ACCEPTED MANUSCRIPT

Chaos
An Interdisciplinary Journal
of Nonlinear Science

This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

Assessing the robustness of permutation Jensen-Shannon divergence estimated values for time series analysis

Luciano Zunino,[1,2] Pedro Rueda Suescun,[1] Aurelio F. Bariviera,[3] and Diego M. Mateos[4,5]
[1)]*Centro de Investigaciones Ópticas (CONICET La Plata - CIC - UNLP), 1897 Gonnet, La Plata, Argentina*
[2)]*Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de La Plata (UNLP), 1900 La Plata, Argentina*
[3)]*Department of Business, ECO-SOS, Universitat Rovira i Virgili, 43204 Reus, Spain*
[4)]*Instituto de Matemática aplicada del Litoral (IMAL-CONICET-UNL), Santa Fe, Argentina*
[5)]*Facultad de Ciencia y Tecnología, Universidad Autónoma de Entre Ríos (UADER), Oro Verde, Entre Ríos, Argentina*

(*Electronic mail: lucianoz@ciop.unlp.edu.ar)

(Dated: 16 July 2026)

The permutation Jensen-Shannon divergence has been shown to be a versatile tool of great value for diverse applications within the time-series analysis field. Here, we discuss the robustness of this ordinal measure by analyzing the finite-size bias of its estimated values. Synthetic data generated from several stochastic and chaotic systems were carefully analyzed as testbeds, and relevant conclusions were extracted. Finally, to demonstrate the implementation and utility of these findings in a practical context, we analyze the data generated from a complex real-world system of high interest today: Bitcoin prices.

Distinguishing whether two time series come from the same system or share underlying dynamics is critical in time series analysis. Ordinal tools, such as the permutation Jensen-Shannon divergence, are effective for quantifying dissimilarity. However, determining if low divergence values indicate genuine similarity or merely finite-size effects remains challenging. This work addresses this gap by proposing a heuristic approach to evaluate the empirical baseline deviations of permutation Jensen-Shannon divergence values. We offer robust guidelines for implementing pJSd in real-world applications, ensuring reliable discrimination even with limited data.

I. INTRODUCTION

It is clear that the robust characterization of time series obtained from complex systems represents a vital task in revealing the mechanisms that govern the dynamics of the underlying systems. This is even more important in those cases in which time series from an observable are the only information available from the complex system under study. Actually, time series analysis comprises a fruitful set of linear and nonlinear techniques to meet this challenge. Within this research field, ordinal pattern tools have been shown to be particularly successful. The vast list of heterogeneous applications in which these ordinal methods have been implemented since the foundational paper by Bandt and Pompe was published more than 20 years ago[1] is unquestionable evidence of this fact—we refer interested readers to several reviews[2–7] for further details.

Many times a comparison between time series is performed to test similarities and differences. Measuring the "distance" between two time series is critical to determine whether they originate from the same system or share specific underlying characteristics. Actually, the widely implemented surrogate analysis[8] can be mentioned as a relevant example of the latter. Quantifying similarity is also a vital ingredient in time series clustering and classification systems. The Jensen-Shannon divergence (JSd) is a tool that has been proposed to address this comparison issue. Arising from information theory, it measures the similarity between two probability distributions by calculating the entropy of mixing[9]. The interpretations and statistical properties of this divergence measure have previously been analyzed by Grosse et al.[10]. Particularly, and bearing in mind what is of interest in our current research, the expected value of the JSd when computed from an ensemble of finite independent and identically distributed (i.i.d.) sequences of symbols of length N has been approximated.

As a step further, the permutation Jensen-Shannon divergence (pJSd) has recently been proposed by integrating the two previous notions[11]. It is just the JSd between the ordinal probability distributions obtained after symbolizing the two original time series via permutation patterns. This merging approach naturally inherits all the important advantages associated with both the ordinal symbolic analysis and the information theory tool, providing a versatile and practical tool for heterogeneous applications. Moreover, the pJSd outperforms other quantifiers of similarity between signals such as the information-based similarity index[12] and the alphabetic Jensen-Shannon divergence[13].

