$ the σ-algebras generated by the random variables inside the parentheses. We define:

$$\alpha(m) = \sup_t \alpha\left(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty\right),$$
$$\varphi(m) = \sup_t \varphi\left(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty\right).$$

We say that the process is *strong* or *α-mixing* if $\lim_{m\to\infty} \alpha(m) = 0$ and *uniform* or *φ-mixing* if $\lim_{m\to\infty} \varphi(m) = 0$.

III. Main Results

Consider a stochastic process $\{x_1, \dots\}$ with finite support, i.e. such that each $x_i \in \{1, 2, \dots, B\}$. This process can be genuinely discrete or can come from the symbolization of a stochastic process $\{\widetilde{x}_1, \dots\}$ whose support is divided into B intervals. The choice of a finite and bounded B may seem restrictive at first, as several recent papers consider in detail what happens when B is infinite or diverges with N (see [6], [54], [55], [119]). However, this assumption will turn out to be natural in Section IV as the estimation procedures we use require finite B and, moreover, it allows us to concentrate on the main aim of this paper, ruling out several interesting but unexpected behaviors (see, in particular, [6]).

We suppose that the process $\{x_1, \dots\}$ is stationary. Note that under this assumption the probability on $\{1, 2, \dots, B\}^{\mathbb{N}}$ can be extended to a probability on $\{1, 2, \dots, B\}^{\mathbb{Z}}$ and this will allow us to refer interchangeably to one-sided $\{x_1, \dots\}$ or two-sided processes $\{\dots, x_1, \dots\}$. The hypothesis of stationarity can be generalized using the concept of asymptotic mean stationarity (see [21], [42], [43], [49]), but we will not pursue this improvement here.

The results we are going to prove are stated for the estimation of the entropy computed on the marginal distribution of the process. They can be easily adapted to the computation of the block entropy for blocks of length k. Indeed, let us consider a process $\{y_1, \dots\}$. We take a process $\{x_1, \dots\}$ where we identify $x_i := (y_i, \dots y_{i+k-1})$. If $y_i \in \{1, 2, \dots, b\}$, $x_i \in \{1, 2, \dots, b\}^k$ and it is easy to reorder the elements of this set in such a way that $x_i \in \{1, 2, \dots, B\}$ where $B = b^k$. If $\{y_1, \dots\}$ is stationary, ergodic and mixing with mixing coefficients $\alpha(m)$ ($\varphi(m)$) for $m \in \mathbb{N}$, then the

process $\{x_1, \dots\}$ is respectively stationary, ergodic and mixing with mixing coefficients $\alpha(m-k+1)$ $(\varphi(m-k+1))$ for $m \in \mathbb{N}$. However, this estimator of the block entropy may suffer from some drawbacks: as the number of cells whose probability is small increases with k, the bias tends to increase and the bias corrections are less reliable.

The proportions of values equal to i is:

$$q_i = \frac{n_i}{N} = \frac{\sum_{j=1}^{N} 1\{x_j = i\}}{N}.$$

The observed entropy is therefore:

$$H_N = -\sum_{i=1}^{B} \frac{n_i}{N} \ln \frac{n_i}{N} = -\sum_{i=1}^{B} q_i \ln q_i.$$

With respect to the case in which the observations are from a sequence of iid random variables, the asymptotic theory is quite different.

We will need the following quantities, characterizing the distribution of the process:

$$p_i = \mathbb{P}\{x_1 = i\} \quad i = 1, \dots, B$$
$$p_{ij}^{(h)} = \mathbb{P}\{x_1 = i, x_{h+1} = j\} \quad i, j = 1, \dots, B,\ h \in \mathbb{N}_0.$$

It is clear that $p_{ii}^{(0)} \equiv p_i$ and that $p_{ij}^{(0)} \equiv 0$ if $i \neq j$. Moreover, we will use the notation $p_i^{(h)} \equiv p_{ii}^{(h)}$. Stationarity allows us to extend $p_{ij}^{(h)}$ to $h \in \mathbb{Z}$, in which case $p_{ij}^{(h)} = p_{ji}^{(-h)}$. We also define the vector of dichotomic variables $\mathbf{x}_j = (1\{x_j = 1\}, 1\{x_j = 2\}, \dots, 1\{x_j = B\})'$.

In the following, we outline two examples that will be used throughout the paper to show and support our main results. The two examples concern a dichotomized first-order autoregressive process and the Gilbert–Shannon–Reed model whose behavior follows a Markov chain.

Example 1 (Dichotomized AR*(1) Process):* Let us consider a process $\{\widetilde{x}_1, \dots\}$ defined by the first-order autoregressive (i.e. AR (1)) equation:

$$\widetilde{x}_i = \alpha \cdot \widetilde{x}_{i-1} + \varepsilon_i \quad i = 2, \dots$$

where $\{\varepsilon_1, \dots\}$ is an iid process of normally distributed random variables with mean 0 and variance $1-\alpha^2$. The initial value has the distribution $\widetilde{x}_1 \sim \mathcal{N}(0,1)$ that guarantees that the process is strictly stationary. A symbolized process requires the choice of a partition of the real line $\{\mathcal{I}_1, \dots, \mathcal{I}_B\}$. Then:

$$p_i^{(h)} = \mathbb{P}\{x_1 = i, x_{h+1} = i\} = \mathbb{P}\{\widetilde{x}_1 \in \mathcal{I}_i, \widetilde{x}_{h+1} \in \mathcal{I}_i\}$$
$$= \mathbb{P}\left\{\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \alpha^h \\ \alpha^h & 1 \end{bmatrix}\right) \in \mathcal{I}_i \times \mathcal{I}_i\right\}.$$

Here we consider only a dichotomized process, i.e. a symbolized process with $B = 2$, $\mathcal{I}_1 = (-\infty, 0)$ and $\mathcal{I}_2 = [0, +\infty)$. This process retains only the signs of the original process, i.e.:

$$x_i = 1 + 1\{\widetilde{x}_i \geq 0\} \quad i \in \mathbb{N}.$$

It is clear that:

$$p_1 = \mathbb{P}\{x_i = 1\} = \mathbb{P}\{\widetilde{x}_i \geq 0\} = 1/2$$
$$p_2 = 1 - p_1 = 1/2.$$

As to the probabilities of couples separated by h time periods, we first derive the expressions for $\widetilde{x}_i$ as a function of $\widetilde{x}_{i-h}$:

$$\widetilde{x}_i = \alpha^h \cdot \widetilde{x}_{i-h} + \sum_{j=0}^{h-1} \alpha^j \cdot \varepsilon_{i-j}$$

or:

$$\widetilde{x}_{h+1} = \alpha^h \cdot \widetilde{x}_1 + \sum_{\ell=1}^{h} \alpha^{h-\ell} \cdot \varepsilon_{\ell+1}.$$

This implies that $\mathrm{Cov}(\widetilde{x}_1, \widetilde{x}_{h+1}) = \alpha^h \cdot \mathbb{V}(\widetilde{x}_1) = (\alpha^h \sigma^2)/(1-\alpha^2)$ and the correlation is α^h. From [114, p. 189], we have:

$$p_{22}^{(h)} = \mathbb{P}\{x_1 = 2, x_{h+1} = 2\} = \mathbb{P}\{\widetilde{x}_1 \geq 0, \widetilde{x}_{h+1} \geq 0\}$$
$$= \mathbb{P}\left\{\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \alpha^h \\ \alpha^h & 1 \end{bmatrix}\right) \in \mathbb{R}_+^2\right\}$$
$$= \pi_h = 1/4 + 1/2\pi \arcsin\left(\alpha^h\right)$$
$$p_{11}^{(h)} = \pi_h$$
$$p_{12}^{(h)} = p_{21}^{(h)} = 1/2 - \pi_h.$$

It is clear that $\pi_h \to 1/4$, $|\pi_h - 1/4| \leq \alpha^h/4$ and, for large h, $\pi_h \sim 1/4 + 1/2\pi\alpha^h$.

Example 2 (Gilbert–Shannon–Reeds (GSR) Model): We consider the Gilbert–Shannon–Reeds model of shuffles (see, e.g., [12]), but in the following we only need a short description. Let B the number of cards in a deck. First, a number C is chosen from $\{0, 1, \dots, B\}$ according to the binomial distribution with probabilities $\binom{B}{C}/(2B)$. Second, the first C cards are held in the left hand and the remaining $B-C$ cards in the right. Third, cards are dropped from a given hand with probability proportional to packet size. Thus, the first card is dropped from the left hand packet with probability C/B and from the right hand packet with probability $(B-C)/B$. If the first card is dropped from the left packet, the next card is dropped from the left packet with probability $(C-1)/(B-1)$ and from the right packet with probability $(B-C)/(B-1)$. The process continues until there is no card left. This describes a Markov chain whose state space is the set of all possible permutations of the deck of cards, but we will not focus on this process. We will instead consider what happens to a single randomly selected card when the deck is repeatedly shuffled. Even if there is no guarantee that aggregating a Markov chain will result in a Markov chain of the same order (see, e.g., [28]), it is easy to convince oneself that what matters for the position of the card after a shuffle is the position of that same card before the shuffle, the positions of the other cards being irrelevant. The transition matrix of this Markov chain is called *position matrix* in [23]. From Lemma 2.1 in [23] or Proposition 2.1 in [9], the probability of going from state i to state j is:

$$\pi_{ij} = \begin{cases} 2^{-j} + 2^{j-1-B} & \text{if } i = j \\ 2^{j-1-B}\binom{B-j}{i-j} & \text{if } i > j \\ 2^{-j}\binom{j-1}{i-1} & \text{if } j > i \end{cases}$$

The position matrices $\mathbf{P} = [\pi_{ij}]$ are therefore given by the following formulas, valid respectively for $B = 2, 3, 4, 5, 6$:

$$\mathbf{P} = 4^{-1} \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{P} = 8^{-1} \begin{bmatrix} 5 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 5 \end{bmatrix}$$

$$\mathbf{P} = 16^{-1} \begin{bmatrix} 9 & 4 & 2 & 1 \\ 3 & 6 & 4 & 3 \\ 3 & 4 & 6 & 3 \\ 1 & 2 & 4 & 9 \end{bmatrix}$$

$$\mathbf{P} = 32^{-1} \begin{bmatrix} 17 & 8 & 4 & 2 & 1 \\ 4 & 10 & 8 & 6 & 4 \\ 6 & 6 & 8 & 6 & 6 \\ 4 & 6 & 8 & 10 & 4 \\ 1 & 2 & 4 & 8 & 17 \end{bmatrix}$$

$$\mathbf{P} = 64^{-1} \begin{bmatrix} 33 & 16 & 8 & 4 & 2 & 1 \\ 5 & 18 & 16 & 12 & 8 & 5 \\ 10 & 8 & 12 & 12 & 12 & 10 \\ 10 & 12 & 12 & 12 & 8 & 10 \\ 5 & 8 & 12 & 16 & 18 & 5 \\ 1 & 2 & 4 & 8 & 16 & 33 \end{bmatrix}.$$

The stationary probability of these Markov chains is a uniform distribution on the states, so that $p_i = B^{-1}$ for $i = 1, \ldots, B$. The probabilities $p_{ij}^{(h)}$ can then be easily obtained multiplying the position matrices and the stationary probability.

A. Limiting Behavior

A trivial application of the Birkhoff Ergodic Theorem and of, e.g., Corollary 6.3.1 in [96, p. 174] yields almost sure convergence of H_N to a limiting random variable measurable with respect to the invariant σ-algebra $\mathcal{I}$.

Proposition 3: Under stationarity:

$$\lim_{N\to\infty} H_N = H_\infty = -\sum_{i=1}^{B} \mathbb{P}\{x_1 = i \,|\mathcal{I}\} \ln \mathbb{P}\{x_1 = i \,|\mathcal{I}\} \quad \mathbb{P}-\text{as}$$

where H_∞ is an invariant random variable. Under stationary ergodicity, $H_\infty = -\sum_{i=1}^{B} p_i \ln p_i$ is a constant.

B. Bias

The nonlinearity introduced by the logarithm implies that H_N will not be an unbiased estimator of H_∞. Moreover, it is well known that the bias is always negative (see, e.g., [87, Proposition 1] and note that the proof does not depend on the iid assumption). The next result characterizes the bias of H_N.

Proposition 4: If the process $\{x_1, \ldots\}$ is stationary ergodic:

$$\mathbb{E}[H_N] - H_\infty = -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right)}{p_i} + o(1)$$

where $-\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right)}{p_i} \leq 0$ and the right-hand side is $o(1)$. If the process $\{x_1, \ldots\}$ is α-mixing with $\sum_{n=1}^{\infty} \alpha(n) < \infty$, then:

$$\mathbb{E}[H_N] - H_\infty = -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i} + o\left(N^{-1}\right)$$

where the right-hand side is indeed $O\left(N^{-1}\right)$.

Remark 5: (i) Positive values of the covariance $\mathrm{Cov}\left(1\{x_j = i\}, 1\{x_\ell = i\}\right) = p_i^{(j-\ell)} - p_i^2$, for $i = 1, \ldots, B$ and $j, \ell \in \mathbb{N}$, for most values of the indices are sometimes used as an indicator of persistence of the stochastic process $\{x_1, \ldots\}$, often defined as the tendency to assume in a time period values that are near to the ones of previous time periods. Persistent stochastic processes will usually have $\sum_{i=1}^{B} \frac{\sum_{j=1}^{N-1}\left(p_i^{(j)} - p_i^2\right)}{p_i N} > 0$. This implies not only that the observed entropy is systematically biased downwards from the true entropy, but that this effect is stronger for the case of stochastic processes with persistence. Antipersistence can instead reduce the bias.

(ii) Under ergodic stationarity, we have:

$$\frac{1}{N-1}\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right) \to 0$$

but this term is not necessarily $O\left(N^{-1}\right)$ and so isn't the bias.

(iii) In the rest of the paper, and especially in Section IV, we will use the following definition, valid under α-mixing with $\sum_{n=1}^{\infty} \alpha(n) < \infty$:

$$\mathrm{bias}(H_N) := -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B} \frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i}.$$

The reason is that most results on which we will rely for the estimation of $\mathrm{bias}(H_N)$ require conditions stronger than $\sum_{n=1}^{\infty} \alpha(n) < \infty$ (see, e.g., [5]).

(iv) It is interesting to see what this result implies for the stationary not necessarily ergodic case. For a stationary ergodic process, the time average and the ensemble average coincide and the bias correction is quite simple to understand and, as we will see below, implement. However, in the general case of a stationary process, the limit of H_N is the time average H_∞, that is an $\mathcal{I}$-measurable random variable, and a bias correction for the time average should be an $\mathcal{I}$-measurable random variable too. Nevertheless, in most applications one observes a single time series. It is well known that a stationary process can be written as a mixing of ergodic processes with respect to a measure that, e.g., is called *contingency law* in [115]. This means that each single realization of a stationary process is obtained, first, extracting a random value from the contingency law and, second, extracting a realization from the ergodic process associated with the previous random value. This implies that each time series is extracted from an ergodic process whose properties can be inferred using the Ergodic Theorem, but nothing can be inferred about the contingency

law. This point of view is made very clear in [74, p. 202] with reference to prediction. As a result, the bias correction applies to the entropy computed on the single trajectory, that can be supposed to be extracted from an ergodic law.

Example 6 (Dichotomized $\mathsf{AR}(1)$ *Process - Example 1 Continued):* The process $\{x_1, \dots\}$ is ergodic and mixing with $\alpha(h) \leq {}^1\!/_{2\pi} \arcsin\left(\alpha^h\right) \leq {}^{\alpha^h}\!/_{2\pi}$. Therefore, we have:

$$\begin{aligned}\mathbb{E}[H_N] =& H_\infty - \frac{1}{2N} \\ & - \frac{2}{\pi}\left(\frac{\sum_{j=1}^{N-1}\left(1-\frac{j}{N}\right)\arcsin\left(\alpha^j\right)}{N}\right) + o\left(N^{-1}\right) \\ =& H_\infty - \frac{1}{2N} - \frac{2}{\pi N}\left(\sum_{j=1}^{\infty}\arcsin\left(\alpha^j\right)\right) + o\left(N^{-1}\right).\end{aligned}$$

In Figure 1 we show the performance of the bias correction. For the moment, as we have not yet considered estimation, we correct H_∞ to approximate $\mathbb{E}[H_N]$.[1] The trajectories of H_N are represented by the grey jigsaw lines that oscillate around the dark grey line representing $\mathbb{E}H_N$. They converge from below towards the fixed limiting value H_∞. In finite samples, H_N is a badly biased estimator of H_∞. We show two corrections to H_∞, i.e. in black dotted line $H_\infty - {}^1\!/_{2N}$, the correction for the iid case (the one proposed in [102] for the time-series case), and in black dashed line $H_\infty - {}^1\!/_{2N} - {}^2\!/_{\pi N}\left(\sum_{j=1}^{\infty}\arcsin\left(\alpha^j\right)\right)$, our correction. It is clear that our correction is much better than the one in [102].[2] On the right plot, we display the empirical cumulative distribution function (cdf) of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black solid line). This shows that in the ergodic case H_N converges (almost surely) to H_∞.

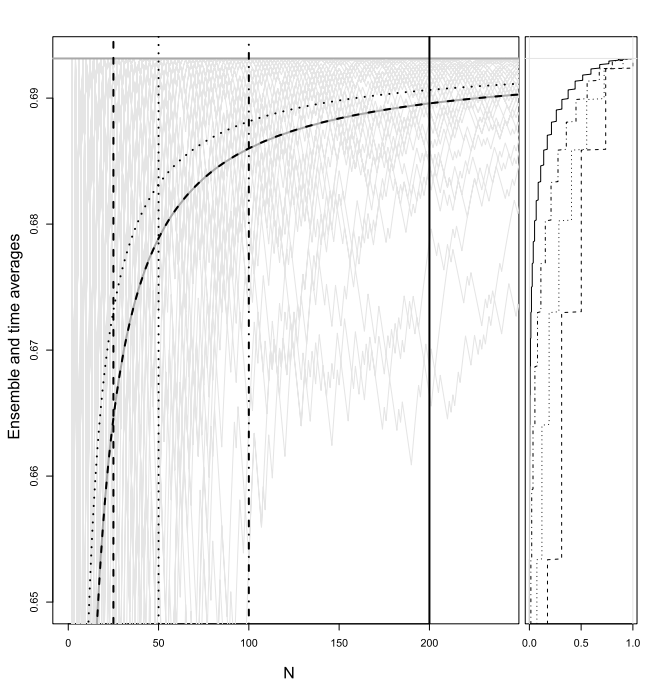

Fig. 1. Ensemble and time averages of the entropy in the ergodic (dichotomized $\mathsf{AR}(1)$) case: on the left plot, 50 trajectories of H_N as a function of N (light grey jigsaw lines), H_∞ (dark grey horizontal line), $\mathbb{E}H_N$ (dark grey curved line), $H_\infty - {}^1\!/_{2N}$ (black dotted line), $H_\infty - {}^1\!/_{2N} - {}^2\!/_{\pi N}\left(\sum_{j=1}^{\infty}\arcsin\left(\alpha^j\right)\right)$ (black dashed line), vertical lines at $N \in \{25, 50, 100, 200\}$ (respectively black dashed, dotted, dash-dot, solid lines); on the right plot, empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black solid line).

C. Central Limit Theorem

The following proposition provides an asymptotic distributional result.

Proposition 7: If the process $\{x_1, \dots\}$ is α-mixing with $\sum_{n=1}^{\infty}\alpha(n) < \infty$, we have:

$$\sqrt{N}\left(H_N - H_\infty\right) \to_{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$$

where

$$\begin{aligned}\sigma^2 :=& \sum_{i=1}^{B} p_i \ln^2 p_i - \left(\sum_{i=1}^{B} p_i \ln p_i\right)^2 \\ & + 2\sum_{i=1}^{B}\sum_{i'=1}^{B}\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right)\ln p_i \ln p_{i'}.\end{aligned}$$

Provided $\sigma^2 \neq 0$ and $\varphi(n) \leq \kappa(n+1)^{-2}$ for any n:

$$\left\| F_{\sqrt{N}(H_N - H_\infty)/\sigma} - \Phi \right\|_\infty = O\left(N^{-1/2}\right).$$

Remark 8: (i) When $p_i = B^{-1}$ for any i, the asymptotic variance annihilates. For the iid case, this was remarked by Miller and Madow in 1954 (see [72, p. 45]) but was later overlooked by [11] (see [47, pp. 326-327]). Here we show that the same result extends to the case in which the observations are dependent. The asymptotic distribution in this case is dealt with in Section III-D.

(ii) The Berry–Esséen bound involves a condition on the uniform mixing coefficients because Berry–Esséen bounds for the strong mixing case are less satisfactory (see [98, Theorem 2, Remark 2, Application 2]).

Example 9 (Dichotomized $\mathsf{AR}(1)$ *process - Examples 1, 6 continued):* In this case we have $\sigma^2 = 0$.

Combining together Propositions 4 and 7, we obtain the following trivial result.