The main purpose of this research is to try to check how well the approximation given in Ref. 10 provides reference values that help to decide whether two ordinal sequences are generated from the same or different dynamical systems. It is clear that the answer to this question is of considerable statistical relevance. We face this issue following basically a heuristic approach, i.e., by testing ordinal sequences obtained from some well-known stochastic and chaotic models. Long-range correlated and non-stationary models, such as the fractional Brownian motion, have also been considered taking into ac-

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

count that they clearly represent a harder challenge. Furthermore, we have particularly analyzed how using overlapping versus non-overlapping schemes for successive ordinal patterns affects the results.

The remainder of the work is structured as follows. In Sec. II, the main concepts to understand how the pJSd is defined are briefly introduced. After that, in Sec. III, several numerical analyses are included to gain some insights on the topic from controlled scenarios. Differences observed in the results for the overlapping and non-overlapping ordinal mapping schemes are analyzed more carefully in Sec. IV. In Sec. V, a real-world application on financial data of high interest is included to illustrate the practical utility of our previous findings. Finally, the main conclusions obtained from this study are summarized in the last section.

II. METHODS

Below, we briefly review the theoretical background required to understand the definition of the pJSd.

A. The Jensen-Shannon divergence

The JSd[9] is a dissimilarity measure between two arbitrary probability distributions $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$ defined as the entropy of the average distribution minus the average of the entropies, i.e.,

$$D_{JS}(P,Q) = S\left(\frac{P+Q}{2}\right) - \frac{1}{2}S(P) - \frac{1}{2}S(Q), \qquad (1)$$

where S is the Shannon entropy function, i.e. $S(P) = -\sum_{i=1}^{n} p_i \log_2 p_i$, and, as usual, the convention $0 \log_2 0 = 0$ is assumed in accordance with its mathematical limit. This quantity is bounded between 0 and 1, with the minimum value reached when the two distributions under comparison are identical, while the maximum value is obtained whenever their supports are disjoints (that is, $p_i q_i = 0$ for $i = 1, \ldots, n$). The JSd, which stands out for being well-defined even when p_i vanishes without vanishing q_i or vice versa, has had successful applications in heterogeneous fields, from the segmentation of nonstationary symbolic sequences, such as DNA sequences, into stationary subsequences[10,14] and the study of proteins dynamics[15] to the statistical analysis of language[16].

The statistical properties of the JSd have been carefully analyzed[10]. Due to finite size effects, this measure is nonzero even when two time series from the same system are compared. Taking into account these statistical fluctuations, a first-order approximation for the mean of the JSd in finite, uncorrelated sequences generated from the same system was obtained. Assuming that two subsequences are compared, it was demonstrated that

$$\langle D_{JS}(P,Q) \rangle \simeq \frac{k-1}{2Nln(2)}, \qquad (2)$$

with k the number of symbols and N the total length of the symbolic sequence. We will evaluate this expression's suitability for ordinal symbolic mapping.

B. The ordinal mapping

The ordinal mapping is a simple, fast and natural way of discretizing real-valued time series, that takes into account the temporal evolution of the underlying data. Leveraging this last property, ordinal patterns-based methodologies have been widely applied to unveil subtle information about the intrinsic dynamics of complex systems. In order to make the manuscript more self-contained, this symbolic approach is briefly described below.

Given a real-valued time series $\{x_t \in \mathbb{R}, t = 1, \ldots, L\}$, vectors of equally spaced D values of the form $(x_s, x_{s+\tau}, \ldots, x_{s+(D-1)\tau})$ with $s = 1, \ldots, L^*$ and $L^* = L - (D-1)\tau$ are mapped to one of the $D!$ possible ordinal permutations of the same size that describe the order relation between these elements. In other words, each element of the original vector is replaced by its respective ranking in the subset. For example, the vector $(2.5, 3.7, 1.2)$ is mapped to the permutation pattern $(2,3,1)$. Given the set of permutations $\Pi_D = \{\pi_1, \pi_2, \ldots, \pi_{D!}\}$ of length D, the original time series is mapped to the coarse-graining sequence $\{y_s \in \Pi_D, s = 1, \ldots, L^*\}$. The ordinal probability distribution