Corollary 10: If the process is α-mixing with $\sum_{n=1}^{\infty}\alpha(n) < \infty$, we have:

$$\sqrt{N}\left(H_N - \operatorname{bias}(H_N) - H_\infty\right) \to_{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right).$$

D. Asymptotic Distribution Under Degeneracy

Now we turn to the properties when $p_i = B^{-1}$ for any i. In the following proposition we need the definition of a matrix $\mathbf{\Omega}$. In Section IV we will show that this is a modification of a covariance matrix $\mathbf{\Sigma}$.

[1] In general, the bias correction is applied to H_N in order to reduce its distance with respect to the value H_∞ that is being estimated (see Example 16).

[2] Our correction to the bias looks even better because in this example the $o\left(N^{-1}\right)$ term in the equation above can be shown to be $O\left(N^{-2}\right)$. This easily derives from the development of Lemma 30: using the fact that the random variables q_i are symmetric and have odd moments equal to 0, the first non-null term after the bias is $-\frac{4}{3}\mathbb{E}(q_1 - {}^1\!/_2)^4 = O\left(N^{-2}\right)$.

Proposition 11: Suppose that the process $\{x_1, \dots\}$ is α-mixing with $\sum_{n=1}^{\infty} \alpha(n) < \infty$. Consider the matrix $\mathbf{\Omega}$, whose elements are given by:

$$\mathbf{\Omega}_{ii} = \frac{2\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right) + p_i(1-p_i)}{2p_i},$$

$$\mathbf{\Omega}_{ii'} = \frac{2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'}}{2\left(p_i p_{i'}\right)^{1/2}}.$$

Let $(\lambda_1, \dots, \lambda_B)$ be the eigenvalues of the matrix $\mathbf{\Omega}$ arranged in decreasing order. Therefore:

$$N(H_N - H_\infty) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2.$$

Remark 12: (i) The derivation of a precise rate of convergence of the finite-sample distribution to the asymptotic one, namely $\left\| F_{N(H_N - H_\infty)} - F_{-\sum_{i=1}^{B}\lambda_i\chi_{1,i}^2} \right\|_\infty$, seems to be out of reach given the state of the literature. According to the proof of Proposition 11, the rate of convergence of $N(H_N - H_\infty)$ to its asymptotic distribution can be linked to the rate of convergence of the chi-square statistic $-N\sum_{i=1}^{B}\frac{(q_i - p_i)^2}{2p_i}$. In the dependent case there seems to be no available result for the lattice case, but one can consider what happens in the independent case as a benchmark. In that case, [27] showed that the convergence rate is $O\left(N^{-\frac{B-1}{B}}\right)$ (see also [16]), while [35] showed that the rate of convergence is $O\left(N^{-1}\right)$ for $B \geq 6$. We investigate the rate of convergence in Examples 13 and 14 below.

Example 13 (Dichotomized AR*(1) Process - Examples 1, 6, 9 Continued):* We have:

$$\mathbf{\Omega}_{11} = \mathbf{\Omega}_{22} = 1/4 + 1/\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right),$$

$$\mathbf{\Omega}_{12} = \mathbf{\Omega}_{21} = -1/4 - 1/\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right).$$

Therefore:

$$\mathbf{\Omega} = \left\{ 1/4 + 1/\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right) \right\} \cdot \begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}.$$

This matrix is singular and therefore $\lambda_1 = \operatorname{tr}(\mathbf{\Omega}) = 1/2 + 2/\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right)$ and $\lambda_2 = 0$. We have:

$$N(H_N - H_\infty) \to_{\mathcal{D}} -\left\{ 1/2 + 2/\pi \sum_{h=1}^{\infty} \arcsin\left(\alpha^h\right) \right\} \cdot \chi_1^2.$$

In Figure 2 we show the difference between the cdf of the entropy and its asymptotic approximation for $N \in \{250, 1000, 4000\}$. The finite-sample distribution of the entropy is discrete as q_1 and q_2 can assume only $N+1$ values. These distribution are obtained through 1,000,000 samplings. The Kolmogorov distances between the finite-sample distributions with $N \in \{250, 1000, 4000\}$ and the asymptotic one are respectively 0.04209815, 0.02110595 and 0.01056157, thus suggesting a rate of convergence of $O\left(N^{-1/2}\right)$, that is in line with Remark 12 for $B = 2$.

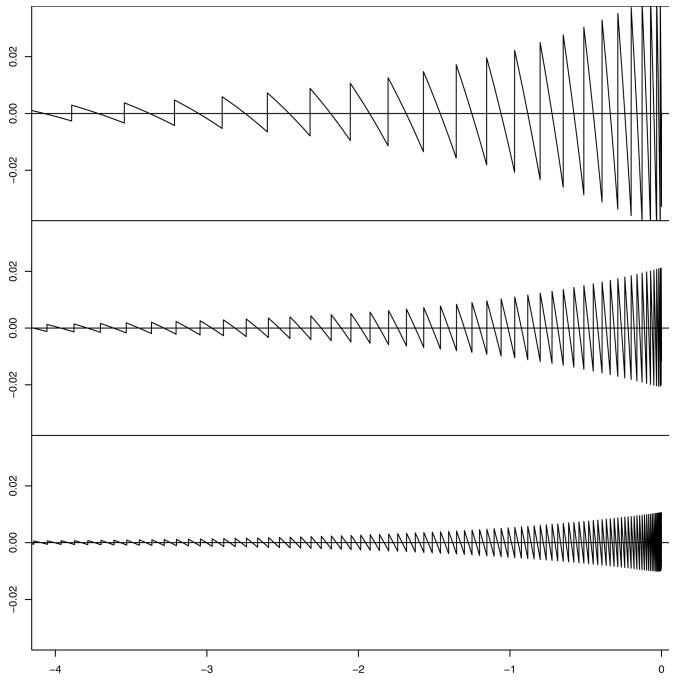

Fig. 2. Difference between the cdf of the entropy and its asymptotic approximation for $N = 250$, $N = 1000$ and $N = 4000$ (from above to below).

Example 14 (GSR Model - Example 2 Continued): For $B \in \{2, 3, 4, 5, 6\}$ we compute the asymptotic distribution and we compare it with the finite-sample distributions for $N \in \{10, 11, \dots, 250\}$. These curves are represented in Figure 3 and are consistent with an increase in the rate of convergence when B increases. The jigsaw profile of the curves in the figure does not seem to be an artifact of our simulations as it appears consistently across different replications. The curves for B running from 2 to 6 are respectively based on over $4 \cdot 10^7$, $4.5 \cdot 10^7$, $4.5 \cdot 10^7$, $2 \cdot 10^8$ and $5 \cdot 10^8$ (non-independent) observations.

The combination of Propositions 4 and 11 gives the following result.

Corollary 15: Suppose that the process $\{x_1, \dots\}$ is α-mixing with $\sum_{n=1}^{\infty} \alpha(n) < \infty$. Consider the matrix $\mathbf{\Omega}$ defined in Proposition 11. Therefore:

$$N(H_N - \operatorname{bias}(H_N) - H_\infty) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \left(\chi_{1,i}^2 - 1\right).$$

IV. Estimation

When correcting for bias or computing the asymptotic variance of the entropy, we need to compute the matrix $\mathbf{\Sigma}$ whose elements are (see Lemma 32 in Section VII-B):

$$\mathbf{\Sigma}_{ii} = p_i(1-p_i) + 2\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right),$$

$$\mathbf{\Sigma}_{ii'} = 2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'}.$$

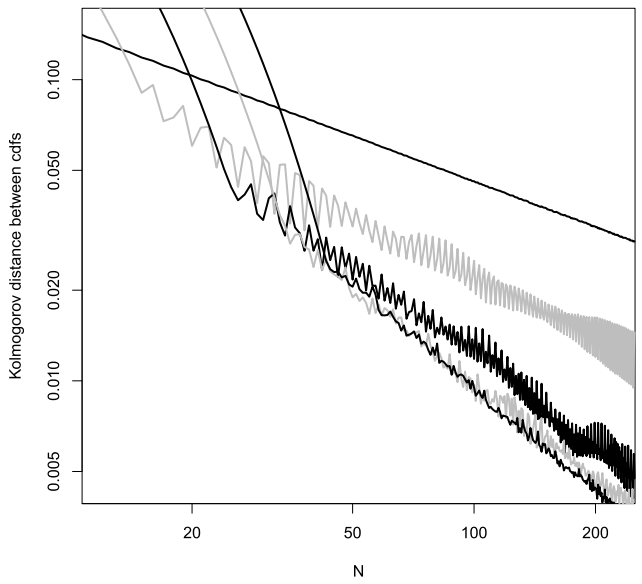

Fig. 3. Rate of convergence to zero of the Kolmogorov distance between the cdf of the entropy and its asymptotic approximation for $N \in \{10, \ldots, 250\}$ and $B \in \{2, \ldots, 6\}$.

We can rewrite the elements as:

$$\boldsymbol{\Sigma}_{ii} = p_i\left(1-p_i\right) + 2\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right) \tag{IV.1}$$

$$= \sum_{h=-\infty}^{\infty}\left(p_i^{(h)} - p_i^2\right) \tag{IV.2}$$

$$\boldsymbol{\Sigma}_{ii'} = 2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'}\right) - p_i p_{i'} \tag{IV.3}$$

$$= \sum_{h=-\infty}^{\infty}\left(p_{ii'}^{(h)} - p_i p_{i'}\right) \tag{IV.4}$$

where we have used the fact that, under stationarity, $p_{i\ell}^{(h)} = p_{\ell i}^{(-h)}$.

The computation of the bias is performed as follows rewriting it as:

$$\operatorname{bias}\left(H_N\right) := -\frac{B-1}{2N} - \frac{1}{N}\sum_{i=1}^{B}\frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i}$$

$$= -\frac{1}{2N}\sum_{i=1}^{B}\frac{\boldsymbol{\Sigma}_{ii}}{p_i} = -\frac{\operatorname{tr}\left(\operatorname{dg}\left(\bar{\mathbf{p}}\right)\boldsymbol{\Sigma}\right)}{2N}. \tag{IV.5}$$

The matrix $\boldsymbol{\Omega}$, used to define the distribution in the degenerate case (see Proposition 11), is defined as:

$$\boldsymbol{\Omega} = \frac{1}{2}\operatorname{dg}\left(\mathbf{p}^{\odot\left(-\frac{1}{2}\right)}\right)\boldsymbol{\Sigma}\operatorname{dg}\left(\mathbf{p}^{\odot\left(-\frac{1}{2}\right)}\right). \tag{IV.6}$$

In the following we propose two methods aimed at correcting the bias of the estimator of the entropy and at computing its variance.

A. First Method

A first method that holds with generality is to estimate the elements of the matrix $\boldsymbol{\Sigma}$ through an autocorrelation-consistent (AC) covariance-matrix estimator, like the ones considered in [5], [80], [93]. We show their general structure.

We define:

$$\boldsymbol{\Pi}^{(h)} = \operatorname{Cov}\left(\mathbf{x}_1, \mathbf{x}_{1+h}'\right)$$

whose generic element is $\left[\boldsymbol{\Pi}^{(h)}\right]_{ii'} = \operatorname{Cov}\left(1\left\{x_1 = i\right\}, 1\left\{x_{1+h} = i'\right\}\right) = p_{ii'}^{(h)} - p_i p_{i'}$. Thus, from (IV.1) and (IV.3):

$$\boldsymbol{\Sigma} = \sum_{h=-\infty}^{\infty}\boldsymbol{\Pi}^{(h)}.$$

Estimators take the form:

$$\hat{\boldsymbol{\Sigma}} = \sum_{h=-N+1}^{N-1} k\left(\frac{h}{S_N}\right)\hat{\boldsymbol{\Pi}}^{(h)}$$

where k is a kernel function, S_N is a bandwidth parameter and:

$$\hat{\boldsymbol{\Pi}}^{(h)} = \begin{cases} \frac{1}{N}\sum_{n=h+1}^{N}\left(\mathbf{x}_n - \mathbf{q}\right)\left(\mathbf{x}_{n-h} - \mathbf{q}\right)' & h \geq 0, \\ \frac{1}{N}\sum_{n=-h+1}^{N}\left(\mathbf{x}_{n+h} - \mathbf{q}\right)\left(\mathbf{x}_n - \mathbf{q}\right)' & h < 0. \end{cases}$$

A plug-in estimator of $\boldsymbol{\Omega}$ is:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{2}\operatorname{dg}\left(\mathbf{q}^{\odot\left(-\frac{1}{2}\right)}\right)\hat{\boldsymbol{\Sigma}}\operatorname{dg}\left(\mathbf{q}^{\odot\left(-\frac{1}{2}\right)}\right).$$

Therefore, a plug-in estimator of the bias is:

$$\widehat{\operatorname{bias}\left(H_N\right)} = -\frac{\operatorname{tr}\left(\operatorname{dg}\left(\bar{\mathbf{q}}\right)\hat{\boldsymbol{\Sigma}}\right)}{2N}.$$

Example 16 (Dichotomized AR (1) *Process - Examples 1, 6, 9, 13 Continued):* Here we consider bias correction using the estimator of [80] and [5]. The light grey jigsaw lines are the trajectories of $H_N - \widehat{\operatorname{bias}\left(H_N\right)}$ with Newey–West bias correction. They oscillate around a curved solid grey line that is $\mathbb{E}\left(H_N - \widehat{\operatorname{bias}\left(H_N\right)}\right)$ with Newey–West bias correction, a curved dashed grey line that is $\mathbb{E}\left(H_N - \widehat{\operatorname{bias}\left(H_N\right)}\right)$ with Andrews bias correction, an horizontal dark grey line that is H_∞, a black solid curve that is $\mathbb{E}\left(H_N\right)$, and a black dashed curve (almost indistinguishable from H_∞) that is $\mathbb{E}\left(H_N\right) - \widehat{\operatorname{bias}\left(H_N\right)}$. It is apparent from the plot that $H_N - \widehat{\operatorname{bias}\left(H_N\right)}$ is in both cases less biased than H_N (see also Figure 1). However, the replacement of $\operatorname{bias}\left(H_N\right)$ with $\widehat{\operatorname{bias}\left(H_N\right)}$ is not without consequences. Indeed, the quantity $\mathbb{E}\left(H_N\right) - \operatorname{bias}\left(H_N\right)$ is represented by a black dashed line that is almost undistinguishable from H_∞, while $\mathbb{E}\left(H_N - \widehat{\operatorname{bias}\left(H_N\right)}\right)$ is not. The vertical lines at $N \in \{100, 125, 150, 200\}$ (respectively black dashed, dotted, dash-dot, solid lines) represent the values of N at which the empirical cdfs of $H_N - \widehat{\operatorname{bias}\left(H_N\right)}$ with Newey–West bias correction (black lines) and with Andrews bias correction (grey lines) are computed.

In the following we will prove our results under the following assumption.

AC Let $q > 0$ be such that:

$$\sum_{h=-\infty}^{\infty}|h|^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} < \infty$$

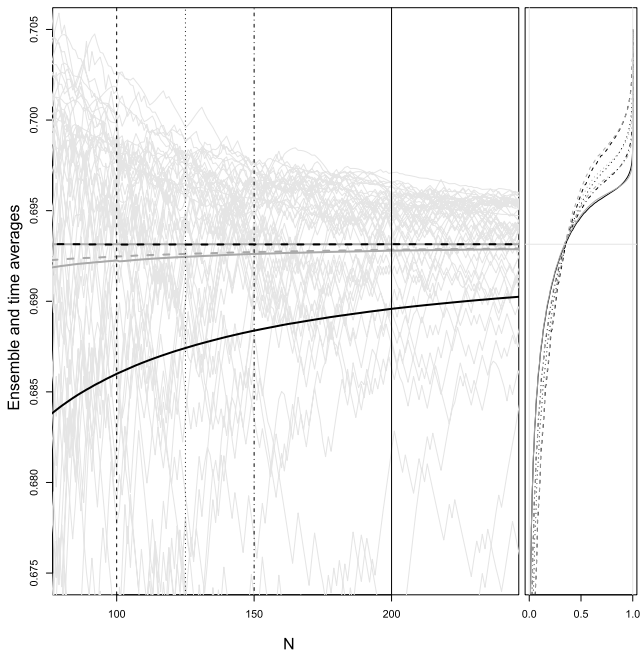

Fig. 4. Ensemble and time averages of the entropy in the ergodic (dichotomized AR (1)) case with bias corrections: on the left plot, 50 trajectories of $H_N - \widehat{\text{bias}}(H_N)$ with Newey–West bias correction as a function of N (light grey jigsaw lines), H_∞ (dark grey horizontal line), $\mathbb{E}(H_N)$ (black solid curve), $\mathbb{E}\left(H_N - \widehat{\text{bias}}(H_N)\right)$ with Newey–West bias correction (grey solid curve), $\mathbb{E}\left(H_N - \widehat{\text{bias}}(H_N)\right)$ with Andrews bias correction (grey dashed curve), $\mathbb{E}(H_N) - \text{bias}(H_N)$ (black dashed curve), vertical lines at $N \in \{100, 125, 150, 200\}$ (respectively black dashed, dotted, dash-dot, solid lines); on the right plot, empirical cdf of $H_N - \widehat{\text{bias}}(H_N)$ with Newey–West bias correction with $N = 100$ (black dashed line), $N = 125$ (black dotted line), $N = 150$ (black dash-dot line), $N = 200$ (black solid line), with Andrews bias correction with $N = 100$ (grey dashed line), $N = 125$ (grey dotted line), $N = 150$ (grey dash-dot line), $N = 200$ (grey solid line).

and:

$$\lim_{x\to 0} \frac{1-k(x)}{|x|^q} = k_q < \infty.$$

Then, the following conditions hold:

1) the process is α-mixing with $\sum_{n=1}^{\infty} n^2 \alpha(n) < \infty$;
2) $S_N/N \to 0$ as $N \to \infty$;
3) $k : \mathbb{R} \to [-1, 1]$ is symmetric, continuous at 0 and for all but a finite number of points, and satisfies $k(0) = 1$ and $\int_{-\infty}^{\infty} k^2(x)\,\mathrm{d}x < \infty$;
4) if $S_N \not\to \infty$, $S_N^{-1} \sum_{j=-N+1}^{N-1} |k(j/S_N)| = O(1)$;
5) if $q < 1/2$, $N^{1/2-q} S_N^{-1/2} = O(1)$;
6) one of the following three sets of conditions hold true:
   a) if $S_N \to \infty$ and $k_q \neq 0$, $S_N^{-q-1/2} N^{1/2} = O(1)$;
   b) if $S_N \to \infty$ and $k_q = 0$:

$$S_N^{-1/2} N^{1/2} \sum_{h=1}^{N-1} \left(1 - k\left(\frac{h}{S_N}\right)\right) \mathbf{\Pi}^{(h)} = O(1)$$

   and this is a fortiori true if $S_N^{-q-1/2} N^{1/2} = O(1)$;
   c) if $S_N \not\to \infty$:

$$S_N^{-1/2} N^{1/2} \sum_{h=1}^{N-1} \left(1 - k\left(\frac{h}{S_N}\right)\right) \mathbf{\Pi}^{(h)} = O(1).$$

Remark 17: (i) The case in which $S_N \not\to \infty$ is required by some recent results on the estimation of AC covariance matrices (see Theorem 2.1 in [93, p. 707]).
(ii) The case in which $S_N \to \infty$ is surely the most interesting. If $k_q \neq 0$, conditions 2 and 6 imply that $N^{\frac{1}{2q+1}} \lesssim S_N \ll N$. In this case condition 4 is always verified and condition 5 is redundant, as $N^{1-2q} \lesssim N^{\frac{1}{2q+1}}$.
(iii) If the function k is non-negative and non-increasing over $[0, \infty)$, one can adapt the reasoning in Theorem 1 in [8, p. 410] to show that assumption 4 is automatically true:

$$\begin{aligned}
&S_N^{-1} \sum_{j=-N+1}^{N-1} |k(j/S_N)| \\
&\quad = S_N^{-1} \left\{1 + 2\sum_{j=1}^{N-1} k(j/S_N)\right\} \\
&\quad \leq S_N^{-1} \left\{1 + 2\int_1^N k(x/S_N)\,\mathrm{d}x + 2k(1/S_N)\right\} \\
&\quad \leq S_N^{-1} + S_N^{-1} \int_{-\infty}^{\infty} k(x/S_N)\,\mathrm{d}x + 2S_N^{-1} k(1/S_N) \\
&\quad \leq \sqrt{\int_{-\infty}^{\infty} k^2(y)\,\mathrm{d}y} + 3S_N^{-1} = O(1).
\end{aligned}$$

This holds irrespective of the fact that $S_N \not\to \infty$ or $S_N \to \infty$.

B. Second Method

Whenever the process is a Markov chain, an alternative is to use the transition matrix in order to compute the probabilities appearing in the formulas above. We suppose below that the Markov chain is ergodic and regular, i.e. irreducible and aperiodic.

We have $\mathbf{p}' = \mathbf{p}'\mathbf{P}$, i.e. $\mathbf{p}$ is a normalized right eigenvector of the stochastic transition matrix $\mathbf{P}$ corresponding to the eigenvalue equal to 1. We define $\mathbf{H} := (\mathbf{I} - \mathbf{P} + \boldsymbol{\iota}\mathbf{p}')^{-1}$, the *fundamental matrix* of [62] (see also [104]).

Proposition 18: For a Markov chain with transition matrix $\mathbf{P}$ and ergodic distribution $\mathbf{p}$, we have:

$$\text{bias}(H_N) = -\frac{2\text{tr}(\mathbf{H}) - B - 1}{2N}$$

and:

$$\mathbf{\Omega} = -\frac{1}{2}\mathbf{I} + \frac{1}{2}\text{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right)\left(\mathbf{H}\text{dg}(\bar{\mathbf{p}}) + \text{dg}(\bar{\mathbf{p}})\mathbf{H}' - \mathbf{U}\right)\text{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right).$$

Example 19 (GSR Model - Examples 2 and 14 Continued): We can characterize the quantities appearing in the GSR model. From Theorem 2.2 in [23], the matrix $\mathbf{P}$ has eigenvalues given by 2^{-m} for $0 \leq m \leq B-1$. Using Theorem 1 in [120], the eigenvalues of $\mathbf{H}^{-1}$ are 1 and $1 - 2^{-m}$ for

$1 \leq m \leq B-1$, and the column eigenvector associated with 1 is proportional to $\boldsymbol{\iota}$. Therefore, the eigenvalues of $\mathbf{H}$ are 1 and $\left(1-2^{-m}\right)^{-1}$ for $1 \leq m \leq B-1$. This implies that $\operatorname{tr}(\mathbf{H}) = B + \sum_{m=1}^{B-1} \frac{1}{2^m-1}$. Now, $\mathbf{p} = B^{-1}\boldsymbol{\iota}$ from which the matrix $\boldsymbol{\Omega}$ in Proposition 18 becomes:

$$\boldsymbol{\Omega} = \frac{1}{2}\left(\mathbf{H} + \mathbf{H}' - \mathbf{I} - B^{-1}\mathbf{U}\right).$$

We suppose to estimate $\mathbf{P}$ through $\hat{\mathbf{P}}$ defined as:

$$\left[\hat{\mathbf{P}}\right]_{ii'} = \frac{\sum_{j=1}^{N-1} \mathbb{1}\left\{x_j = i, x_{j+1} = i'\right\}}{\sum_{i'=1}^{B}\sum_{j=1}^{N-1} \mathbb{1}\left\{x_j = i, x_{j+1} = i'\right\}}$$

and $\mathbf{p}$ through $\hat{\mathbf{p}}$, the normalized left eigenvector of $\hat{\mathbf{P}}$, i.e. $\hat{\mathbf{p}}'\hat{\mathbf{P}} = \hat{\mathbf{p}}'$ (in general $\hat{\mathbf{p}}$ does not coincide with $\mathbf{q}$). Moreover we define $\hat{\mathbf{H}} := \left(\mathbf{I} - \hat{\mathbf{P}} + \boldsymbol{\iota}\hat{\mathbf{p}}'\right)^{-1}$.

C. Error in the Estimation of Bias

We provide a result on the average error induced by the estimation of the bias.

Proposition 20: For the method in Section IV-A, under AC:

$$\widehat{\operatorname{bias}}(H_N) = \operatorname{bias}(H_N) + O_{\mathbb{P}}\left(S_N^{1/2} N^{-3/2}\right).$$

For the method in Section IV-B:

$$\widehat{\operatorname{bias}}(H_N) = \operatorname{bias}(H_N) + O_{\mathbb{P}}\left(N^{-3/2}\right).$$

Remark 21: (i) As $S_N = o(N)$ in AC, for the method in Section IV-A, we get that the error is $o_{\mathbb{P}}\left(N^{-1}\right)$.
(ii) The optimal rate of divergence of S_N for the Newey–West estimator in [80] is $S_N \asymp N^{1/3}$, and for the second-order kernels in [5] it is $S_N \asymp N^{1/5}$. The rate of error in the bias decreases respectively as $O_{\mathbb{P}}\left(N^{-4/3}\right)$ and $O_{\mathbb{P}}\left(N^{-7/5}\right)$.

Corollary 22: Under the conditions of Propositions 7 and 20, if $S_N = o(N)$, we have:

$$\sqrt{N}\left(H_N - \widehat{\operatorname{bias}}(H_N) - H_\infty\right) \to_{\mathcal{D}} \mathcal{N}\left(0, \sigma^2\right)$$

and

$$N\left(H_N - \widehat{\operatorname{bias}}(H_N) - H_\infty\right) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i\left(\chi_{1,i}^2 - 1\right).$$

Using the previous results we can derive the following corollary concerning the MSE.

Corollary 23: Under the hypotheses of Propositions 7, 11 and 20:

$$\operatorname{MSE}(H_N) = \begin{cases} O\left(N^{-1}\right) & \text{if } \sigma^2 > 0 \\ O\left(N^{-2}\right) & \text{if } \sigma^2 = 0 \end{cases}$$

$$\operatorname{MSE}(H_N) - \operatorname{MSE}\left(H_N - \operatorname{bias}(H_N)\right) = O\left(N^{-2}\right)$$

$$\operatorname{MSE}\left(H_N - \widehat{\operatorname{bias}}(H_N)\right) - \operatorname{MSE}\left(H_N - \operatorname{bias}(H_N)\right) = \begin{cases} O\left(S_N^{1/2} N^{-2}\right) & \text{if } \sigma^2 > 0 \\ O\left(S_N^{1/2} N^{-5/2}\right) & \text{if } \sigma^2 = 0 \end{cases}$$

where the left-hand sides of the first two expressions are always non-negative. For the method in Section IV-B, the formulas still hold with $S_N \equiv 1$.

Example 24 (GSR Model - Examples 2, 14 and 19 Continued): We consider a deck of $B = 10$ cards and we shuffle it $N = 1,000$ times. We record the position taken by the card occupying the first position in the original order of the deck. It is expected that, in a series of shuffles, the card will visit each integer number between 1 and B with a probability converging to B^{-1}. Therefore, the limit value of the entropy is $H_\infty = \ln B = \ln 10 \doteq 2.302585$. We have simulated 1,000,000 times the process of shuffling. The average $\mathbb{E}H_N$ without bias correction is 2.296466. We have then computed the Newey–West, Andrews and Markov bias corrections using each time series of N observations. One should note that the matrix $\hat{\mathbf{P}}$ for the Markov bias correction is estimated using only $N-1$ observations. The values of $\mathbb{E}\left(H_N - \widehat{\operatorname{bias}}(H_N)\right)$ with Newey–West, Andrews and Markov bias corrections are respectively 2.302135, 2.302413 and 2.302571, thus confirming the order suggested by Proposition 20. The empirical cdfs of H_N and $H_N - \widehat{\operatorname{bias}}(H_N)$ with Newey–West, Andrews and Markov bias corrections confirm these findings. Note that the Markov bias correction is more precise than the other two. It is also possible to estimate the variance $\mathbb{V}(H_N)$ as $1.121541 \cdot 10^{-5}$, as well as the variances $\mathbb{V}\left(H_N - \widehat{\operatorname{bias}}(H_N)\right)$ with Newey–West, Andrews and Markov bias corrections respectively as $1.143687 \cdot 10^{-5}$, $1.128285 \cdot 10^{-5}$ and $1.126361 \cdot 10^{-5}$. This means that the bias correction slightly increases the variance of the estimator, but the MSE is still improved by the corrections; indeed, the MSE is respectively $4.865827 \cdot 10^{-5}$, $1.163983 \cdot 10^{-5}$, $1.131258 \cdot 10^{-5}$ and $1.126382 \cdot 10^{-5}$ for H_N and $H_N - \widehat{\operatorname{bias}}(H_N)$ with Newey–West, Andrews and Markov bias corrections.

D. Error in the Estimation of the Distribution Under Degeneracy

One of the problems raised by the previous result is to determine what is the effect of estimating the weights on the significance level of tests.

Proposition 25: Let $\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_B\right)$ be the eigenvalues of the matrix $\hat{\boldsymbol{\Omega}}$ defined in Sections IV-A and IV-B. The following bound holds true:

$$\left\| F_{-\sum_{i=1}^{B} \hat{\lambda}_i \chi_{1,i}^2} - F_{-\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2} \right\|_\infty = O\left(\left\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\right\|_1\right).$$

For the method in Section IV-A, under AC the bound is $O_{\mathbb{P}}\left(\left(S_N/N\right)^{1/2}\right)$. For the method in Section IV-B, the bound is $O_{\mathbb{P}}\left(N^{-1/2}\right)$.

Remark 26: (i) This result can be used as follows. Suppose that we determine the quantile q_α of a test of level α using $-\sum_{i=1}^{B} \hat{\lambda}_i \chi_{1,i}^2$. Then:

$$F_{-\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2}(q_\alpha) = \alpha + O\left(\left\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\right\|_1\right).$$

(ii) The distance $\left\| F_{-\sum_{i=1}^{B} \hat{\lambda}_i \chi_{1,i}^2} - F_{-\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2} \right\|_\infty$ is $O_{\mathbb{P}}\left(N^{-1/3}\right)$ for the Newey–West estimator in [80], $O_{\mathbb{P}}\left(N^{-2/5}\right)$ for the second-order kernels in [5], and $O_{\mathbb{P}}\left(N^{-1/2}\right)$ for the flat-top kernels in [93].

Example 27 (GSR Model - Examples 2, 14, 19 and 24 Continued): We compute the asymptotic distribution

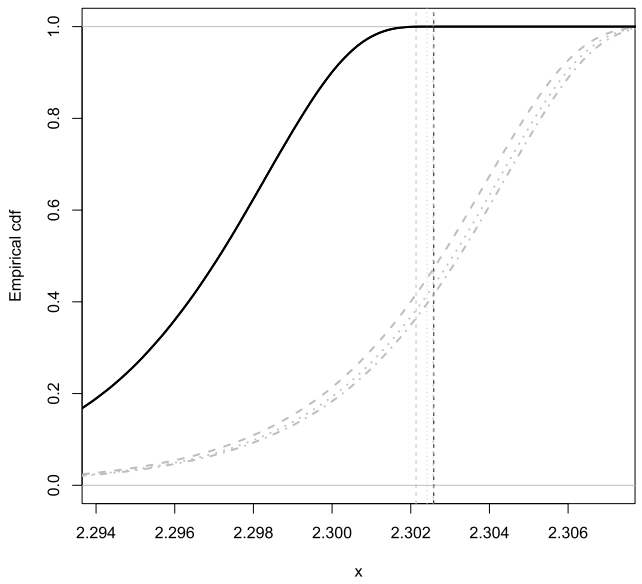

Fig. 5. Empirical cdf of H_N (black line), $H_N - \widehat{\text{bias}}(H_N)$ with Newey–West bias correction (grey dashed line), $H_N - \widehat{\text{bias}}(H_N)$ with Andrews bias correction (grey dotted line), $H_N - \widehat{\text{bias}}(H_N)$ with Markov bias correction (grey dash-dot line), in comparison with the expected values $\mathbb{E}H_N$ (vertical black solid line), $\mathbb{E}\left(H_N - \widehat{\text{bias}}(H_N)\right)$ with Newey–West (vertical grey dashed line), Andrews (vertical grey dotted line) and Markov (vertical grey dash-dot line) bias corrections, and the true value H_∞ (vertical black dashed line).

TABLE I

AVERAGE KOLMOGOROV DISTANCE BETWEEN THE EXACT ASYMPTOTIC DISTRIBUTION AND THE ONE OBTAINED ESTIMATING $\hat{\lambda}_i$, FOR $i = 1, \ldots, B$ THROUGH N OBSERVATIONS

		B				
		2	3	4	5	6
N	125	0.04074774	0.04793574	0.05325378	0.05832340	0.06242307
	250	0.02851654	0.03318252	0.03648672	0.03949900	0.04239093
	500	0.02004530	0.02348522	0.02570183	0.02735191	0.02925993
	1000	0.01416706	0.01646886	0.01806233	0.01931552	0.02067011
	2000	0.009979191	0.011574740	0.012672534	0.013527968	0.014549110
	4000	0.007065414	0.008097934	0.008901423	0.009557864	0.010190659

$F_{-\sum_{i=1}^{B}\lambda_i\chi^2_{1,i}}$ and we compare it with the distributions using matrices $\hat{\mathbf{\Sigma}}$ based on time series of length N for several values B, as described in Example 24. In Table I we compute the quantity:

$$\mathbb{E}\left\|F_{-\sum_{i=1}^{B}\hat{\lambda}_i\chi^2_{1,i}} - F_{-\sum_{i=1}^{B}\lambda_i\chi^2_{1,i}}\right\|_\infty$$

where the expectation is computed over the distribution of the weights based on a series of length N. It is apparent that when the number of observations N is multiplied by 4 there is a division by 2 of the average Kolmogorov distance, coherently with the $O_{\mathbb{P}}\left(N^{-1/2}\right)$ rate predicted by Proposition 25.

V. A GOODNESS-OF-FIT TEST

In this section, we propose a test of goodness-of-fit based on the entropy.

We first describe the setup, then we give two different interpretations of the test procedure. Suppose to observe a stationary time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ with $\widetilde{x}_1$ taking its values in $\mathbb{R}$. Suppose that the process is α-mixing with $\sum_{n=1}^{\infty}\alpha(n) < \infty$. Its marginal distribution has a density f with respect to a measure σ. We want to test the null hypothesis $\mathsf{H}_0 : f \equiv f_0$, where f_0 is a completely specified density function with respect to σ. We identify a partition of the real line $\{\mathcal{I}_1, \ldots, \mathcal{I}_B\}$ such that:

$$\int_{\mathcal{I}_b} f_0(x)\,\sigma(\mathrm{d}x) = B^{-1}, \quad b = 1, \ldots, B.$$

We introduce the symbolized time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ defined by:

$$x_i = \sum_{b=1}^{B} b \cdot 1\{\widetilde{x}_i \in \mathcal{I}_b\}.$$

The symbolization is clearly much simpler when the measure σ is the Lebesgue measure and the density with respect to σ is a classical probability density function. Note that, by the very definition of α-mixing, the mixing coefficients of the symbolized process are majorized by the ones of the original process.

The first justification of the test uses the different behavior of the entropy under the null and the alternative hypotheses. Under the null hypothesis, the entropy computed on the time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ is degenerate as in Proposition 11. The asymptotic distribution of the entropy based on $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ satisfies:

$$N(H_N - H_\infty) \to_{\mathcal{D}} -\sum_{i=1}^{B}\lambda_i\chi^2_{1,i}$$

where $H_\infty = \ln B$. Therefore, under H_0, an acceptance region $\mathcal{A} = [q_\alpha, 0]$ corresponding to a significance level α for $N(H_N - \ln B)$ can be built using the quantile q_α of $-\sum_{i=1}^{B}\lambda_i\chi^2_{1,i}$ such that $F_{-\sum_{i=1}^{B}\lambda_i\chi^2_{1,i}}(q_\alpha) = \alpha$.

Now, we investigate what happens under the alternative hypothesis $\mathsf{H}_1 : f \equiv f_1 \neq f_0$. The test appears to have no power against any f_1 such that:

$$\int_{\mathcal{I}_b} f_1(x)\,\sigma(\mathrm{d}x) = B^{-1}, \quad b = 1, \ldots, B. \tag{V.1}$$

If, however, this does not hold true, Proposition 7 implies that σ is strictly positive and that:

$$\begin{aligned}
&\mathbb{P}\{N(H_N - \ln B) \in \mathcal{A}\} \\
&\quad = \mathbb{P}\left\{H_N \geq \ln B + \frac{q_\alpha}{N}\right\} \\
&\quad = \mathbb{P}\left\{\sqrt{N}\frac{H_N - H_\infty}{\sigma} \geq \sqrt{N}\frac{\ln B - H_\infty}{\sigma} + \frac{q_\alpha}{\sqrt{N}\sigma}\right\} \\
&\quad \leq \left\|F_{\sqrt{N}\frac{H_N - H_\infty}{\sigma}} - \Phi\right\|_\infty \\
&\qquad + \Phi\left(-\sqrt{N}\frac{\ln B - H_\infty}{\sigma} - \frac{q_\alpha}{\sqrt{N}\sigma}\right) \\
&\quad = O\left(N^{-1/2}\right) + \Phi\left(\sqrt{N}\frac{H_\infty - \ln B}{\sigma} - \frac{q_\alpha}{\sqrt{N}\sigma}\right).
\end{aligned}$$

Now, $H_\infty \leq \ln B$ with equality if and only if $\int_{\mathcal{I}_b} f_1(x)\,\sigma(\mathrm{d}x) = B^{-1}$ for $b = 1, \ldots, B$ (see, e.g., [72, p. 27]). Therefore, $\Phi\left(\sqrt{N}\frac{H_\infty - \ln B}{\sigma} - \frac{q_\alpha}{\sqrt{N}\sigma}\right) \downarrow 0$, $\mathbb{P}\{N(H_N - \ln B) \in \mathcal{A}\} \downarrow 0$ and the power of the test converges to 1.

Now we come to the second justification. We build the likelihood of the symbolized time series $\{\widetilde{x}_1, \ldots, \widetilde{x}_N\}$ neglecting the dependence between the values, i.e. supposing that they are independent. This object is sometimes called a pseudolikelihood (see, e.g., [21, Section 2.5] for a general result and [36], [100] for earlier examples). Despite the data are dependent, it is still possible to formulate a LR test of $\mathsf{H}_0 : f \equiv f_0$ that takes the form (see, e.g., [117, p. 252]):

$$\mathsf{LR} = \sum_{i=1}^{B} q_i \ln \frac{q_i}{B^{-1}} = \sum_{i=1}^{B} q_i \ln q_i + \ln B = H_\infty - H_N.$$

In the context of [79, Section 9], this is called a *distance metric statistic*. This test will not have the usual asymptotic distribution of LR tests but its distribution can be obtained from the one of the entropy. Linking this goodness-of-fit test with a LR test also shows that the test enjoys some optimality properties in the case of independent data and outperforms commonly used tests such as the chi-square test (see, e.g., [117, Section 17.6]).

In the following we provide two examples showing the finite-sample properties of the test.

Example 28 (Iid Process): We consider the previous procedure when applied to an iid standard Gaussian sample. We symbolize the process in $B = 4$ equally probable intervals. In Figure 6, we depict the deviation between the actual and the nominal significance level:

$$\begin{aligned}\alpha \mapsto &\mathbb{P}\left\{N\left(H_N - \ln B\right) \notin \mathcal{A}\right\} - \mathbb{P}\left\{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i} \notin \mathcal{A}\right\} \\ = &\mathbb{P}\left\{N\left(H_N - \ln B\right) < q_\alpha\right\} - \mathbb{P}\left\{-\sum_{i=1}^{B} \lambda_i \chi^2_{1,i} < q_\alpha\right\} \\ = &\mathbb{P}\left\{N\left(H_N - \ln B\right) < q_\alpha\right\} - \alpha\end{aligned}$$

under the null hypothesis, for $B = 4$, $N \in \{50, 100, 200, 400\}$ and α ranging from 0.01 to 0.1. As the curves are based on $5 \cdot 10^7$ replications, for both plots the irregular profile of the curves is not an artifact of the simulations. In Figure 7 we depict the statistical power function:

$$\alpha \mapsto \pi = \mathbb{P}\left\{N\left(H_N - \ln B\right) \notin \mathcal{A}\right\} = \mathbb{P}\left\{N\left(H_N - \ln B\right) < q_\alpha\right\}$$

under some alternative hypotheses, i.e. when the data are from a sample of iid Gaussian random variables with mean $c \in \{0.1, 0.2, 0.3\}$ and variance 1. These curves are based on 10^7 replications.

Example 29 (Symbolized AR*(1) Process):* We consider the process described in Example 1. We want to test that its marginal distribution is standard Gaussian. In order to do so, we symbolize the process as explained above. We apply the Newey–West variance estimator with bandwidth equal to $S_N = \lceil N^{1/3} \rceil$ and the Andrews quadratic spectral variance estimator with bandwidth equal to $S_N = \lceil N^{1/5} \rceil$. The second choice is advocated in [31, pp. 551, 573] and criticized in [4, p. 17]. The first choice is rather similar to other ones proposed in the literature, such as the commonly used $S_N = \lceil 0.75 \cdot N^{1/3} \rceil$, but we have chosen the present one for simplicity. We have considered the adaptive procedures of [5] and [81], but in a small percentage of cases they fail to deliver reliable results, and this can be a problem in large simulations. Figure 8 represents the deviation between the actual and the nominal significance level for $B = 4$, $N \in \{50, 100, 200, 400\}$, α ranging from 0.01 to 0.1 with Newey–West (on the left) and Andrews (on the right) estimators. The curves look smoother than the ones for the independent case

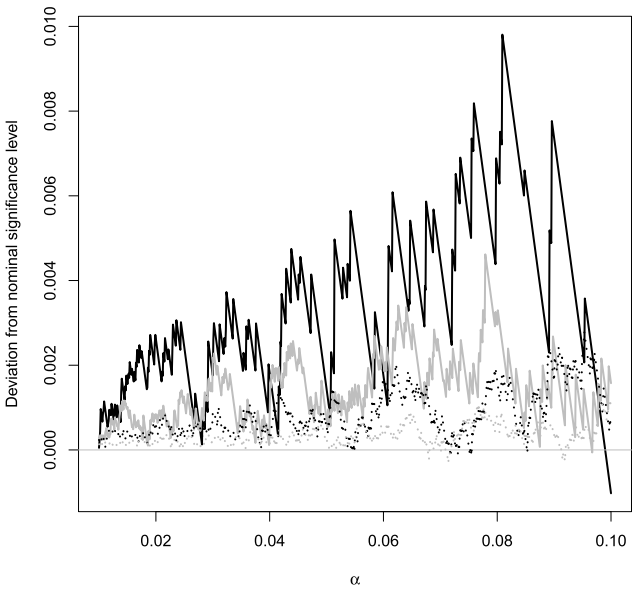

Fig. 6. Difference between the actual and the nominal significance level in the independent case, for $\alpha \in [0.01, 0.1]$, $B = 4$ and $N \in \{50, 100, 200, 400\}$ (continuous black line for $N = 50$, continuous grey line for $N = 100$, dotted black line for $N = 200$, dotted grey line for $N = 400$).