$$P_\pi^{D,\tau} = \{p(\pi_i), i = 1, \ldots, D!\}, \qquad (3)$$

is then estimated by computing the relative frequency of each ordinal pattern in the symbolized sequence. Two parameters, the number of elements in the permutation patterns D (called order or embedding dimension, $D \geq 2$ with $D \in \mathbb{N}$) and the time separation between the elements in the subsequence τ (called lag or embedding delay, $\tau \in \mathbb{N}$), must be fixed before implementing the method. Regarding their selection, on the one hand, a rule of thumb requires $L \gg D!$, with L the number of data in the original time series, for a robust estimation of $P_\pi^{D,\tau}$. It is also clear that larger values of D offer improved characterization of the underlying temporal structures. On the other hand, a value of the lag $\tau = 1$ (consecutive data) is often used in discrete systems, and also when the chosen sampling frequency is the optimal to characterize the underlying dynamics of continuous systems[5]. However, this arbitrary choice can lead to erroneous conclusions for systems with scale-dependent dynamics[17]. In these cases, a multiscale analysis, in which a given descriptor of the ordinal distribution is estimated as a function of τ, can offer a more comprehensive view[18]. Permutation entropy, defined just as the Shannon entropy of the ordinal probability distribution given by Eq. (3), is the most representative and widely implemented quantifier.

Given that ordinal mapping relies on the relative ranking of data points, the treatment of ties in the analyzed signal requires careful consideration. A significant presence of equal values—common in signals with low amplitude resolution—can lead to spurious conclusions if they are simply symbolized by their order of appearance[19]. While adding low-

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

level noise to break ties is a common solution, the signals analyzed in this work exhibited a negligible number of equal values. Consequently, when ties did occur, they were ranked according to their temporal order.

The previously described overlapping approach for the ordinal mapping considers every consecutive D elements of the time series for encoding into permutation patterns. Due to this overlap, successive ordinal patterns are not independent, and a spurious clustering effect unavoidably appears. For example, with $D = 3$, the fully increasing ordinal pattern $(1,2,3)$ can only be followed by the following three ordinal motifs: $(1,2,3)$, $(1,3,2)$ or $(2,3,1)$, since the first two elements of the vector should respect the increasing order. A non-overlapping scheme is an alternative to avoid this spurious dependency, but with the tradeoff of reducing the amount of ordinal patterns by a factor approximately equal to D. It has been previously shown that the overlapping scheme yields a lower permutation entropy variance for a given time-series length, but at the same time with a higher uncertainty in the variance estimates[20]. The maximum amount of overlap between successive ordinal patterns also turns out to be the optimal option within the ordinal partition network paradigm[21]. In the following, we will compare the pros and cons of both setups regarding the pJSd.

C. The permutation Jensen-Shannon divergence

The pJSd is a statistic used to compare the ordinal mapping between two time series[11]. It is defined as the classical JSd but is estimated employing the ordinal pattern distributions associated with the two time series under analysis, i.e., $D_{JS}\left(P_\pi^{D,\tau_1}, Q_\pi^{D,\tau_2}\right)$ with P_π^{D,τ_1} and Q_π^{D,τ_2} the ordinal probability distributions obtained from the two time series. When calculating pJSd, the order D chosen to implement the BP symbolization recipe should be the same for both time series in order to have the same number of possible permutation patterns in the ordinal probability distributions to be compared. However, different lags (τ_1 and τ_2) can potentially be selected. This opens the possibility of contrasting the ordinal mappings of the same time series at two different temporal scales.

In order to robustly conclude that two time series are generated from different systems, it is necessary to show that the pJSd estimated from them is significantly larger than the baseline value obtained when time series from the same system are compared. When ordinal mapping is applied, the number of symbols k in Eq. (2) is at most equal to $D!$. It is worth noting here that the number of symbols could be potentially lower than this value due to finite-size effects (some ordinal patterns do not appear just because the length N is too small; they are also known as missing ordinal patterns[22]) or due to the fact that chaotic systems are analyzed (true forbidden ordinal patterns are a signature of these dynamical regimes[23]). Consequently, Eq. (2) should be applied with caution when handling analysis under these conditions.

Since Eq. (2) is an approximation obtained from symbolic i.i.d. sequences, the fundamental question is whether the ordinal mapping generates i.i.d. symbols, independent of the dynamical system that produced the original time series. As a first approach to answering this question, in the next section, ordinal symbolic sequences obtained from several stochastic and chaotic processes have been tested.