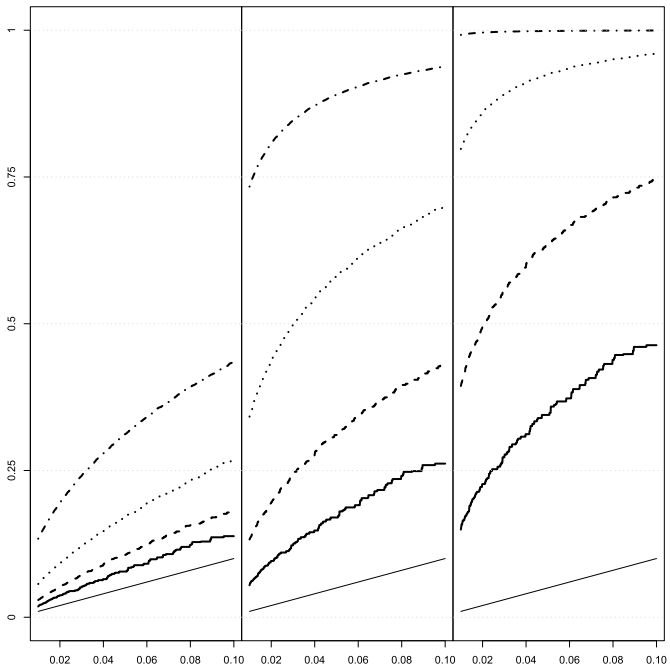

Fig. 7. Power function in the independent case, for $\alpha \in [0.01, 0.1]$, $B = 4$, $N \in \{50, 100, 200, 400\}$, $c \in \{0.1, 0.2, 0.3\}$ (thin line for power equal to α, continuous line for $N = 50$, dashed line for $N = 100$, dotted line for $N = 200$, dash-dot line for $N = 400$; left column for $c = 0.1$, central column for $c = 0.2$, right column for $c = 0.3$).

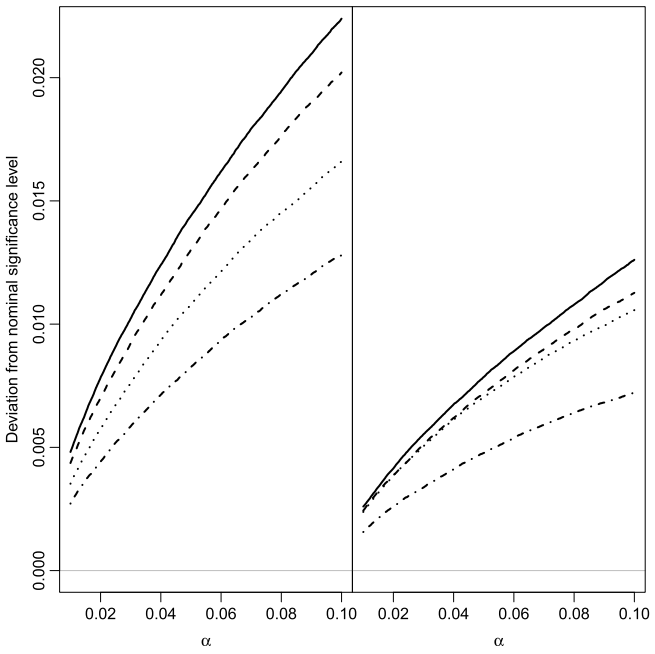

Fig. 8. Difference between the actual and the nominal significance level in the dependent case with Newey–West estimator (on the left) and Andrews estimator (on the right), for $\alpha \in [0.01, 0.1]$, $B = 4$ and $N \in \{50, 100, 200, 400\}$ (continuous line for $N = 50$, dashed line for $N = 100$, dotted line for $N = 200$, dash-dot line for $N = 400$).

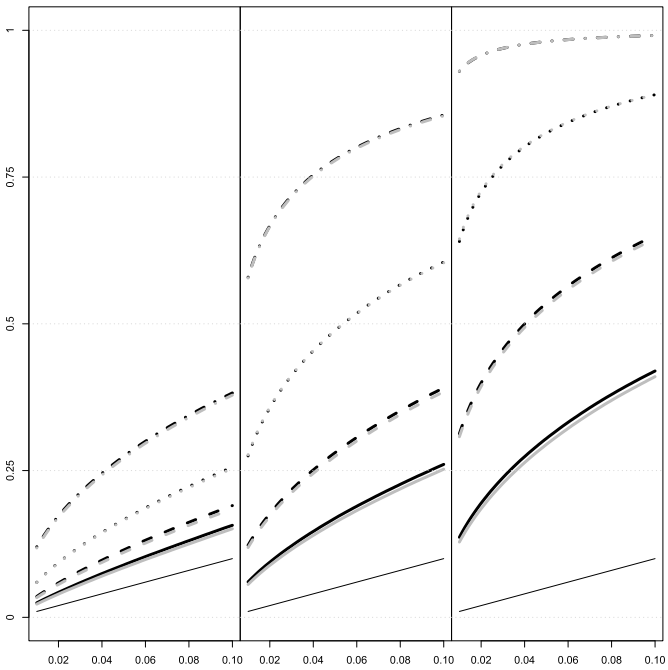

Fig. 9. Power function in the dependent case with Newey–West (in black) and Andrews (in grey) estimators, for $\alpha \in [0.01, 0.1]$, $B = 4$, $N \in \{50, 100, 200, 400\}$, $c \in \{0.1, 0.2, 0.3\}$ (thin line for power equal to α, continuous line for $N = 50$, dashed line for $N = 100$, dotted line for $N = 200$, dash-dot line for $N = 400$; left column for $c = 0.1$, central column for $c = 0.2$, right column for $c = 0.3$).

because the weights of the distribution are computed on the basis of the data and differ for each replication. Even for $N = 50$, the error in the significance level is rather small. Figure 9 represents the statistical power functions for Newey–West (in black) and Andrews (in grey) estimators. In this case the plots are more similar to the ones for the independent case.

VI. Conclusion

In this contribution we consider the estimation of the entropy of data coming from a discretely-supported stochastic process. In order to do so, we use the plug-in estimator of the entropy, in which the probabilities of the different values are replaced by their empirical estimators. With respect to the state of the art, we provide new results concerning asymptotic normality and bias, and we fix an error about formulas for bias correction and variance that, started in [102], has propagated through the literature. We demonstrate that our correction of the bias removes the $O\left(N^{-1}\right)$ part of the bias of the observed entropy H_N. One of the central outcomes of the paper is represented by the behavior of the distribution under degeneracy, i.e. when the marginal distribution of the process assume equal probabilities for each value of the time series. Indeed, at odds with the general case, under degeneracy the statistic–with a different scaling–converges in distribution to a weighted sum of chi-square random variables. We finally introduce some estimators of the distribution under degeneracy and we provide results on the error in the estimation. To complete our analysis, we showcase an application of the entropy to a goodness-of-fit test for the marginal distribution of the process. The simulation studies performed throughout the paper to investigate the finite-sample properties of the estimators enhance the theoretical conclusions.

The present study has some limitations. First of all, we only consider strictly stationary stochastic processes under ergodicity and, for some results, mixing. However, some processes used in signal processing and information theory exhibit limited amounts of nonstationarity such as cyclostationarity (see [33]). Further generalizations of the properties could be obtained through asymptotic mean stationarity, a more general concept (see, e.g., [49]) than stationarity, encompassing cyclostationarity. Second, we consider the properties of the plug-in estimator of the entropy in the discrete or discretized case. This has the consequence, among other things, that the goodness-of-fit test that we propose has no power against some alternative hypotheses (see (V.1)). It could be possible to circumvent this problem by letting the number of classes B to diverge together with the number of observations N.

VII. Proofs

A. Preliminary Results

Here we collect two results for future reference.

The first result (Lemma 30) is a general expansion of H_N already introduced in [47]. To keep the paper self-contained, we reproduce the proof here. The second result (Lemma 31) is a multivariate second-order delta method that will be used in the proof of the asymptotic distribution under degeneracy.

Lemma 30: We have:

$$\begin{aligned}H_N =&H_\infty - \sum_{i=1}^{B}(q_i-p_i)\ln p_i \\ &+\sum_{m=2}^{r}\frac{(-1)^{m-1}}{m(m-1)}\sum_{i=1}^{B}\frac{(q_i-p_i)^m}{p_i^{m-1}}+R_{r+1}\end{aligned}$$

where $|R_{r+1}|\leq \frac{1}{r(r+1)}\sum_{i=1}^{B}\frac{|q_i-p_i|^{r+1}}{(\lambda^\star p_i)^r}$ for $\lambda^\star>0$ independent of any q_i.

Proof: We take a limited development of $-q_i\ln q_i$ for q_i around p_i with Lagrange remainder. We have:

$$\begin{aligned}-q_i\ln q_i =&-p_i\left(1+\frac{q_i-p_i}{p_i}\right)\ln\left[p_i\left(1+\frac{q_i-p_i}{p_i}\right)\right]\\ =&-p_i\ln p_i-(q_i-p_i)\ln p_i\\ &+\sum_{m=2}^{r}\frac{(-1)^{m-1}}{m(m-1)}\frac{(q_i-p_i)^m}{p_i^{m-1}}+R_{r+1,i}\end{aligned}$$

where:

$$\begin{aligned}R_{r+1,i}=&\frac{(-1)^r}{r(r+1)}\frac{(q_i-p_i)^{r+1}}{\xi_i^r},\\ \xi_i=&\lambda_i p_i+(1-\lambda_i)q_i,\quad 0<\lambda_i<1.\end{aligned}$$

This implies that:

$$\begin{aligned}H_N =&H_\infty - \sum_{i=1}^{B}(q_i-p_i)\ln p_i \\ &+\sum_{m=2}^{r}\frac{(-1)^{m-1}}{m(m-1)}\sum_{i=1}^{B}\frac{(q_i-p_i)^m}{p_i^{m-1}}+R_{r+1}\end{aligned}$$

where $R_{r+1}=\sum_{i=1}^{B}R_{r+1,i}$. Now we majorize R_{r+1}.

Let us first suppose that $|q_i-p_i|<(1-\varepsilon)p_i$ for $0<\varepsilon<1$. This implies that $\left|\frac{q_i-p_i}{p_i}\right|<1-\varepsilon$ and $-q_i\ln q_i$ can be expanded in an infinite series:

$$\begin{aligned}-q_i\ln q_i =&-p_i\ln p_i-(q_i-p_i)\ln p_i\\ &+\sum_{m=2}^{r}\frac{(-1)^{m-1}}{m(m-1)}\frac{(q_i-p_i)^m}{p_i^{m-1}}\\ &+\sum_{m=r+1}^{\infty}\frac{(-1)^{m-1}}{m(m-1)}\frac{(q_i-p_i)^m}{p_i^{m-1}}\end{aligned}$$

so that $R_{r+1,i}=\sum_{m=r+1}^{\infty}\frac{(-1)^{m-1}}{m(m-1)}\frac{(q_i-p_i)^m}{p_i^{m-1}}$. Now:

$$\begin{aligned}|R_{r+1,i}|\leq&\sum_{m=r+1}^{\infty}\frac{1}{m(m-1)}\frac{|q_i-p_i|^m}{p_i^{m-1}}\leq\sum_{m=r+1}^{\infty}\frac{|q_i-p_i|^m}{p_i^{m-1}}\\ =&\frac{|q_i-p_i|^{r+1}}{p_i^r}\sum_{j=0}^{\infty}\frac{|q_i-p_i|^j}{p_i^j}\leq\frac{|q_i-p_i|^{r+1}}{\varepsilon p_i^r}\end{aligned}$$

where we have used the fact that, as $\left|\frac{q_i-p_i}{p_i}\right|<1-\varepsilon$, $\sum_{j=0}^{\infty}\frac{|q_i-p_i|^j}{p_i^j}=\left(1-\frac{|q_i-p_i|}{p_i}\right)^{-1}\leq\varepsilon^{-1}$.

Now, let us consider the case $|q_i-p_i|\geq(1-\varepsilon)p_i$ for $0<\varepsilon<1$. Let $A_\varepsilon(p_i):=\{q_i\in[0,1]:|q_i-p_i|\geq(1-\varepsilon)p_i\}$.[3] Then, $R_{r+1,i}=\frac{(-1)^r}{r(r+1)}\frac{(q_i-p_i)^{r+1}}{\xi_i^r}$ will not be zero on $A_\varepsilon(p_i)$. We have:

$$|R_{r+1,i}|=\frac{1}{r(r+1)}\frac{|q_i-p_i|^{r+1}}{(\lambda_i p_i+(1-\lambda_i)q_i)^r}$$

or:

$$\lambda_i=\frac{\left(\frac{|q_i-p_i|^{r+1}}{r(r+1)|R_{r+1,i}|}\right)^{\frac{1}{r}}-q_i}{p_i-q_i}.$$

The set $A_\varepsilon(p_i)$ is compact and $q_i\mapsto\lambda_i$ is a continuous positive function, therefore it attains its minimum on that set and the minimum must be positive. We define:

$$\lambda_B^\star:=\min_{1\leq i\leq B}\min_{q_i\in A_\varepsilon(p_i)}\lambda_i(q_i)>0$$

and we note that this is independent of N.

We define $\lambda^\star:=\min\left\{\lambda_B^\star,\varepsilon^{\frac{1}{r}}\right\}>0$ and we note it is independent of N. The final formula is easily obtained.

Lemma 31: Let $\{\mathbf{X}_n\}$ be a sequence of vectors in $\mathbb{R}^k$. Assume that $\tau_n(\mathbf{X}_n-\boldsymbol{\mu})\to_{\mathcal{D}}\mathbf{X}$ where $\boldsymbol{\mu}$ is a constant vector and $\{\tau_n\}$ is a sequence of constants such that $\tau_n\to\infty$. Let $g:\mathbb{R}^k\to\mathbb{R}$ be twice differentiable at $\boldsymbol{\mu}$ with continuous derivatives and suppose that $\left.\frac{\partial g(\mathbf{x})}{\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})\equiv 0$ for $\mathbf{x}$ in a neighborhood of $\boldsymbol{\mu}$. Then:

$$\tau_n^2(g(\mathbf{X}_n)-g(\boldsymbol{\mu}))\to_{\mathcal{D}}\frac{1}{2}\mathbf{X}'\left.\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}\mathbf{X}.$$

Proof: The proof follows the one of [69, Theorem 11.2.14 (i), p. 436]. A limited development gives the following formula:

$$\begin{aligned}g(\mathbf{x})=&g(\boldsymbol{\mu})+\left.\frac{\partial g(\mathbf{x})}{\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})\\ &+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\left.\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})+R(\mathbf{x}-\boldsymbol{\mu})\end{aligned}$$

where $R(\mathbf{y})=o\left(\|\mathbf{y}\|_{L^2}^2\right)$ as $\|\mathbf{y}\|_{L^2}\downarrow 0$. Now, from $\left.\frac{\partial g(\mathbf{x})}{\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}(\mathbf{x}-\boldsymbol{\mu})\equiv 0$:

$$\begin{aligned}\tau_n^2(g(\mathbf{X}_n)-g(\boldsymbol{\mu}))=&\frac{1}{2}\tau_n^2(\mathbf{X}_n-\boldsymbol{\mu})'\left.\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}(\mathbf{X}_n-\boldsymbol{\mu})\\ &+\tau_n^2 R(\mathbf{X}_n-\boldsymbol{\mu}).\end{aligned}$$

By the Continuous Mapping Theorem, the first term on the right-hand side yields:

$$\begin{aligned}&\frac{1}{2}\tau_n^2(\mathbf{X}_n-\boldsymbol{\mu})'\left.\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}(\mathbf{X}_n-\boldsymbol{\mu})\\ &\quad\to_{\mathcal{D}}\frac{1}{2}\mathbf{X}'\left.\frac{\partial^2 g(\mathbf{x})}{\partial\mathbf{x}\partial\mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}\mathbf{X}.\end{aligned}$$

We then show that $\tau_n^2 R(\mathbf{X}_n-\boldsymbol{\mu})=o_{\mathbb{P}}(1)$. We define the function $h(\mathbf{y}):=R(\mathbf{y})/\|\mathbf{y}\|_{L^2}^2$ for $\mathbf{y}\neq\mathbf{0}$ and $h(\mathbf{0}):=0$. This function is continuous at $\mathbf{0}$ and, therefore:

$$\tau_n^2 R(\mathbf{X}_n-\boldsymbol{\mu})=\tau_n^2\|\mathbf{X}_n-\boldsymbol{\mu}\|_{L^2}^2\,h(\mathbf{X}_n-\boldsymbol{\mu})$$

where $\tau_n^2\|\mathbf{X}_n-\boldsymbol{\mu}\|_{L^2}^2=O_{\mathbb{P}}(1)$, by definition, and $h(\mathbf{X}_n-\boldsymbol{\mu})=o_{\mathbb{P}}(1)$, by the fact that $\tau_n(\mathbf{X}_n-\boldsymbol{\mu})\to_{\mathcal{D}}\mathbf{X}$ implies that $\mathbf{X}_n\to_{\mathbb{P}}\boldsymbol{\mu}$ and by the Continuous Mapping Theorem. By Slutsky's theorem, we get the final result.

[3] We amend the notation used by [47], $A_\varepsilon(p_i,q_i)$, as the set does not depend on q_i.

B. Variances

The following lemma contains some formulas for the variances and covariances of $\mathbf{q}$ and a central limit theorem for $\mathbf{q}$.

Lemma 32: Under stationarity:

$$\begin{aligned}
&\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]\\
&\quad=p_i\left(1-p_i\right)+2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(p_i^{(h)}-p_i^2\right),\\
&\operatorname{Cov}\left[\sqrt{N}\left(q_i-p_i\right),\sqrt{N}\left(q_{i'}-p_{i'}\right)\right]\\
&\quad=2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(\frac{p_{ii'}^{(h)}+p_{i'i}^{(h)}}{2}-p_ip_{i'}\right)-p_ip_{i'}.
\end{aligned}$$

Under α-mixing, if $\sum_{n=1}^{\infty}\alpha\left(n\right)<\infty$, $\sqrt{N}\left(\mathbf{q}-\mathbf{p}\right)\rightarrow_{\mathcal{D}}\mathcal{N}\left(\mathbf{0},\mathbf{\Sigma}\right)$ where:

$$\begin{aligned}
\mathbf{\Sigma}_{ii}&=\lim_{N\rightarrow\infty}\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]\\
&=p_i\left(1-p_i\right)+2\sum_{h=1}^{\infty}\left(p_i^{(h)}-p_i^2\right),\\
\mathbf{\Sigma}_{ii'}&=\lim_{N\rightarrow\infty}\operatorname{Cov}\left[\sqrt{N}\left(q_i-p_i\right),\sqrt{N}\left(q_{i'}-p_{i'}\right)\right]\\
&=2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)}+p_{i'i}^{(h)}}{2}-p_ip_{i'}\right)-p_ip_{i'}.
\end{aligned}$$

Proof: In the following, we will frequently use the rewriting:

$$\sum_{k=1}^{N}\sum_{\ell=1}^{N}p_{ij}^{(k-\ell)}=Np_i\cdot 1\left\{i=j\right\}+\sum_{h=1}^{N-1}\left(N-h\right)\left(p_{ij}^{(h)}+p_{ji}^{(h)}\right)\tag{VII.1}$$

where we have used the equality $p_{i\ell}^{(h)}=p_{\ell i}^{(-h)}$, valid under stationarity.

We have:

$$\begin{aligned}
&\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]=N\mathbb{V}\left(q_i\right)\\
&\quad=\frac{1}{N}\mathbb{V}\left(\sum_{j=1}^{N}1\left\{x_j=i\right\}\right)\qquad\text{(VII.2)}\\
&\quad=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\operatorname{Cov}\left(1\left\{x_j=i\right\},1\left\{x_{j'}=i\right\}\right)\\
&\quad=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left\{\mathbb{E}\left(1\left\{x_j=i\right\}1\left\{x_{j'}=i\right\}\right)-p_i^2\right\}\\
&\quad=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left(p_i^{(j-j')}-p_i^2\right)\\
&\quad=p_i\left(1-p_i\right)+2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(p_i^{(h)}-p_i^2\right)\qquad\text{(VII.3)}
\end{aligned}$$

and:

$$\begin{aligned}
&\operatorname{Cov}\left[\sqrt{N}\left(q_i-p_i\right),\sqrt{N}\left(q_{i'}-p_{i'}\right)\right]\\
&\quad=N\operatorname{Cov}\left(q_i,q_{i'}\right)\qquad\text{(VII.4)}\\
&\quad=\frac{1}{N}\operatorname{Cov}\left(\sum_{j=1}^{N}1\left\{x_j=i\right\},\sum_{j'=1}^{N}1\left\{x_{j'}=i'\right\}\right)\\
&\quad=\frac{1}{N}\sum_{j=1}^{N}\sum_{j'=1}^{N}\left(p_{ii'}^{(j-j')}-p_ip_{i'}\right)\\
&\quad=2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(\frac{p_{ii'}^{(h)}+p_{i'i}^{(h)}}{2}-p_ip_{i'}\right)-p_ip_{i'}
\end{aligned}\tag{VII.5}$$

where we have used repeatedly (VII.1) and $2\sum_{h=1}^{N}\left(1-\frac{h}{N}\right)=N-1$.