III. NUMERICAL ANALYSIS

The following stochastic and chaotic systems are tested: white Gaussian noise, ordinary Brownian motion, fractional Gaussian noise (fGn) and fractional Brownian motion (fBm) with Hurst exponent $H = 0.75$, and logistic map with bifurcation parameter $r = 4$, i.e., in the fully chaotic regime. Further details about these systems can be found in the Appendix. One-hundred synthetic symbolic realizations of length N were obtained from each tested system, and then the mean value of the JSd, $\langle D_{JS}\rangle$, between the first half and the second half of these ordinal sequences, was estimated. This averaged quantity is plotted as a function of N following Fig. 2 in Ref. 10. Ten different lengths N between 10^3 and 10^6, equidistant on the logarithmic scale, have been considered on the horizontal axis, and the theoretical reference, Eq. (2), is displayed in dashed lines. Orders D between 3 and 6, and lag $\tau = 1$ were implemented to generate the ordinal symbolic sequences. It is important to remark once again that for each model, results obtained for overlapping and non-overlapping ordinal sequences were included to contrast them. It is clear that temporal correlations between symbols are unavoidably present in the overlapping sequences due to the recipe implemented to define consecutive ordinal patterns. On the contrary, these spurious correlations are avoided when a non-overlapping scheme is applied.

Figures 1 and 2 show the results obtained for white Gaussian noises when implementing overlapping and non-overlapping setups, respectively.

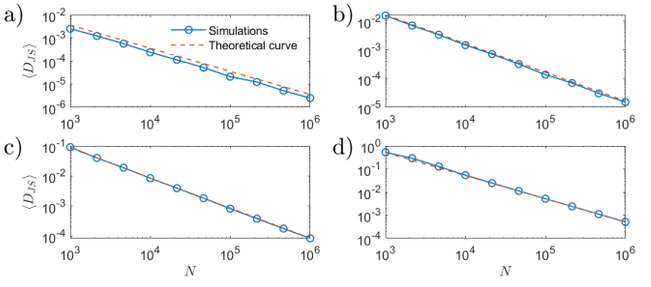


FIG. 1. Mean value of the pJSd, $\langle D_{JS}\rangle$, for 100 independent ordinal sequences of length N obtained from synthetic realizations of normally distributed pseudorandom numbers. JSd between the first half and the second half of these symbolic sequences has been estimated. Overlapping ordinal symbolic sequences with orders D between 3 and 6, increasing from a) to d), and lag $\tau = 1$ are analyzed. Dashed lines show the analytical approximations, Eq. (2), adapted to each case, i.e., $k = D!$.

By contrasting the results obtained in these two figures, it can be concluded that worse results are obtained when overlapping ordinal sequences are implemented, especially for

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

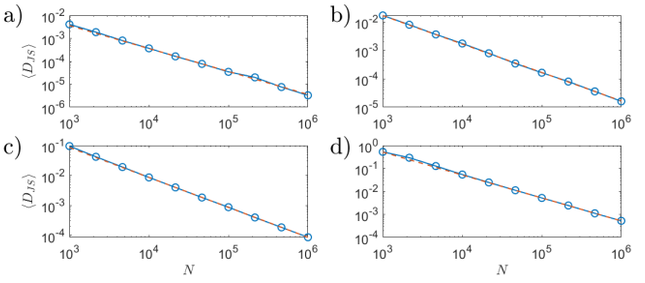


FIG. 2. The same as in Fig. 1 but when non-overlapping ordinal symbolic sequences are studied.

$D = 3$. This is due to the spurious temporal correlations introduced by the overlapping between consecutive ordinal patterns. Thus, the mapped ordinal symbols are not i.i.d., and a shift is observed between the analytical approximation, Eq. (2), and the numerical results. This shift is smaller for larger orders D. Actually, the results for both schemes are analogous for $D \geq 4$. Consequently, a higher number of ordinal symbols seems to mitigate this unwanted effect.

The results obtained for the ordinary Brownian motion are shown in Figs. 3 and 4. The numerical realizations of this fractal stochastic process are obtained as cumulative sums of normally distributed pseudorandom numbers. We have confirmed that the same behavior is observed when the MATLAB code *wfbm* with Hurst exponent $H = 0.5$ is implemented to generate the simulations.