Now, we can apply Lemma 1.1 in [99, p. 2] to $\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]$. If $\lim_{\ell\rightarrow\infty}p_i\left(1-p_i\right)+2\sum_{h=1}^{\ell}\left(p_i^{(h)}-p_i^2\right)$ exists, then $\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]$ converges to the same limit. Now, it is clear that $\left|p_i^{(h)}-p_i^2\right|\leq\alpha\left(h\right)$ for any i. If the process is α-mixing with $\sum_{n=1}^{\infty}\alpha\left(n\right)<\infty$, we can apply Lemma 1.2 in [99, p. 3]. Then $\lim_{N\rightarrow\infty}\sum_{h=1}^{N-1}\left(p_i^{(h)}-p_i^2\right)$ exists and can be written as $\sum_{h=1}^{\infty}\left(p_i^{(h)}-p_i^2\right)$. Therefore:

$$\lim_{N\rightarrow\infty}\mathbb{V}\left[\sqrt{N}\left(q_i-p_i\right)\right]=p_i\left(1-p_i\right)+2\sum_{h=1}^{\infty}\left(p_i^{(h)}-p_i^2\right).\tag{VII.6}$$

The same reasoning allows us to write:

$$\begin{aligned}
&\lim_{N\rightarrow\infty}\operatorname{Cov}\left[\sqrt{N}\left(q_i-p_i\right),\sqrt{N}\left(q_{i'}-p_{i'}\right)\right]\\
&\quad=2\sum_{h=1}^{\infty}\left(\frac{p_{ii'}^{(h)}+p_{i'i}^{(h)}}{2}-p_ip_{i'}\right)-p_ip_{i'}.
\end{aligned}$$

To prove the asymptotic normality of $\sqrt{N}\left(\mathbf{q}-\mathbf{p}\right)$, it is enough to apply the CLT in Theorem 18.5.4 of [53]. A vector version of the same result is in [26, p. 67].

C. Bias

Proof of Proposition 4: From Lemma 30 in Section VII-A with $r=2$, we have:

$$\mathbb{E}H_N=H_\infty-\frac{1}{2}\sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i-p_i\right)^2}{p_i}+\mathbb{E}R_3$$

where:

$$\mathbb{E}\left|R_3\right|\leq\frac{1}{6}\sum_{i=1}^{B}\frac{\mathbb{E}\left|q_i-p_i\right|^3}{\left(\lambda^\star p_i\right)^2}.$$

Under stationarity:

$$\begin{aligned}
&\sum_{i=1}^{B}\frac{\mathbb{E}\left(q_i-p_i\right)^2}{2p_i}\\
&\quad=\frac{1}{2N}\sum_{i=1}^{B}\left\{\left(1-p_i\right)+\frac{2\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(p_i^{(h)}-p_i^2\right)}{p_i}\right\}\\
&\quad=\frac{B-1}{2N}+\frac{1}{N}\sum_{i=1}^{B}\frac{\sum_{h=1}^{N-1}\left(1-\frac{h}{N}\right)\left(p_i^{(h)}-p_i^2\right)}{p_i}
\end{aligned}$$

$$= \frac{B-1}{2N} + \frac{1}{N}\sum_{i=1}^{B}\frac{\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right)}{p_i} - \frac{1}{N^2}\sum_{i=1}^{B}\frac{\sum_{h=1}^{N-1} h\left(p_i^{(h)} - p_i^2\right)}{p_i}.$$

If the process is ergodic stationary, then:

$$\frac{1}{N-1}\sum_{h=1}^{N-1}\left(p_i^{(h)} - p_i^2\right) \to 0$$

(see, e.g., [25, Theorem 13.13]). Now, we show that $\frac{1}{N^2}\sum_{h=1}^{N-1} h\left(p_i^{(h)} - p_i^2\right) \to 0$. Note that most majorizations do not work here as they would involve taking the absolute value of $p_i^{(h)} - p_i^2$. Let us define $s_n := p_i^{(n)} - p_i^2$ for $n \in \mathbb{N}_0$ and $s_0 := 0$, and define $\{a_n\}$ as the sequence whose partial sums are given by $\{s_n\}$, i.e. $s_n = \sum_{j=0}^{n} a_j$ or $a_n = s_n - s_{n-1}$. Now we define the Cesáro averaging methods (C,α) for $\alpha \geq 0$ (see [108, Section 2.2]). Using Definitions 2.9 and 2.10 and Lemma 2.11 in [108], we are led to consider:

$$\frac{A_n^\alpha}{E_n^\alpha} = \frac{\sum_{k=0}^{n}\binom{n-k+\alpha}{\alpha}a_k}{\binom{n+\alpha}{\alpha}}.$$

If $\lim_{n\to\infty} {}^{A_n^\alpha}/_{E_n^\alpha}$ converges to a limit, we say that $\{a_n\}$ is summable (C,α), where summability (C,k) implies summability $(C,k+1)$ to the same limit (see [127, Vol. I, p. 76]). Now:

$$\begin{aligned}
\frac{A_n^0}{E_n^0} &= \sum_{k=0}^{n} a_k = s_n \\
\frac{A_n^1}{E_n^1} &= \frac{\sum_{k=0}^{n}(n-k+1)a_k}{n+1} = \frac{\sum_{k=0}^{n} s_k}{n+1} \\
\frac{A_n^2}{E_n^2} &= \frac{\sum_{k=0}^{n}(n-k+1)(n-k+2)a_k}{(n+1)(n+2)} \\
&= 2\left\{\frac{\sum_{k=0}^{n} s_k}{n+2} - \frac{\sum_{k=0}^{n} k s_k}{(n+1)(n+2)}\right\}.
\end{aligned}$$

The sequence $\{a_n\}$ is summable $(C,1)$ with limit 0 as $\lim_{n\to\infty} {}^{A_n^1}/_{E_n^1} = 0$. Therefore it is also summable $(C,2)$, i.e.:

$$\lim_{n\to\infty} 2\left\{\frac{\sum_{k=0}^{n} s_k}{n+2} - \frac{\sum_{k=0}^{n} k s_k}{(n+1)(n+2)}\right\} = 0$$

and, as a result, $\lim_{n\to\infty} n^{-2}\sum_{k=0}^{n} k s_k = 0$. In our case, $\frac{1}{N^2}\sum_{h=1}^{N-1} h\left(p_i^{(h)} - p_i^2\right) \to 0$.

As to the remainder term:

$$\mathbb{E}|R_3| \leq \frac{1}{6}\sum_{i=1}^{B}\frac{\mathbb{E}|q_i - p_i|^3}{(\lambda^\star p_i)^3} \leq \frac{1}{6}\sum_{i=1}^{B}\frac{\mathbb{E}|q_i - p_i|^2}{(\lambda^\star p_i)^3} = o(1).$$

This can also be proved in a different way as in [87, Section 4]. Let us start from the formula:

$$H_N = H_\infty - \sum_{i=1}^{B}(q_i - p_i)\ln p_i - D_{KL}(\mathbf{q};\mathbf{p})$$

where $D_{KL}(\mathbf{q};\mathbf{p}) := \sum_{i=1}^{B} q_i \ln \frac{q_i}{p_i}$ is the Kullback–Leibler divergence. Therefore, $\mathbb{E}D_{KL}(\mathbf{q};\mathbf{p}) = H_\infty - \mathbb{E}H_N$ and, as $D_{KL}(\mathbf{q};\mathbf{p}) \geq 0$ (see [34, p. 422]), $\mathbb{E}H_N \leq H_\infty$ and the bias of H_N is always negative. Now, from [34, Theorem 5]:

$$D_{KL}(\mathbf{q};\mathbf{p}) \leq \ln\left(1 + \sum_{i=1}^{B}\frac{(q_i - p_i)^2}{p_i}\right)$$

and, through Jensen inequality:

$$\begin{aligned}
\mathbb{E}D_{KL}(\mathbf{q};\mathbf{p}) &\leq \mathbb{E}\ln\left(1 + \sum_{i=1}^{B}\frac{(q_i - p_i)^2}{p_i}\right) \\
&\leq \ln\left(1 + \sum_{i=1}^{B}\frac{\mathbb{E}(q_i - p_i)^2}{p_i}\right).
\end{aligned}$$

At last:

$$\begin{aligned}
0 \leq H_\infty - \mathbb{E}H_N &\leq \ln\left(1 + \mathbb{E}\sum_{i=1}^{B}\frac{(q_i - p_i)^2}{p_i}\right) \\
&\leq \sum_{i=1}^{B}\frac{\mathbb{E}(q_i - p_i)^2}{p_i}
\end{aligned}$$

and:

$$\left|H_\infty - \mathbb{E}H_N - \frac{1}{2}\sum_{i=1}^{B}\frac{\mathbb{E}(q_i - p_i)^2}{p_i}\right| \leq \frac{1}{2}\sum_{i=1}^{B}\frac{\mathbb{E}(q_i - p_i)^2}{p_i}.$$

Now we turn to the mixing case. We can apply the reasoning leading to (VII.6) in Lemma 32 in Section VII-B to show that

$$\lim_{N\to\infty} N\sum_{i=1}^{B}\frac{\mathbb{E}(q_i - p_i)^2}{2p_i} = \frac{B-1}{2} + \sum_{i=1}^{B}\frac{\sum_{h=1}^{\infty}\left(p_i^{(h)} - p_i^2\right)}{p_i}.$$

As far as $\mathbb{E}|R_3|$ is concerned, we use Theorem 6.3 in [99]:

$$\begin{aligned}
N^3\mathbb{E}|q_i - p_i|^3 &= \mathbb{E}|N(q_i - p_i)|^3 \\
&\leq 2^{11}3\left\{s_N^3 + N\int_0^1 \left[\alpha^{-1}(u) \wedge N\right]^2 Q^3(u)\,\mathrm{d}u\right\}
\end{aligned}$$

where $s_N^2 := \sum_{j=1}^{N}\sum_{\ell=1}^{N}|\mathrm{Cov}(1\{x_j = i\}, 1\{x_\ell = i\})|$ and $Q(\cdot)$ is the quantile function of the random variable $1\{x_j = i\}$. Now:

$$\begin{aligned}
s_N^2 &:= \sum_{j=1}^{N}\sum_{\ell=1}^{N}\left|p_i^{(j-\ell)} - p_i^2\right| \\
&\leq N p_i(1-p_i) + 2N\sum_{h=1}^{N-1}\alpha(h) - 2\sum_{h=1}^{N-1} h\alpha(h) \\
&\leq N p_i(1-p_i) + 2N\sum_{h=1}^{N-1}\alpha(h).
\end{aligned}$$

As $\sum_{h=1}^{\infty}\alpha(h) < \infty$, $s_N^2 = O(N)$. From (C.10) in [99], using the fact that $Q(\cdot) \leq 1$:

$$\begin{aligned}
M_{p,\alpha,N}(Q) &= \int_0^1 \left[\alpha^{-1}(u) \wedge N\right]^{p-1} Q^p(u)\,\mathrm{d}u \\
&\leq \max(1, p-1)\sum_{h=1}^{N-1}(h+1)^{p-2}\alpha(h)
\end{aligned}$$

and $\mathbb{E}|q_i - p_i|^3 \lesssim N^{-\frac{3}{2}} + N^{-2}\sum_{h=1}^{N-1} h\alpha(h)$. From $\sum_{h=1}^{\infty}\alpha(h) < \infty$ we get $\alpha(h) = o\left(h^{-1}\right)$ and $\mathbb{E}|R_3| = o\left(N^{-1}\right)$.

D. Asymptotic Normality and Berry–Esséen Bound

Proof of Proposition 7: Asymptotic normality of H_N follows from asymptotic normality of $\sqrt{N}(\mathbf{q}-\mathbf{p})$ (see Lemma 32 in Section VII-B) and the delta method (see, e.g., Example 6.1 (b) in [109, p. 279]). The quantity $\sqrt{N}(H_N - H_\infty)$ is asymptotically equivalent to:

$$\sum_{i=1}^{B} \frac{\partial H_\infty}{\partial p_i} \cdot \sqrt{N}(q_i - p_i) = -\sum_{i=1}^{B} (1+\ln p_i) \cdot \sqrt{N}(q_i - p_i) = -\sum_{i=1}^{B} \ln p_i \cdot \sqrt{N}(q_i - p_i).$$

This is asymptotically normal with variance:

$$\begin{aligned}
&\sum_{i=1}^{B}\sum_{i'=1}^{B} \mathbf{\Sigma}_{ii'} \ln p_i \ln p_{i'} \\
&\quad = \sum_{i=1}^{B} p_i \ln^2 p_i \\
&\qquad + \sum_{i=1}^{B}\sum_{i'=1}^{B} \left\{ 2\sum_{h=1}^{\infty} \left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'} \right) - p_i p_{i'} \right\} \\
&\qquad \cdot \ln p_i \ln p_{i'} \\
&\quad = \sum_{i=1}^{B} p_i \ln^2 p_i - \left(\sum_{i=1}^{B} p_i \ln p_i \right)^2 \\
&\qquad + 2\sum_{i=1}^{B}\sum_{i'=1}^{B}\sum_{h=1}^{\infty} \left(\frac{p_{ii'}^{(h)} + p_{i'i}^{(h)}}{2} - p_i p_{i'} \right) \ln p_i \ln p_{i'}
\end{aligned}$$

where $\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{ii'}$ are defined in Lemma 32 in Section VII-B, and the first equality uses the fact that $\mathbf{\Sigma}_{ii}$ is identical to p_i plus the expression for $\mathbf{\Sigma}_{ii'}$ in which i' is formally replaced by i.

Now we turn to the Berry–Esséen bound. We will apply Lemma 1.3 in [109, p. 261], i.e. the inequality:

$$\begin{aligned}
\|F_{W+\Delta} - \Phi\|_\infty &\le \|F_W - \Phi\|_\infty + 4\mathbb{E}|W\Delta| + 4\mathbb{E}|\Delta| \\
&\le \|F_W - \Phi\|_\infty + 4\sqrt{\mathbb{E}W^2\mathbb{E}\Delta^2} + 4\sqrt{\mathbb{E}\Delta^2},
\end{aligned}$$

valid for any random variables W and Δ. We identify $W+\Delta$ with $\sqrt{N}(H_N - H_\infty)/\sigma$ and W with $-\sigma^{-1}\sum_{i=1}^{B}\sqrt{N}(q_i - p_i)\cdot \ln p_i$. As far as $\|F_W - \Phi\|_\infty$ is concerned, if $\sum_{j=1}^{\infty} j\varphi(j) < \infty$, it is shown to be $O\left(N^{-1/2}\right)$ in [97, Théorème 1]. Now we turn to the second term, and we remark that $\mathbb{E}W^2 = 1$. Therefore:

$$\|F_{W+\Delta} - \Phi\|_\infty \le O\left(N^{-1/2}\right) + 8\sqrt{\mathbb{E}\Delta^2}.$$

From Lemma 30 in Section VII-A, we have:

$$H_N = H_\infty - \sum_{i=1}^{B} (q_i - p_i)\ln p_i + R_2$$

where $|R_2| \le \frac{1}{2}\sum_{i=1}^{B} \frac{(q_i-p_i)^2}{\lambda^\star p_i}$. Therefore, $\Delta = -\frac{\sqrt{N}}{\sigma} R_2$ and:

$$\Delta^2 \le \frac{N}{4\lambda^{\star,2}\sigma^2} \left(\sum_{i=1}^{B} \frac{(q_i-p_i)^2}{p_i} \right)^2 \le \frac{NB}{4\lambda^{\star,2}\sigma^2} \sum_{i=1}^{B} \frac{(q_i-p_i)^4}{p_i^2}.$$

Using the fact that $\alpha(n) \le \varphi(n)$ and $\varphi(n) \le \kappa(n+1)^{-2}$ for any n, by Remark 6.3 in [99] or Eq. (2.10) in [99, p. 36], $\mathbb{E}(q_i - p_i)^4 = O\left(N^{-2}\right)$ and $\mathbb{E}\Delta^2 = O\left(N^{-1}\right)$. As a consequence the whole bound is $O\left(N^{-1/2}\right)$.

E. Distribution Under Degeneracy

Proof of Proposition 11: A seldom observed fact is that, if $p_i \equiv 1/B$ for any i, the first-order term in Lemma 30 in Section VII-A is:

$$\sum_{i=1}^{B} (q_i - p_i)\ln p_i = -\ln B \cdot \sum_{i=1}^{B} \left(q_i - \frac{1}{B} \right) = -\ln B \cdot \left(\sum_{i=1}^{B} q_i - 1 \right) = 0.$$

In this case the asymptotic distribution is a degenerate normal random variable with null variance. Therefore, we apply Lemma 31 in Section VII-A identifying $k = B$, $\tau_n = \sqrt{N}$, $\mathbf{X}_n = \mathbf{q}$, $\boldsymbol{\mu} = \mathbf{p}$ and $g(\mathbf{x}) = -\sum_{i=1}^{B} x_i \ln x_i$. The convergence $\sqrt{N}(\mathbf{q}-\mathbf{p}) \to_{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ (that is $\tau_n(\mathbf{X}_n - \boldsymbol{\mu}) \to_{\mathcal{D}} \mathbf{X}$) is proved in Lemma 32 in Section VII-B. We need to compute $\frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'}$ that is the diagonal matrix with $\left[\frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'}\right]_{ii} = -\frac{1}{x_i}$, so that $\left.\frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}'}\right|_{\mathbf{x}=\boldsymbol{\mu}}$ is $-\mathrm{dg}(\overline{\mathbf{p}})$. If we write $\mathbf{G}$ for a standard normal vector, we have:

$$N(H_N - H_\infty) \to_{\mathcal{D}} -\frac{1}{2}\mathbf{G}'\mathbf{\Sigma}^{\frac{1}{2}}\mathrm{dg}(\overline{\mathbf{p}})\,\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{G}.$$

It can be shown (see, e.g., [112]) that the asymptotic distribution of $N(H_N - H_\infty)$ is minus a weighted sum of chi-square random variables whose weights are the eigenvalues, arranged in decreasing order, $(\lambda_1, \ldots, \lambda_B)$ of the matrix $\mathbf{\Omega}$ where:

$$\begin{aligned}
\mathbf{\Omega}_{ii} &= \frac{1}{2p_i}\mathbf{\Sigma}_{ii}, \\
\mathbf{\Omega}_{ij} &= \frac{1}{2(p_i p_j)^{1/2}}\mathbf{\Sigma}_{ij},
\end{aligned}$$

where $\mathbf{\Sigma}_{ii}$ and $\mathbf{\Sigma}_{ii'}$ are defined in Lemma 32 in Section VII-B. At last:

$$N(H_N - H_\infty) \to_{\mathcal{D}} -\sum_{i=1}^{B} \lambda_i \chi_{1,i}^2.$$

Proof of Corollary 15: We have:

$$N(H_N - \mathrm{bias}(H_N) - H_\infty) = N(H_N - H_\infty) - N\mathrm{bias}(H_N).$$

From (IV.5) and (IV.6):

$$\mathrm{bias}(H_N) = -\frac{\mathrm{tr}(\mathrm{dg}(\overline{\mathbf{p}})\,\mathbf{\Sigma})}{2N} = -\frac{\mathrm{tr}(\mathbf{\Omega})}{N} = -\frac{\sum_{i=1}^{B}\lambda_i}{N},$$

from which we get the result.