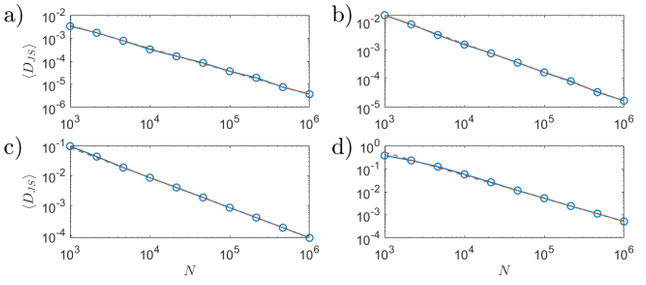


FIG. 3. The same as in Fig. 1 but when overlapping ordinal symbolic mappings from ordinary Brownian motions are studied. The stochastic simulations were obtained as cumulative sum of normally distributed pseudorandom numbers.

Curiously, when ordinary Brownian motions are tested, numerical results for both overlapping and non-overlapping sequences agree very well with the analytical expressions, even for the smaller order $D = 3$. This seems to suggest that this particular stochastic model is immune to the spurious temporal correlations introduced by the overlapping scheme. We will dive deeper into the reasons behind this specific finding in the next section.

It is well-known that stochastic processes with long-range dependence are ubiquitous mathematical models for a large number of complex systems[24]. Consequently, characterizing the power-law decay of their temporal correlations through the analysis of associated time series is a relevant challenge that is often difficult to address[25,26]. In the following, we analyze the performance of the analytical approximation in these cases.

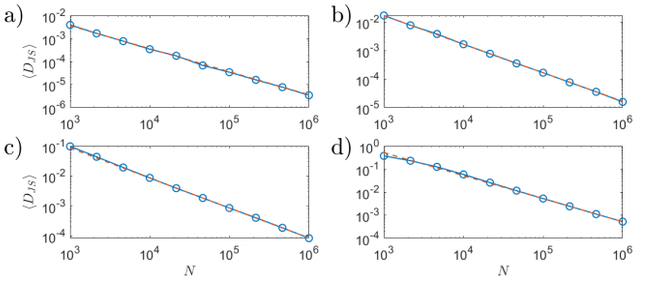


FIG. 4. The same as in Fig. 3 but when non-overlapping ordinal symbolic sequences are studied.

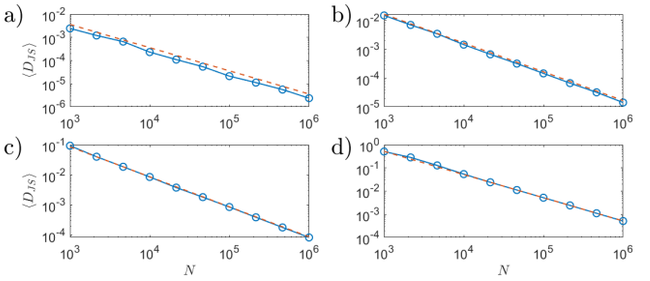


FIG. 5. The same as in Fig. 1 but when overlapping ordinal symbolic mappings from fGn with Hurst exponent $H = 0.75$ are studied. The stochastic simulations were obtained as approximate derivatives of fBm realizations with the same Hurst exponent H (obtained by implementing the MATLAB code *wfbm*).

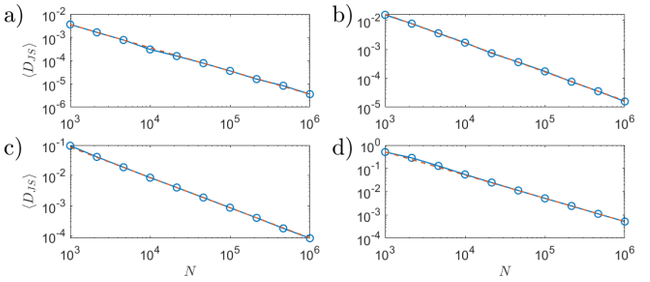


FIG. 6. The same as in Fig. 5 but when non-overlapping ordinal symbolic sequences are studied.

In Figs. 5 and 6 the results obtained for numerical realizations of fGn with Hurst exponent $H = 0.75$ are plotted for the overlapping and non-overlapping scheme, respectively. Analogously, Figs. 7 and 8 show the results obtained when simulations from fBm with the same H are tested. The behavior found for fGn are very similar to those observed for white Gaussian noises. The numerically estimated reference values are slightly below the analytical curve when an overlapping configuration is applied for the ordinal mapping, mainly for the lowest order D.