F. Preliminary Results on Markov Chain Estimation

Proof of Proposition 18: We can write:

$$\begin{aligned}
&\left[p_{ii'}^{(h)} - p_i p_{i'} \right] \\
&\quad = \mathrm{dg}(\mathbf{p})\,\mathbf{P}^h - \mathbf{p}\mathbf{p}'
\end{aligned}$$

$$\cdot\left[\sum_{h=1}^{\infty}\left(p_{ii'}^{(h)}-p_ip_{i'}\right)+\sum_{h=1}^{\infty}\left(p_{i'i}^{(h)}-p_ip_{i'}\right)\right.$$
$$\left.+\,p_i1\left\{i=i'\right\}-p_ip_{i'}\right]$$
$$=\sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^h-\mathbf{p}\mathbf{p}'\right)+\sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^h-\mathbf{p}\mathbf{p}'\right)'$$
$$+\,\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'.$$

We then note that $\mathbf{p}\mathbf{p}'=\mathrm{dg}\left(\mathbf{p}\right)\boldsymbol{\iota}\mathbf{p}'$ allows us to write:

$$\sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^h-\mathbf{p}\mathbf{p}'\right)=\mathrm{dg}\left(\mathbf{p}\right)\sum_{h=1}^{\infty}\left(\mathbf{P}^h-\boldsymbol{\iota}\mathbf{p}'\right).$$

It is well known that:

$$\sum_{h=0}^{\infty}\left(\mathbf{P}^h-\boldsymbol{\iota}\mathbf{p}'\right)=\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right)^{-1}=\mathbf{H}$$

from which:

$$\sum_{h=1}^{\infty}\left(\mathbf{P}^h-\boldsymbol{\iota}\mathbf{p}'\right)=\mathbf{H}-\mathbf{I}. \tag{VII.7}$$

Therefore:

$$\begin{aligned}\boldsymbol{\Sigma}&=\sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^h-\mathbf{p}\mathbf{p}'\right)+\sum_{h=1}^{\infty}\left(\mathrm{dg}\left(\mathbf{p}\right)\mathbf{P}^h-\mathbf{p}\mathbf{p}'\right)'\\&\quad+\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'\\&=\mathrm{dg}\left(\mathbf{p}\right)\sum_{h=1}^{\infty}\left(\mathbf{P}^h-\boldsymbol{\iota}\mathbf{p}'\right)+\sum_{h=1}^{\infty}\left(\mathbf{P}^h-\boldsymbol{\iota}\mathbf{p}'\right)'\mathrm{dg}\left(\mathbf{p}\right)\\&\quad+\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'\\&=\mathrm{dg}\left(\mathbf{p}\right)\left(\mathbf{H}-\mathbf{I}\right)+\left(\mathbf{H}'-\mathbf{I}\right)\mathrm{dg}\left(\mathbf{p}\right)+\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'\\&=\mathrm{dg}\left(\mathbf{p}\right)\mathbf{H}+\mathbf{H}'\mathrm{dg}\left(\mathbf{p}\right)-\mathrm{dg}\left(\mathbf{p}\right)-\mathbf{p}\mathbf{p}'.\end{aligned}$$

The bias can be computed as:

$$\begin{aligned}\mathrm{bias}\left(H_N\right)&=-\frac{\mathrm{tr}\left(\mathrm{dg}\left(\bar{\mathbf{p}}\right)\boldsymbol{\Sigma}\right)}{2N}\\&=-\frac{2\mathrm{tr}\mathbf{H}-B-1}{2N}\end{aligned}$$

where we have used the equalities $\mathrm{tr}\left(\mathbf{AB}\right)=\mathrm{tr}\left(\mathbf{BA}\right)$, $\mathrm{dg}\left(\mathbf{a}\right)\mathrm{dg}\left(\bar{\mathbf{a}}\right)=\mathbf{I}$, $\mathrm{tr}\left(\mathbf{A}\right)=\mathrm{tr}\left(\mathbf{A}'\right)$, $\mathbf{p}\mathbf{p}'=\mathbf{p}\boldsymbol{\iota}'\mathrm{dg}\left(\mathbf{p}\right)$ and $\mathrm{tr}\left(\mathbf{p}\boldsymbol{\iota}'\right)=\mathrm{tr}\left(\boldsymbol{\iota}'\mathbf{p}\right)=1$.

The matrix $\boldsymbol{\Omega}$ used to obtain the distribution in the degenerate case is then defined as:

$$\boldsymbol{\Omega}=-\frac{1}{2}\mathbf{I}+\frac{1}{2}\mathrm{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right)\left(\mathbf{H}\mathrm{dg}\left(\bar{\mathbf{p}}\right)+\mathrm{dg}\left(\bar{\mathbf{p}}\right)\mathbf{H}'-\mathbf{U}\right)\mathrm{dg}\left(\mathbf{p}^{\odot\frac{1}{2}}\right).$$

Lemma 33: For the method in Section IV-B:

$$\widehat{\mathrm{bias}\left(H_N\right)}\simeq\mathrm{bias}\left(H_N\right)+O_{\mathbb{P}}\left(N^{-3/2}\right).$$

Proof: Using the matrix differential notation (see [14], [50], [75]), we write $\mathrm{d}\mathbf{P}:=\hat{\mathbf{P}}-\mathbf{P}$, where $\mathrm{d}\mathbf{P}$ is asymptotically negligible of order $O_{\mathbb{P}}\left(N^{-1/2}\right)$. It is easy to see that:

$$\left(\hat{\mathbf{p}}-\mathbf{p}\right)'\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right)=\hat{\mathbf{p}}'\left(\hat{\mathbf{P}}-\mathbf{P}\right).$$

We have:

$$\left(\hat{\mathbf{p}}-\mathbf{p}\right)'\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right)=\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}$$
$$\left(\hat{\mathbf{p}}-\mathbf{p}\right)'=\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\left(\mathbf{I}-\mathbf{P}+\boldsymbol{\iota}\mathbf{p}'\right)^{-1}=\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}$$
$$\hat{\mathbf{p}}'=\mathbf{p}'+\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}$$
$$\hat{\mathbf{p}}=\mathbf{p}+\mathbf{H}'\mathrm{d}\mathbf{P}'\hat{\mathbf{p}}.$$

Replacing the expression for $\hat{\mathbf{p}}$ in the last formula we get:

$$\hat{\mathbf{p}}=\mathbf{p}+\mathbf{H}'\mathrm{d}\mathbf{P}'\hat{\mathbf{p}}\simeq\mathbf{p}+\mathbf{H}'\mathrm{d}\mathbf{P}'\mathbf{p}.$$

From this:

$$\begin{aligned}\hat{\mathbf{H}}&=\left(\mathbf{I}-\hat{\mathbf{P}}+\boldsymbol{\iota}\hat{\mathbf{p}}'\right)^{-1}\\&=\left(\mathbf{I}-\mathbf{P}-\mathrm{d}\mathbf{P}+\boldsymbol{\iota}\left(\mathbf{p}'+\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}\right)\right)^{-1}\\&=\left(\mathbf{H}^{-1}+\left(\boldsymbol{\iota}\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)^{-1}\\&=\mathbf{H}\left(\mathbf{I}+\mathbf{H}\left(\boldsymbol{\iota}\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)^{-1}\\&\simeq\mathbf{H}\left(\mathbf{I}-\mathbf{H}\left(\boldsymbol{\iota}\hat{\mathbf{p}}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)\\&\simeq\mathbf{H}\left(\mathbf{I}-\mathbf{H}\left(\boldsymbol{\iota}\mathbf{p}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right)\end{aligned}$$

Then:

$$\begin{aligned}\widehat{\mathrm{bias}\left(H_N\right)}&=-\frac{2\mathrm{tr}\hat{\mathbf{H}}-B-1}{2N}\\&\simeq-\frac{2\mathrm{tr}\mathbf{H}-B-1}{2N}+\frac{\mathrm{tr}\left[\mathbf{H}^2\left(\boldsymbol{\iota}\mathbf{p}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right]}{N}\\&=\mathrm{bias}\left(H_N\right)+\frac{\mathrm{tr}\left[\mathbf{H}^2\left(\boldsymbol{\iota}\mathbf{p}'\mathrm{d}\mathbf{P}\mathbf{H}-\mathrm{d}\mathbf{P}\right)\right]}{N}.\end{aligned}$$

This implies that $\widehat{\mathrm{bias}\left(H_N\right)}-\mathrm{bias}\left(H_N\right)=O_{\mathbb{P}}\left(N^{-3/2}\right)$.

G. Error in the Estimation of Bias

We first adapt Lemma A.1 in [93, p. 739] to our case. This result is a version of Theorem 10 in [46, p. 283], Lemma 2 in [4, p. 15], and Proposition 1 in [5, p. 825]. It generalizes these results as it allows for a bandwidth not diverging to ∞, as required by some recent results (see Theorem 2.1 in [93, p. 707]). We define:

$$\boldsymbol{F}^{(q)}:=\frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\left|h\right|^q\boldsymbol{\Pi}^{(h)}.$$

Lemma 34: Assume AC. Then, $\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}=O_{\mathbb{P}}\left(S_N^{1/2}N^{-1/2}\right)$.

Proof: We first restate Assumptions A, B and C in [5]. We identify T with N, t with n, θ with $\mathbf{p}$, $\widehat{\theta}$ with $\mathbf{q}$, $V_t\left(\theta\right)$ with $\mathbf{x}_n-\mathbf{p}$, $V_t\left(\widehat{\theta}\right)$ with $\mathbf{x}_n-\mathbf{q}$.

In order to verify Assumption A in [5, p. 823], we use his Lemma 1. Our process $\left\{\mathbf{x}_1-\mathbf{p},\mathbf{x}_2-\mathbf{p},\dots\right\}$ is zero-mean, fourth-order stationary and bounded. Therefore, we can take $\nu=\infty$ and it's enough to require $\sum_{n=1}^{\infty}n^2\alpha\left(n\right)<\infty$, as in our condition 1 of assumption AC.

Now we consider Assumption B in [5, p. 825]. Assumption B (i) is verified by an application of Theorem 18.5.4 of [53] or of [26, p. 67] under $\sum_{n=1}^{\infty}\alpha\left(n\right)<\infty$. Assumption B (ii) is trivially true because of the boundedness of $\mathbf{x}_n$. For Assumption B (iii) it is enough to identify $\left(\partial/\partial\theta'\right)V_t\left(\theta\right)$ with $\left(\partial/\partial\mathbf{p}'\right)\left(\mathbf{x}_n-\mathbf{p}\right)=-\mathbf{I}$. Assumption B (iv) is true under condition 3 of assumption AC.

As far as Assumption C (i) in [5, p. 826] is concerned, this is true by the reasoning reported just after the statement of his Assumption C. Part (ii) of the assumption is true as $\left(\partial^2/\partial\theta\partial\theta'\right)V_{ta}\left(\theta\right)$ is 0.

Now we turn to the requirements stated in [93, p. 739]. Conditions 2, 3 and 4 in our result come from Lemma A.1 in [93, p. 739], respectively, as a statement in the text, as Eq. (A.1), and as condition (i). We just note that condition (i) is not needed explicitly in [5] because, when $S_N \to \infty$, $S_N^{-1}\sum_{j=-N+1}^{N-1}|k(j/S_N)| \to \int |k(x)|\,\mathrm{d}x$ (see [5, p. 852]). This is finite by our condition 3 of Assumption AC. Instead, [93] requires his condition (i) (see [93, p. 744]) because S_N may not diverge. As we allow S_N to be bounded, we require it but change its statement. Condition (iii) of Lemma A.1 in [93, p. 739] is automatically verified as $\mathbb{E}V_t(\theta)(\partial/\partial\theta')V_{t-j}(\theta) = -\mathbb{E}V_t(\theta)$ and this is a zero vector. Condition (ii) of Lemma A.1 in [93, p. 739] is trickier. We first note that the matrix denoted $\hat{\Omega}$ in that source corresponds to the matrix $\tilde{\boldsymbol{\Sigma}}$ defined as:

$$\tilde{\boldsymbol{\Sigma}} = \sum_{h=-N+1}^{N-1} k\left(\frac{h}{S_N}\right)\tilde{\boldsymbol{\Pi}}^{(h)}$$

where:

$$\tilde{\boldsymbol{\Pi}}^{(h)} = \begin{cases} \frac{1}{N}\sum_{n=h+1}^{N}(\mathbf{x}_n - \mathbf{p})(\mathbf{x}_{n-h} - \mathbf{p})' & h \geq 0, \\ \frac{1}{N}\sum_{n=-h+1}^{N}(\mathbf{x}_{n+h} - \mathbf{p})(\mathbf{x}_n - \mathbf{p})' & h < 0. \end{cases}$$

(Note that $\tilde{\boldsymbol{\Pi}}^{(h)}$ is similar to $\boldsymbol{\Pi}^{(h)}$, but the centering is different.) Now, we have $\mathbb{E}\tilde{\boldsymbol{\Pi}}^{(h)} = \frac{N-h}{N}\boldsymbol{\Pi}^{(h)}$ and:

$$\mathbb{E}\tilde{\boldsymbol{\Sigma}} = \mathbb{E}\tilde{\boldsymbol{\Pi}}^{(0)} + 2\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right)\left(1 - \frac{h}{N}\right)\boldsymbol{\Pi}^{(h)}.$$

This means that:

$$\begin{aligned}\mathbb{E}\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = &-2\sum_{h=1}^{N-1}\left(1 - k\left(\frac{h}{S_N}\right)\right)\boldsymbol{\Pi}^{(h)} \\ &- \frac{2}{N}\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right)h\boldsymbol{\Pi}^{(h)} - 2\sum_{h=N}^{\infty}\boldsymbol{\Pi}^{(h)}.\end{aligned}$$

The requirement in [93, Lemma A.1] is that $\mathbb{E}\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = O\left(S_N^{1/2}N^{-1/2}\right)$. Let $\boldsymbol{\gamma}$ be a vector with $\|\boldsymbol{\gamma}\|_{L^2} = 1$.

Let us start from the second term. For $q \geq 1$, we have:

$$\begin{aligned}&\frac{1}{N}\left|\boldsymbol{\gamma}'\left\{\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right)h\boldsymbol{\Pi}^{(h)}\right\}\boldsymbol{\gamma}\right| \\ &\quad\leq \frac{1}{N}\sum_{h=1}^{N-1}\left|k\left(\frac{h}{S_N}\right)\right| h\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\quad\leq \frac{1}{N}\sum_{h=1}^{N-1} h\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \leq \frac{1}{N}\sum_{h=1}^{\infty} h\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2}.\end{aligned}$$

This is automatically $O\left(S_N^{1/2}N^{-1/2}\right)$. For $q < 1$, we use the fact that $|h/N| \leq |h/N|^q$ for $h < N$:

$$\begin{aligned}&\frac{1}{N}\left|\boldsymbol{\gamma}'\left\{\sum_{h=1}^{N-1} k\left(\frac{h}{S_N}\right)h\boldsymbol{\Pi}^{(h)}\right\}\boldsymbol{\gamma}\right| \\ &\quad\leq \frac{1}{N}\sum_{h=1}^{N-1}\left|k\left(\frac{h}{S_N}\right)\right| h\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\quad\leq N^{-q}\sum_{h=1}^{N-1} h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2}.\end{aligned}$$

When $q \geq 1/2$, this is automatically $O\left(S_N^{1/2}N^{-1/2}\right)$. When $q < 1/2$, it requires $N^{1/2-q}S_N^{-1/2} = O(1)$.

The last term can be majorized as:

$$\begin{aligned}\left|\boldsymbol{\gamma}'\left\{\sum_{h=N}^{\infty}\boldsymbol{\Pi}^{(h)}\right\}\boldsymbol{\gamma}\right| &\leq \sum_{h=N}^{\infty}\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\leq N^{-q}\sum_{h=N}^{\infty} h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2}.\end{aligned}$$

We need $N^{-q}\sum_{h=N}^{\infty} h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} = O\left(S_N^{1/2}N^{-1/2}\right)$. For $q \geq 1/2$, this is automatically verified. For $q < 1/2$, it is true under $N^{1/2-q}S_N^{-1/2} = O(1)$.

Now we turn to the first term. If $S_N \not\to \infty$ as $N \to \infty$, we just require it to be $O\left(S_N^{1/2}N^{-1/2}\right)$. If $S_N \to \infty$ as $N \to \infty$, we can reason as in [46, p. 284] and in [4, pp. A4-A5]. We have:

$$\begin{aligned}&S_N^q\sum_{h=1}^{N-1}\left(1 - k\left(\frac{h}{S_N}\right)\right)\boldsymbol{\Pi}^{(h)} \\ &\quad= \sum_{h=1}^{N-1}\left(\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right)h^q\boldsymbol{\Pi}^{(h)} + k_q\sum_{h=1}^{N-1} h^q\boldsymbol{\Pi}^{(h)}.\end{aligned}$$

Now, the function defined by $\frac{1-k(x)}{|x|^q}$ for $x \neq 0$ and by k_q for $x = 0$ is non-negative and bounded by a constant M. Hence, $\frac{1-k(x)}{|x|^q} \leq M$. Let us choose a fixed N_0 such that $\sum_{h=N_0}^{\infty} h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \leq \varepsilon/(2M)$ for $\varepsilon > 0$. Then:

$$\begin{aligned}&\left\|\sum_{h=1}^{N-1}\left(\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right)h^q\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\quad= \left\|\sum_{h=1}^{N_0-1}\left(\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right)h^q\boldsymbol{\Pi}^{(h)}\right. \\ &\qquad\left. + \sum_{h=N_0}^{N-1}\left(\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right)h^q\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\quad\leq \sum_{h=1}^{N_0-1}\left|\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right| h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\qquad + \sum_{h=N_0}^{N-1}\left|\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right| h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\quad\leq \sum_{h=1}^{N_0-1}\left|\frac{1 - k(h/S_N)}{(h/S_N)^q} - k_q\right| h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\qquad + 2M\sum_{h=N_0}^{N-1} h^q\left\|\boldsymbol{\Pi}^{(h)}\right\|_{L^2} \\ &\quad\lesssim o(1) + \varepsilon = o(1)\end{aligned}$$

where the first term is $o(1)$ due to the bounded convergence of $\frac{1-k(x)}{|x|^q} - k_q$ to 0 and the second is $o(1)$ due to the arbitrariness

of ε. When $N \to \infty$, $\sum_{h=1}^{N-1} h^q \mathbf{\Pi}^{(h)}$ converges to $\mathbf{F}^{(q)}$. As a result:

$$\begin{aligned}
&-2\sum_{h=1}^{N-1}\left(1-k\left(\frac{h}{S_N}\right)\right)\mathbf{\Pi}^{(h)}\\
&\quad=-2S_N^{-q}\sum_{h=1}^{N-1}\left(\frac{1-k\left(h/S_N\right)}{\left(h/S_N\right)^q}-k_q\right)h^q\mathbf{\Pi}^{(h)}\\
&\qquad-2k_qS_N^{-q}\sum_{h=1}^{N-1}h^q\mathbf{\Pi}^{(h)}\\
&\quad=o\left(S_N^{-q}\right)-2k_qS_N^{-q}\mathbf{F}^{(q)}.
\end{aligned}$$

If $k_q \neq 0$, the second term dominates and we need $-2k_qS_N^{-q}\mathbf{F}^{(q)} = O\left(S_N^{1/2}N^{-1/2}\right)$. If $k_q = 0$, the condition boils down to the one for $S_N \not\to \infty$.

Proof of Proposition 20: We consider a limited development of $\widehat{\text{bias}\left(H_N\right)}$ with respect to $\widehat{\mathbf{\Sigma}}_{ii}$ and q_i respectively around $\mathbf{\Sigma}_{ii}$ and p_i:

$$\begin{aligned}
\widehat{\text{bias}\left(H_N\right)}=&\frac{1}{2N}\sum_{i=1}^{B}\frac{\widehat{\mathbf{\Sigma}}_{ii}}{q_i}=\frac{1}{2N}\sum_{i=1}^{B}\frac{\mathbf{\Sigma}_{ii}+\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)}{p_i\left(1+\frac{q_i-p_i}{p_i}\right)}\\
\simeq&\frac{1}{2N}\sum_{i=1}^{B}\frac{\mathbf{\Sigma}_{ii}+\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)}{p_i}\\
&\cdot\left(1-\frac{q_i-p_i}{p_i}+\left(\frac{q_i-p_i}{p_i}\right)^2\right)\\
=&\frac{1}{2N}\sum_{i=1}^{B}\left(\frac{\mathbf{\Sigma}_{ii}}{p_i}+\frac{\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}}{p_i}-\frac{\mathbf{\Sigma}_{ii}\left(q_i-p_i\right)}{p_i^2}\right.\\
&-\frac{\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)\left(q_i-p_i\right)}{p_i^2}+\frac{\mathbf{\Sigma}_{ii}\left(q_i-p_i\right)^2}{p_i^3}\\
&\left.+\frac{\left(\widehat{\mathbf{\Sigma}}_{ii}-\mathbf{\Sigma}_{ii}\right)\left(q_i-p_i\right)^2}{p_i^3}\right).
\end{aligned}$$

Now, $q_i - p_i = O_{\mathbb{P}}\left(N^{-1/2}\right)$ and, under AC, Lemma 34 implies that $\widehat{\mathbf{\Sigma}}_{ii} - \mathbf{\Sigma}_{ii} = O_{\mathbb{P}}\left(\left(S_N/N\right)^{1/2}\right)$. At last we get:

$$\widehat{\text{bias}\left(H_N\right)} \simeq \text{bias}\left(H_N\right) + O_{\mathbb{P}}\left(S_N^{1/2}N^{-3/2}\right). \qquad \text{(VII.8)}$$

For the Markov case, we refer to Lemma 33 in Section VII-F.

Proof of Corollary 22: We only consider the case of Section IV-A. It is simple to see that:

$$\begin{aligned}
&\sqrt{N}\left(H_N-\widehat{\text{bias}\left(H_N\right)}-H_\infty\right)\\
&\quad=\sqrt{N}\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)\\
&\qquad+\sqrt{N}\left(\text{bias}\left(H_N\right)-\widehat{\text{bias}\left(H_N\right)}\right)\\
&\quad=\sqrt{N}\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)+O_{\mathbb{P}}\left(S_N^{1/2}N^{-1}\right)
\end{aligned}$$

and:

$$\begin{aligned}
&N\left(H_N-\widehat{\text{bias}\left(H_N\right)}-H_\infty\right)\\
&\quad=N\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)\\
&\qquad+N\left(\text{bias}\left(H_N\right)-\widehat{\text{bias}\left(H_N\right)}\right)\\
&\quad=N\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)+O_{\mathbb{P}}\left(S_N^{1/2}N^{-1/2}\right)
\end{aligned}$$

and, provided $S_N = o\left(N\right)$, the results of Corollaries 10 and 15 hold.

Proof of Corollary 23: Let us first consider the case when $\sigma^2 > 0$. Then, from Proposition 7 it is trivial to see that:

$$\text{MSE}\left(H_N\right) = \mathbb{E}\left(H_N - H_\infty\right)^2 = O\left(N^{-1}\right).$$

We have:

$$\begin{aligned}
&\text{MSE}\left(H_N\right)-\text{MSE}\left(H_N-\text{bias}\left(H_N\right)\right)\\
&\quad=\mathbb{E}\left(H_N-H_\infty\right)^2-\mathbb{E}\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)^2\\
&\quad=\text{bias}\left(H_N\right)\left(2\mathbb{E}H_N-\text{bias}\left(H_N\right)-2H_\infty\right)\\
&\quad\simeq\left[\text{bias}\left(H_N\right)\right]^2=O\left(N^{-2}\right).
\end{aligned}$$

At last:

$$\begin{aligned}
&\text{MSE}\left(H_N-\widehat{\text{bias}\left(H_N\right)}\right)-\text{MSE}\left(H_N-\text{bias}\left(H_N\right)\right)\\
&\quad=\mathbb{E}\left(H_N-\widehat{\text{bias}\left(H_N\right)}-H_\infty\right)^2\\
&\qquad-\mathbb{E}\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)^2\\
&\quad=\mathbb{E}\left(\text{bias}\left(H_N\right)-\widehat{\text{bias}\left(H_N\right)}\right)\\
&\qquad\cdot\left(2H_N-\widehat{\text{bias}\left(H_N\right)}-\text{bias}\left(H_N\right)-2H_\infty\right)\\
&\quad=O\left(\left[\mathbb{E}\left(\text{bias}\left(H_N\right)-\widehat{\text{bias}\left(H_N\right)}\right)^2\right]^{1/2}\right.\\
&\qquad\cdot\left[\mathbb{E}\left(\left(H_N-\widehat{\text{bias}\left(H_N\right)}-H_\infty\right)\right.\right.\\
&\qquad\left.\left.+\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)\right)^2\right]^{1/2}\Bigg)\\
&\quad=O\left(\left[\mathbb{E}\left(\text{bias}\left(H_N\right)-\widehat{\text{bias}\left(H_N\right)}\right)^2\right]^{1/2}\right.\\
&\qquad\cdot\left[2\left(\mathbb{E}\left(H_N-\widehat{\text{bias}\left(H_N\right)}-H_\infty\right)^2\right.\right.\\
&\qquad\left.\left.+\ \mathbb{E}\left(H_N-\text{bias}\left(H_N\right)-H_\infty\right)^2\right)\right]^{1/2}\Bigg)\\
&\quad=O\left(S_N^{1/2}N^{-2}\right)
\end{aligned}$$

where the third step comes from Cauchy-Schwarz inequality, the fourth from the inequality $\left(a+b\right)^2 \leq 2\left(a^2+b^2\right)$, and the fifth from Proposition 20 and Corollaries 10 and 22.

When $\sigma^2 = 0$, from Proposition 11:

$$\text{MSE}\left(H_N\right) = \mathbb{E}\left(H_N - H_\infty\right)^2 = O\left(N^{-2}\right).$$

The formula for $\text{MSE}\left(H_N\right) - \text{MSE}\left(H_N - \text{bias}\left(H_N\right)\right)$ remains unchanged. The formula for $\text{MSE}\left(H_N - \widehat{\text{bias}\left(H_N\right)}\right) - \text{MSE}\left(H_N - \text{bias}\left(H_N\right)\right)$ becomes:

$$\begin{aligned}
&\text{MSE}\left(H_N-\widehat{\text{bias}\left(H_N\right)}\right)-\text{MSE}\left(H_N-\text{bias}\left(H_N\right)\right)\\
&\quad=O\left(\left[\mathbb{E}\left(\text{bias}\left(H_N\right)-\widehat{\text{bias}\left(H_N\right)}\right)^2\right]^{1/2}\right.\\
&\qquad\cdot\left[\mathbb{E}\left(\left(H_N-\widehat{\text{bias}\left(H_N\right)}-H_\infty\right)\right.\right.
\end{aligned}$$

$$+ \left.\left(H_N - \operatorname{bias}(H_N) - H_\infty\right)\right)^2\Big]^{1/2}\Big)$$
$$= O\left(S_N^{1/2} N^{-5/2}\right)$$

if Corollary 10 is replaced by Corollary 15.

H. Error in the Estimation of the Distribution Under Degeneracy

Proof of Proposition 25: Suppose that the matrix $\mathbf{\Omega}$ is estimated through $\hat{\mathbf{\Omega}}$, where $\hat{\mathbf{\Omega}} - \mathbf{\Omega} = o_{\mathbb{P}}(1)$. This means that we will replace the distribution of $\sum_{i=1}^B \lambda_i \chi_{1,i}^2$ by the distribution of $\sum_{i=1}^B \hat{\lambda}_i \chi_{1,i}^2$. We would like to characterize the error in this replacement through a bound on:

$$\left\| F_{\sum_{i=1}^B \hat{\lambda}_i \chi_{1,i}^2} - F_{\sum_{i=1}^B \lambda_i \chi_{1,i}^2} \right\|_\infty.$$

The eigenvalues of both $\mathbf{\Omega}$ and $\hat{\mathbf{\Omega}}$ are non-negative as both of them are variance matrices. Let $B^\star$ be the number of non-zero eigenvalues of $\mathbf{\Sigma}^\star$, so that $\lambda_{B^\star} > 0$. We need to differentiate the case $B^\star = 1$ from the case $B^\star > 1$.

By the Wielandt-Hoffman inequality (see [60, p. 126]), we have:

$$\left|\lambda_i - \hat{\lambda}_i\right| \leq \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1, \quad i = 1, \ldots, B$$

and therefore $\left|\lambda_{B^\star} - \hat{\lambda}_{B^\star}\right| \leq \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1$ or $\lambda_{B^\star} - \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1 \leq \hat{\lambda}_{B^\star}$. As $\left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1 = o_{\mathbb{P}}(1)$, for N large enough, $\hat{\lambda}_{B^\star} > 0$, so that the two matrices have ultimately the same rank.

We start from the case $B^\star > 1$. We define $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_B)$, $\hat{\boldsymbol{\lambda}} = \left(\hat{\lambda}_1, \ldots, \hat{\lambda}_B\right)$, $\mathbf{\Lambda} = \operatorname{dg}(\boldsymbol{\lambda})$ and $\hat{\mathbf{\Lambda}} = \operatorname{dg}\left(\hat{\boldsymbol{\lambda}}\right)$. The techniques in [22], [106] do not work directly here, as they require part of the eigenvalues to coincide. However, we can write:

$$\begin{aligned}
&\left\| F_{\sum_{i=1}^B \hat{\lambda}_i \chi_{1,i}^2} - F_{\sum_{i=1}^B \lambda_i \chi_{1,i}^2} \right\|_\infty \\
&= \sup_{x \geq 0} \left| \mathbb{P}\left\{ \sum_{i=1}^B \hat{\lambda}_i \chi_{1,i}^2 \leq x \right\} - \mathbb{P}\left\{ \sum_{i=1}^B \lambda_i \chi_{1,i}^2 \leq x \right\} \right| \\
&= \sup_{x \geq 0} \Big| \mathbb{P}\left\{ \left\| \mathcal{N}\left(\mathbf{0}, \hat{\mathbf{\Lambda}}\right) \right\|_{L_2} \leq \sqrt{x} \right\} \\
&\qquad - \mathbb{P}\left\{ \left\| \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}) \right\|_{L_2} \leq \sqrt{x} \right\}. \Big|
\end{aligned}$$

We use Theorem 1 in [78]:

$$\begin{aligned}
&\sup_{x \geq 0} \left| \mathbb{P}\left\{ \left\| \mathcal{N}\left(\mathbf{0}, \hat{\mathbf{\Lambda}}\right) \right\|_{L_2} \leq x \right\} - \mathbb{P}\left\{ \left\| \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}) \right\|_{L_2} \leq x \right\} \right| \\
&\leq C \cdot \left\{ \left(\sum_{i=1}^B \lambda_i^2 \cdot \sum_{i=2}^B \lambda_i^2 \right)^{-1/4} + \left(\sum_{i=1}^B \hat{\lambda}_i^2 \cdot \sum_{i=2}^B \hat{\lambda}_i^2 \right)^{-1/4} \right\} \\
&\quad \cdot \sum_{i=1}^B \left| \lambda_i - \hat{\lambda}_i \right|
\end{aligned}$$

for an absolute constant $C > 0$. By the Wielandt-Hoffman inequality (see [60, p. 126]), $\sum_{i=1}^B \left|\lambda_i - \hat{\lambda}_i\right| \leq \left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1$. Now, $\hat{\mathbf{\Omega}} - \mathbf{\Omega} = o_{\mathbb{P}}(1)$ implies that also $\hat{\lambda}_i - \lambda_i = o_{\mathbb{P}}(1)$, so that, for N large enough:

$$\left\| F_{\sum_{i=1}^B \hat{\lambda}_i \chi_{1,i}^2} - F_{\sum_{i=1}^B \lambda_i \chi_{1,i}^2} \right\|_\infty \leq C \frac{\left\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\right\|_1}{\left(\sum_{i=1}^B \lambda_i^2 \cdot \sum_{i=2}^B \lambda_i^2 \right)^{1/4}}$$

where the constant C is here different from the one seen above.

When $B^\star = 1$, we have:

$$\begin{aligned}
&\left\| F_{\hat{\lambda}_1 \chi_1^2} - F_{\lambda_1 \chi_1^2} \right\|_\infty \\
&= \sup_x \left| \mathbb{P}\left\{ \hat{\lambda}_1 \chi_1^2 \leq x \right\} - \mathbb{P}\left\{ \lambda_1 \chi_1^2 \leq x \right\} \right| \\
&= \sup_x \left| \mathbb{P}\left\{ \left| \hat{\lambda}_1^{1/2} Z \right| \leq x \right\} - \mathbb{P}\left\{ \left| \lambda_1^{1/2} Z \right| \leq x \right\} \right| \\
&= \sup_x \Big| \mathbb{P}\left\{ \hat{\lambda}_1^{1/2} Z \leq x \right\} - \mathbb{P}\left\{ \hat{\lambda}_1^{1/2} Z \leq -x \right\} \\
&\quad - \mathbb{P}\left\{ \lambda_1^{1/2} Z \leq x \right\} + \mathbb{P}\left\{ \lambda_1^{1/2} Z \leq -x \right\} \Big| \\
&\leq 2 \sup_{x > 0} \left| \Phi(x) - \Phi\left(\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} x \right) \right|.
\end{aligned}$$

We are only interested in the case in which $\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} \simeq 1$, therefore we write $\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} = 1 + \varepsilon$ and we get:

$$\Phi(x) - \Phi\left(\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} x \right) = \Phi(x) - \Phi((1+\varepsilon)x) \simeq \phi(x) x \varepsilon.$$

This implies that $\left| \Phi(x) - \Phi\left(\hat{\lambda}_1^{1/2} \lambda_1^{-1/2} x \right) \right| \lesssim \sup_{x \in \mathbb{R}} |\phi(x) x| \cdot |\varepsilon| \leq C |\varepsilon|$ (where C can be taken equal to or larger than $\sup_{x \in \mathbb{R}} |\phi(x) x| = 1/\sqrt{2\pi e} \doteq 0.2419707$). Therefore:

$$\left\| F_{\hat{\lambda}_1 \chi_1^2} - F_{\lambda_1 \chi_1^2} \right\|_\infty \lesssim 2C \left| \frac{\hat{\lambda}_1 - \lambda_1}{\lambda_1^{1/2} \left(\hat{\lambda}_1^{1/2} + \lambda_1^{1/2} \right)} \right| \lesssim C \left| \frac{\hat{\lambda}_1 - \lambda_1}{\lambda_1} \right|.$$

In this case too, $\left\| F_{\sum_{i=1}^B \hat{\lambda}_i \chi_{1,i}^2} - F_{\sum_{i=1}^B \lambda_i \chi_{1,i}^2} \right\|_\infty = O\left(\left\| \hat{\mathbf{\Omega}} - \mathbf{\Omega} \right\|_1 \right)$.

The final part of the statement comes from the results in Lemma A.1 in [93] under assumption AC.

APPENDIX

Consider an iid standard Gaussian sequence $\{y_1, y_2, \ldots\}$ and a standard Gaussian random variable z independent of the previous sequence. Then we define:

$$\widetilde{x}_i = \beta z + \left(1 - \beta^2\right)^{1/2} y_i.$$

As above, the dichotomized process is based on the signs of the original process:

$$x_i = 1 + 1\{\widetilde{x}_i \geq 0\}.$$

It is clear that:

$$p_1 = \mathbb{P}\{x_i = 1\} = \mathbb{P}\{\widetilde{x}_i < 0\} = 1/2$$
$$p_2 = 1 - p_1 = 1/2.$$

However, $\operatorname{Cov}(\widetilde{x}_1, \widetilde{x}_{h+1}) = \operatorname{Cov}\left(\beta z + \left(1 - \beta^2\right)^{1/2} y_1, \beta z + \left(1 - \beta^2\right)^{1/2} y_{h+1} \right) = \beta^2$. From [114, p. 189], we have:

$$p_{22}^{(h)} = p_{11}^{(h)} = 1/4 + 1/2\pi \arcsin\left(\beta^2\right)$$

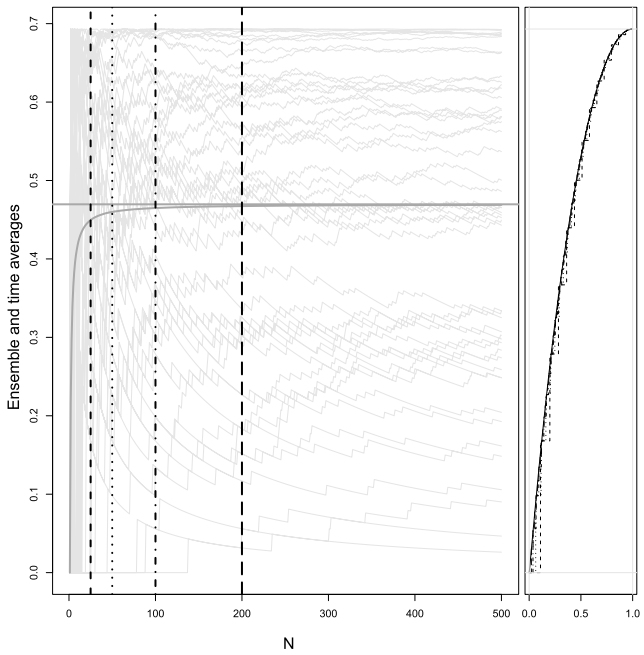

Fig. 10. Ensemble and time averages of the entropy in the non-ergodic case: on the left plot, 50 trajectories of H_N as a function of N (light grey jigsaw lines), $\mathbb{E}H_\infty$ (dark grey horizontal line), $\mathbb{E}H_N$ (dark grey curved line), vertical lines at $N \in \{25, 50, 100, 200\}$ (respectively black dashed, dotted, dash-dot, long dashed lines); on the right plot, empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black long dashed line) and $N = \infty$ (black solid line).

$$p_{12}^{(h)} = p_{21}^{(h)} = 1/2 - p_{22}^{(h)}.$$

In this case, the process $\{\widetilde{x}_1, \widetilde{x}_2, \dots\}$ is stationary but non-ergodic (see, e.g., Example 13.9 in [25, p. 196]), and so is $\{x_1, x_2, \dots\}$. Therefore:

$$\begin{aligned} q_1 = \frac{n_1}{N} &= \frac{\sum_{j=1}^{N} 1\{x_j = 1\}}{N} = \frac{\sum_{j=1}^{N} 1\{\widetilde{x}_j < 0\}}{N} \\ &= \frac{\sum_{j=1}^{N} 1\left\{y_i < -\frac{\beta}{(1-\beta^2)^{1/2}} z\right\}}{N} \\ &\to \Phi\left(-\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \quad \mathbb{P}-\text{as.} \end{aligned}$$

Note that z is an invariant random variable. At last the limit of the observed entropy is:

$$\begin{aligned} H_\infty = &- \Phi\left(-\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \ln \Phi\left(-\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \\ &- \Phi\left(\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \ln \Phi\left(\frac{\beta}{(1-\beta^2)^{1/2}} z\right). \end{aligned}$$

In Figure 10 we show what is the behavior of the statistic in this case. The grey jigsaw lines represent some trajectories of H_N, while the dark grey curved line represents $\mathbb{E}H_N$ (based on 250,000 samples) and the dark grey horizontal line represents $\mathbb{E}H_\infty$. On the right plot, we display the empirical cdf of H_N with $N = 25$ (black dashed line), $N = 50$ (black dotted line), $N = 100$ (black dash-dot line), $N = 200$ (black long dash line). The black solid line represents the empirical cdf of H_∞. In the non-ergodic case H_N converges (almost surely) to the random variable H_∞.

In this case, $H_\infty \neq \mathbb{E}H_\infty$. The first quantity has been obtained above. The second one is given by:

$$\begin{aligned} \mathbb{E}H_\infty = &- \mathbb{E}\left\{\Phi\left(-\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \ln \Phi\left(-\frac{\beta}{(1-\beta^2)^{1/2}} z\right)\right\} \\ &+ \mathbb{E}\left\{\Phi\left(\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \ln \Phi\left(\frac{\beta}{(1-\beta^2)^{1/2}} z\right)\right\} \\ = &- 2\mathbb{E}\left\{\Phi\left(\frac{\beta}{(1-\beta^2)^{1/2}} z\right) \ln \Phi\left(\frac{\beta}{(1-\beta^2)^{1/2}} z\right)\right\}. \end{aligned}$$

The first-order bias correction of $\mathbb{E}H_N$ takes the form:

$$\begin{aligned} &- \frac{B-1}{2N} - \frac{1}{N} \sum_{i=1}^{B} \frac{\sum_{h=1}^{N-1} \left(p_i^{(h)} - p_i^2\right)}{p_i} \\ &= -\frac{1}{2N} - \frac{2(N-1)}{\pi N} \arcsin\left(\beta^2\right) = O(1). \end{aligned}$$

Note that, while the first-order bias correction reduces the bias in the estimation of $\mathbb{E}H_\infty$ through $\mathbb{E}H_N$, nothing guarantees that this reduces the bias in the estimation of H_∞ through H_N.

ACKNOWLEDGMENT

The authors are grateful to the associate editor, two reviewers, and Emmanuel Rio for insightful comments which improved the original manuscript. Moreover, they would like to thank the participants of the Institute of Economics Seminar Series, Sant'Anna School of Advanced Studies, and the Economics Online Seminars, University of Insubria, for helpful discussions.

REFERENCES

[1] I. Ahmad and P.-E. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions (Corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 3, pp. 372–375, May 1976.

[2] D. J. Albers and G. Hripcsak, "Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series," *Chaos, Solitons Fractals*, vol. 45, no. 6, pp. 853–860, Jun. 2012.

[3] P. H. Algoet and T. M. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *Ann. Probab.*, vol. 16, no. 2, pp. 899–909, Apr. 1988.

[4] D. W. K. Andrews, "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," Cowles Found. Res. Econ., Yale Univ., New Haven, CT, USA, Tech. Rep. 877R, 1989.

[5] D. W. K. Andrews, "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica*, vol. 59, pp. 817–858, May 1991.

[6] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Struct. Algorithms*, vol. 19, nos. 3–4, pp. 163–193, 2001.

[7] A. Antos and I. Kontoyiannis, "Estimating the entropy of discrete distributions," in *Proc. IEEE Int. Symp. Inf. Theory*, Jan. 2001, p. 45.

[8] T. M. Apostol, "An elementary view of Euler's summation formula," *Amer. Math. Monthly*, vol. 106, no. 5, pp. 409–418, May 1999.

[9] S. Assaf, P. Diaconis, and K. Soundararajan, "A rule of thumb for riffle shuffling," *Ann. Appl. Probab.*, vol. 21, no. 3, pp. 843–875, Jun. 2011.

[10] A. D. Back, D. Angus, and J. Wiles, "Determining the number of samples required to estimate entropy in natural sequences," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4345–4352, Jul. 2019.

[11] G. P. Basharin, "On a statistical estimate for the entropy of a sequence of independent random variables," *Theory Probab. Appl.*, vol. 4, no. 3, pp. 333–336, Jan. 1959.

[12] D. Bayer and P. Diaconis, "Trailing the dovetail shuffle to its lair," *Ann. Appl. Probab.*, vol. 2, no. 2, pp. 294–313, May 1992.

[13] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[14] M. Bernasconi, C. Choirat, and R. Seri, "A re-examination of the algebraic properties of the AHP as a ratio-scaling technique," *J. Math. Psychol.*, vol. 55, no. 2, pp. 152–165, Apr. 2011.

[15] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *Ann. Statist.*, vol. 47, no. 1, pp. 288–318, 2019.

[16] R. N. Bhattacharya and N. H. Chan, "Comparisons of chisquare, edgeworth expansions and bootstrap approximations to the distribution of the frequency chisquare," *Sankhyā Indian J. Statist., Ser. A*, vol. 58, no. 1, pp. 57–68, 1996.

[17] T. Biemann and E. Kearney, "Size does matter: How varying group sizes in a sample affect the most common measures of group diversity," *Organizational Res. Methods*, vol. 13, no. 3, pp. 582–599, Jul. 2010.

[18] J. A. Bonachela, H. Hinrichsen, and M. A. Muñoz, "Entropy estimates of small data sets," *J. Phys., Math. Theor.*, vol. 41, no. 20, May 2008, Art. no. 202001.

[19] D. A. Butts and D. S. Rokhsar, "The information content of spontaneous retinal waves," *J. Neurosci.*, vol. 21, no. 3, pp. 961–973, Feb. 2001.

[20] A. G. Carlton, "On the bias of information estimates," *Psychol. Bull.*, vol. 71, no. 2, pp. 108–109, 1969.

[21] C. Choirat, C. Hess, and R. Seri, "A functional version of the Birkhoff ergodic theorem for a normal integrand: A variational approach," *Ann. Probab.*, vol. 31, no. 1, pp. 63–92, Jan. 2003.

[22] C. Choirat and R. Seri, "Computational aspects of Cui-Freeden statistics for equidistribution on the sphere," *Math. Comput.*, vol. 82, no. 284, pp. 2137–2156, Apr. 2013.

[23] M. Ciucu, "No-feedback card guessing for dovetail shuffles," *Ann. Appl. Probab.*, vol. 8, no. 4, pp. 1251–1269, Nov. 1998.

[24] I. P. Cornfeld and G. Ya Sinai, "Entropy theory of dynamical systems," in *Dynamical Systems II*, vol. 2, R. V. Gamkrelidze and Y. G. Sinai, Eds. Berlin, Germany: Springer, 1989, pp. 36–58.

[25] J. Davidson, "Stochastic limit theory: An introduction for econometricians," in *Advanced Texts in Econometrics*. London, U.K.: Oxford Univ. Press, 1994.

[26] P. Doukhan, P. Massart, and E. Rio, "The functional central limit theorem for strongly mixing processes," *Annales l'IHP Probabilités Statistiques*, vol. 30, no. 1, pp. 63–82, 1994.

[27] C.-G. Esseen, "Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law," *Acta Math.*, vol. 77, no. 1, pp. 1–125, 1945.

[28] J. P. Florens, M. Mouchart, and J. M. Rolin, "Noncausality and marginalization of Markov processes," *Econ. Theory*, vol. 9, no. 2, pp. 241–262, Apr. 1993.

[29] A. M. Fraser, "Information and entropy in strange attractors," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 245–262, Mar. 1989.

[30] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A, Gen. Phys.*, vol. 33, no. 2, pp. 1134–1140, Feb. 1986.

[31] A. R. Gallant, "Nonlinear statistical models," in *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: Wiley, 1987.

[32] Y. Gao, I. Kontoyiannis, and E. Bienenstock, "Estimating the entropy of binary time series: Methodology, some theory and a simulation study," *Entropy*, vol. 10, no. 2, pp. 71–99, Jun. 2008.

[33] W. A. Gardner, A. Napolitano, and L. Paura, "Cyclostationarity: Half a century of research," *Signal Process.*, vol. 86, no. 4, pp. 639–697, 2006.

[34] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Statist. Rev.*, vol. 70, no. 3, pp. 419–435, Dec. 2002.

[35] F. Götze and V. V. Ulyanov, "Asymptotic distribution of χ^2-type statistics," in *Preprintreihe der Forschergruppe Spektrale Analysis Und stochastische Dynamik 03–033*. Bielefeld, Germany: Universität Bielefeld, 2003.

[36] C. Gourieroux, A. Monfort, and A. Trognon, "A general approach to serial correlation," *Econ. Theory*, vol. 1, no. 3, pp. 315–340, Dec. 1985.

[37] C. Granger and J.-L. Lin, "Using the mutual information coefficient to identify lags in nonlinear models," *J. Time Ser. Anal.*, vol. 15, no. 4, pp. 371–384, Jul. 1994.

[38] P. Grassberger, "Finite sample corrections to entropy and dimension estimates," *Phys. Lett. A*, vol. 128, nos. 6–7, pp. 369–373, 1988.

[39] P. Grassberger, "Estimating the information content of symbol sequences and efficient codes," *IEEE Trans. Inf. Theory*, vol. 35, no. 3, pp. 669–675, May 1989.

[40] P. Grassberger, "Entropy estimates from insufficient samplings," 2003, *arXiv:physics/0307138*. [Online]. Available: https://arxiv.org/abs/physics/0307138

[41] P. Grassberger and I. Procaccia, "Estimation of the Kolmogorov entropy from a chaotic signal," *Phys. Rev. A, Gen. Phys.*, vol. 28, no. 4, pp. 2591–2593, Oct. 1983.

[42] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, 2nd ed. New York, NY, USA: Springer, 2009.

[43] R. M. Gray and J. C. Kieffer, "Asymptotically mean stationary measures," *Ann. Probab.*, vol. 8, no. 5, pp. 962–973, 1980.

[44] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley, "Analysis of symbolic sequences using the Jensen-Shannon divergence," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 65, no. 4, Mar. 2002, Art. no. 041905.

[45] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, no. 1, pp. 69–88, 1993.

[46] E. J. Hannan Ed., "Multiple time series," in *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: Wiley, 1970.

[47] B. Harris, "The statistical estimation of entropy in the non-parametric case," in *Topics in Information Theory*, 1st ed, E. Csiszar, Ed. Amsterdam, The Netherlands: North-Holland, 1975, pp. 323–355.

[48] H. Herzel, A. O. Schmitt, and W. Ebeling, "Finite sample effects in sequence analysis," *Chaos, Solitons Fractals*, vol. 4, no. 1, pp. 97–113, Jan. 1994.

[49] C. Hess, R. Seri, and C. Choirat, "Ergodic theorems for extended real-valued random variables," *Stochastic Processes Their Appl.*, vol. 120, no. 10, pp. 1908–1919, Sep. 2010.

[50] A. Hjørungnes and D. Gesbert, "Complex-valued matrix differentiation: Techniques and key results," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2740–2746, Jun. 2007.

[51] A. B. Holt *et al.*, "Phase-dependent suppression of beta oscillations in Parkinson's disease patients," *J. Neurosci.*, vol. 39, no. 6, pp. 1119–1134, Feb. 2019.

[52] Y. Hong and H. White, "Asymptotic distribution theory for nonparametric entropy measures of serial dependence," *Econometrica*, vol. 73, no. 3, pp. 837–901, May 2005.

[53] I. A. Ibragimov and Ju. V. Linnik, *Independent and Stationary Sequences of Random Variables*. Groningen, The Netherlands: Wolters-Noordhoff, 1971.

[54] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.

[55] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Maximum likelihood estimation of functionals of discrete distributions," *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6774–6798, Oct. 2017.

[56] Y. Jo and N. Z. Cho, "Acceleration and real variance reduction in continuous-energy Monte Carlo whole-core calculation via p-CMFD feedback," *Nucl. Sci. Eng.*, vol. 189, no. 1, pp. 26–40, Jan. 2018.

[57] H. Joe, "Relative entropy measures of multivariate dependence," *J. Amer. Stat. Assoc.*, vol. 84, no. 405, pp. 157–164, Mar. 1989.

[58] A. Kaltchenko, N. Timofeeva, and E. A. Timofeev, "Bias reduction of the nearest neighbor estimator," *Int. J. Bifurcation Chaos*, vol. 18, no. 12, pp. 3781–3787, 2008.

[59] A. Kaltchenko, E.-H. Yang, and N. Timofeeva, "Entropy estimators with almost sure convergence and an $O(n^{-1})$ variance," in *Proc. IEEE Inf. Theory Workshop*, Sep. 2007, pp. 644–649.

[60] T. Katō, "Perturbation theory for linear operators," in *Number 132 in Die Grundlehren der Mathematischen Wissenschaften in Einzeldarstellungen*. 2nd ed. Berlin, Germany: Springer, 1984.

[61] J. G. Kemeny, "Generalization of a fundamental matrix," *Linear Algebra Appl.*, vol. 38, pp. 193–206, Jun. 1981.

[62] J. G. Kemeny and J. L. Snell, *Finite Markov Chains: With a New Appendix 'Generalization of a Fundamental Matrix' Undergraduate Texts in Mathematics*. New York, NY, USA: Springer, 1983.

[63] M. Kennel, J. Shlens, H. Abarbanel, and E. Chichilnisky, "Estimating entropy rates with Bayesian confidence intervals," *Neural Comput.*, vol. 17, no. 7, pp. 1531–1576, Jul. 2005.

[64] I. Kontoyiannis, "Asymptotic recurrence and waiting times for stationary processes," *J. Theor. Probab.*, vol. 11, no. 3, pp. 795–811, 1998.

[65] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1319–1327, May 1998.

[66] D. Kugiumtzis, "Partial transfer entropy on rank vectors," *Eur. Phys. J. Special Topics*, vol. 222, no. 2, pp. 401–420, Jun. 2013.

[67] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[68] F. Lamperti, "An information theoretic criterion for empirical validation of simulation models," *Econometrics Statist.*, vol. 5, pp. 83–106, Jan. 2018.

[69] E. L. Lehmann and J. P. Romano, "Testing statistical hypotheses," in *Springer Texts in Statistics*, 3rd ed. New York, NY, USA: Springer, 2005.

[70] L. Li, I. M. Park, S. Seth, J. C. Sanchez, and J. C. Principe, "Functional connectivity dynamics among cortical neurons: A dependence analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 1, pp. 18–30, Jan. 2012.

[71] D. Lombardi and S. Pant, "Nonparametric k-nearest-neighbor entropy estimator," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 93, no. 1, 2016, Art. no. 013310.

[72] R. D. Luce, "A survey of the theory of selective information and some of its behavioral applications," in *Developments in Mathematical Psychology*, R. D. Luce, Ed. Glencoe, U.K.: Free Press, 1960, pp. 1–119.

[73] R. D. Luce, "Whatever happened to information theory in psychology," *Rev. Gen. Psychol.*, vol. 7, no. 2, pp. 183–188, 2003.

[74] G. W. Mackey, "Ergodic theory and its significance for statistical mechanics and probability theory," *Adv. Math.*, vol. 12, no. 2, pp. 178–268, Feb. 1974.

[75] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. New York, NY, USA: Wiley, 1999.

[76] M. Mata and J. Machado, "Entropy analysis of monetary unions," *Entropy*, vol. 19, no. 6, p. 245, May 2017.

[77] G. A. Miller, "Note on the bias of information estimates," in *Information Theory in Psychology; Problems and Methods*, H. Quastler, Ed. Glencoe, U.K.: Free Press, 1955, pp. 95–100.

[78] A. A. Naumov, V. G. Spokoiny, Y. E. Tavyrikov, and V. V. Ulyanov, "Nonasymptotic estimates for the closeness of Gaussian measures on balls," *Doklady Math.*, vol. 98, no. 2, pp. 490–493, Sep. 2018.

[79] W. K. Newey and D. McFadden, "Chapter 36 Large sample estimation and hypothesis testing," in *Handbook Econometrics*, vol. 4. Amsterdam, The Netherlands: Elsevier, 1994, pp. 2111–2245.

[80] W. K. Newey and K. D. West, "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix," *Econometrica*, vol. 55, no. 3, pp. 703–708, 1987.

[81] W. K. Newey and K. D. West, "Automatic lag selection in covariance matrix estimation," *Rev. Econ. Stud.*, vol. 61, no. 4, pp. 631–653, Oct. 1994.

[82] M. Paluš, "Testing for nonlinearity using redundancies: Quantitative and qualitative aspects," *Phys. D, Nonlinear Phenomena*, vol. 80, nos. 1–2, pp. 186–205, Jan. 1995.

[83] M. Paluš, "Coarse-grained entropy rates for characterization of complex time series," *Phys. D, Nonlinear Phenomena*, vol. 93, nos. 1–2, pp. 64–77, May 1996.

[84] M. Paluš, "Detecting nonlinearity in multivariate time series," *Phys. Lett. A*, vol. 213, nos. 3–4, pp. 138–147, Apr. 1996.

[85] M. Paluš, V. Albrecht, and I. Dvorák, "Information theoretic test for nonlinearity in time series," *Phys. Lett. A*, vol. 175, nos. 3–4, pp. 203–209, Apr. 1993.

[86] X. Pan, L. Hou, M. Stephen, H. Yang, and C. Zhu, "Evaluation of scaling invariance embedded in short time series," *PLoS ONE*, vol. 9, no. 12, Dec. 2014, Art. no. e116128.

[87] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, 2003.

[88] L. Paninski and M. Yajima, "Undersmoothed kernel entropy estimators," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4384–4388, Sep. 2008.

[89] S. Panzeri and A. Treves, "Analytical estimates of limited sampling biases in different information measures," *Netw., Comput. Neural Syst.*, vol. 7, no. 1, pp. 87–107, Jan. 1996.

[90] S. Papadimitriou, S. Mavroudi, and S. D. Likothanassis, "Mutual information clustering for efficient mining of fuzzy association rules with application to gene expression data analysis," *Int. J. Artif. Intell. Tools*, vol. 15, no. 2, pp. 227–250, Apr. 2006.

[91] M. Papapetrou and D. Kugiumtzis, "Markov chain order estimation with conditional mutual information," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 7, pp. 1593–1601, Apr. 2013.

[92] M. Papapetrou and D. Kugiumtzis, "Markov chain order estimation with parametric significance tests of conditional mutual information," *Simul. Model. Pract. Theory*, vol. 61, pp. 1–13, Feb. 2016.

[93] D. N. Politis, "Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices," *Econ. Theory*, vol. 27, no. 4, pp. 703–744, Aug. 2011.

[94] B. Pompe, "Measuring statistical dependences in a time series," *J. Stat. Phys.*, vol. 73, nos. 3–4, pp. 587–610, Nov. 1993.

[95] H. Ren, Y. Yang, C. Gu, T. Weng, and H. Yang, "A patient suffering from neurodegenerative disease may have a strengthened fractal gait rhythm," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 9, pp. 1765–1772, Sep. 2018.

[96] S. I. Resnick, *A Probability Path*. Boston, MA, USA: Birkhäuser, 1999.

[97] E. Rio, "Sur le théorème de Berry-Esseen pour les suites faiblement dépendantes," *Probab. Theory Rel. Fields*, vol. 104, no. 2, pp. 255–282, Jun. 1996.

[98] E. Rio, "About the Lindeberg method for strongly mixing sequences," *ESAIM, Probab. Statist.*, vol. 1, pp. 35–61, Dec. 1997.

[99] E. Rio, *Asymptotic Theory of Weakly Dependent Random Processes, volume 80 of Probability Theory and Stochastic Modelling*, 1st ed, vol. 80. Berlin, Germany: Springer, 2017.

[100] P. M. Robinson, "On the asymptotic properties of estimators of models containing limited dependent variables," *Econometrica*, vol. 50, no. 1, pp. 27–41, 1982.

[101] M. S. Rogers and B. F. Green, "The moments of sample information when the alternatives are equally likely," in *Information Theory in Psychology; Problems and Methods*, H. Quastler, Ed. Glencoe, U.K.: Free Press, 1955, pp. 101–108.

[102] M. S. Roulston, "Estimating the errors on measured entropy and mutual information," *Phys. D, Nonlinear Phenomena*, vol. 125, nos. 3–4, pp. 285–294, 1999.

[103] T. Schürmann and P. Grassberger, "Entropy estimation of symbol sequences," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 6, no. 3, pp. 414–427, 1996.

[104] E. Seneta, "Sensitivity of finite Markov chains under perturbation," *Statist. Probab. Lett.*, vol. 17, no. 2, pp. 163–168, May 1993.

[105] R. J. Serfling, "Approximation theorems of mathematical statistics," in *Wiley Series in Probability and Statistics*. Hoboken, NJ, USA: Wiley, 1980.

[106] R. Seri, "Statistical properties of b-adic diaphonies," *Math. Comput.*, vol. 86, no. 304, pp. 799–828, 2016.

[107] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.

[108] B. Shawyer and B. Watson, "Borel's methods of summability: Theory and applications," in *Oxford Mathematical Monographs*, 1st ed. Oxford, U.K.: Clarendon Press, 1994.

[109] G. R. Shorack, "Probability for statisticians," in *Springer Texts in Statistics*. New York, NY, USA: Springer, 2000.

[110] K. M. Short, "Direct calculation of metric entropy from time series," *J. Comput. Phys.*, vol. 104, no. 1, pp. 162–172, Jan. 1993.

[111] L. Sun and A. G. Nikolaev, "Mutual information based matching for causal inference with observational data," *J. Mach. Learn. Res.*, vol. 17, p. 31, Jan. 2016.

[112] W. Y. Tan, "On the distribution of quadratic forms in normal random variables," *Can. J. Statist.*, vol. 5, no. 2, pp. 241–250, 1977.

[113] Y.-C. Tian and F. Gao, "Extraction of delay information from chaotic time series based on information entropy," *Phys. D, Nonlinear Phenomena*, vol. 108, nos. 1–2, pp. 113–118, Sep. 1997.

[114] Y. L. Tong, "The multivariate normal distribution," in *Springer Series in Statistics*. New York, NY, USA: Springer, 1990.

[115] M. Valadier, "Stationary stochastic processes are mixing of ergodic ones: Contingency," *J. Convex Anal.*, vol. 18, no. 4, pp. 1127–1140, 2011.

[116] G. Valiant and P. Valiant, "Estimating the unseen: Improved estimators for entropy and other properties," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2013, pp. 2157–2165.

[117] A. W. van der Vaart, "Asymptotic statistics," in *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[118] Q. Wang, Y. Shen, and J. Q. Zhangb, "A nonlinear correlation measure for multivariable data set," *Phys. D, Nonlinear Phenomena*, vol. 200, nos. 3–4, pp. 287–295, 2005.

[119] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3702–3720, Jun. 2016.
[120] L. Xia and P. W. Glynn, "A generalized fundamental matrix for computing fundamental quantities of Markov systems," 2016, *arXiv:1604.04343*. [Online]. Available: http://arxiv.org/abs/1604.04343
[121] W. Xiong, L. Faes, and P. C. Ivanov, "Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 95, nos. 1–6, 2017, Art. no. 062114.
[122] Y. Yang, C. Gu, Q. Xiao, and H. Yang, "Evolution of scaling behaviors embedded in sentence series from a story of the stone," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171776.
[123] W. Zhang, L. Qiu, Q. Xiao, H. Yang, Q. Zhang, and J. Wang, "Evaluation of scale invariance in physiological signals by means of balanced estimation of diffusion entropy," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 86, no. 2, 2012, Art. no. 056107.
[124] Z. Zhang, "Asymptotic normality of an entropy estimator with exponentially decaying bias," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 504–508, Jan. 2013.
[125] Z. Zhang and X. Zhang, "A normal law for the plug-in estimator of entropy," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2745–2747, May 2012.
[126] R. Zhou, R. Cai, and G. Tong, "Applications of entropy in finance: A review," *Entropy*, vol. 15, no. 12, pp. 4909–4931, Nov. 2013.
[127] A. Zygmund, "Trigonometric series," in *Cambridge Mathematical Library*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

Raffaello Seri received the M.S. and Ph.D. degrees in management engineering from the Politecnico di Milano, Milan, Italy, in 1996 and 1999, respectively, the joint DEA degree in mathematics applied to economics from ENSAE and Université Paris Dauphine, Paris, France, in 1999, and the second Ph.D. degree in mathematics from Université Paris Dauphine, in 2005.

He is a Full Professor with the DiECO, Università degli Studi dell'Insubria, Varese, Italy. His research interests include statistics, numerical analysis, operations research, mathematical psychology, and economics.

Prof. Seri is a member of the Center for Nonlinear and Complex Systems, Como, Italy, and the Centre for Computational & Organisational Cognition (CORG), University of Southern Denmark, Slagelse, Denmark. He has held visiting positions, among others, at the Institute for Computational and Experimental Research in Mathematics, Brown University, Providence, RI, USA; the Chair of Microeconomics, FSU Jena, Germany; the Lehrstuhl für Innovationsökonomik, Universität Hohenheim, Germany; and the Department of Language and Communication, University of Southern Denmark.

Mario Martinoli received the M.S. degree in economics and finance and the Ph.D. degree in methods and models for economic decisions from the University of Insubria, Varese, Italy, in 2011 and 2021, respectively, and the executive master's degree in financial risk management from the MIP Politecnico di Milano Graduate School of Business, Milan, in 2015.

He has held a visiting position at the Department of Economics, Stony Brook University, Stony Brook, NY, USA, in 2019. Since November 2020, he has been a Post-Doctoral Researcher with the Institute of Economics and EMbeDS, Sant'Anna School of Advanced Studies, Pisa. His research interests include estimation, calibration, validation, and inference for simulation models.