The results obtained for fBm simulations (Figs. 7 and 8)

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

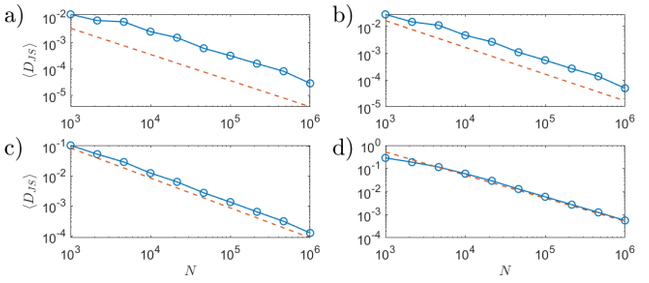


FIG. 7. The same as in Fig. 1 but when overlapping ordinal symbolic mappings from fBm with Hurst exponent $H = 0.75$ are studied. The stochastic simulations were obtained by using the MATLAB code *wfbm* with $H = 0.75$.

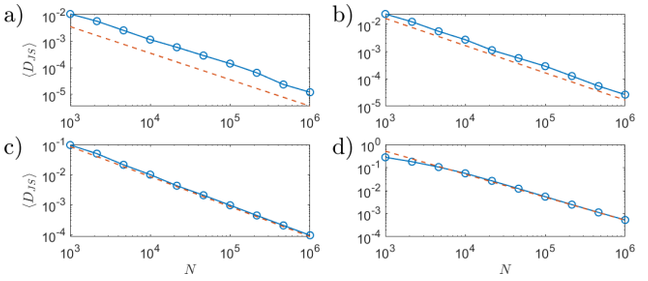


FIG. 8. The same as in Fig. 7 but when non-overlapping ordinal symbolic sequences are studied.

show clear discrepancies with the analytical curve for both, overlapping and non-overlapping cases. The numerically estimated values are higher than those expected according to the analytical expression and the differences decrease as the order D increases. Once again, increasing the number of symbolic elements helps reduce the bias. We conjecture that the mapped ordinal symbolic sequences obtained from this fractal stochastic model with long-range correlations keep part of the underlying memory. Thus, the required hypothesis, i.e., to have i.i.d. symbolic sequences, is not satisfied.

Since chaotic dynamical regimes are also exhibited by many complex systems, we propose to analyze a paradigmatic example of chaotic dynamics: the logistic map (see Appendix). The results obtained for this discrete dynamical system with bifurcation parameter $r = 4$ (chaotic regime) are summarized in Figs. 9 and 10. Large biases are observed between numerical estimations and analytical curves. In this case, the numerical estimations are always lower than the expected analytical approximation. Moreover, the differences increase with the order D.

Since the existence of forbidden ordinal patterns is a signature of chaotic dynamics[23], we hypothesize that it is necessary to take them into account in the analytical expression. That is, the number of symbols k in Eq. (2) should be replaced by the cardinality of the permutation set, $\mathrm{card}(\Pi_D)$. For the logistic map: $\mathrm{card}(\Pi_3) = 5$, $\mathrm{card}(\Pi_4) = 12$, $\mathrm{card}(\Pi_5) = 31$ and $\mathrm{card}(\Pi_6) = 75$[27]. As can be concluded from Figs. 11 and 12, the agreement between the numerical and analytical results is noticeably better with this adjustment. This finding illustrates

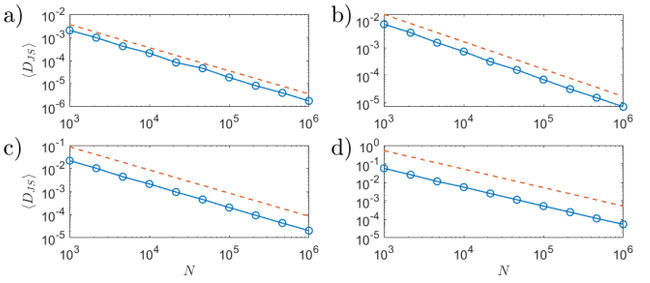


FIG. 9. The same as in Fig. 1 but when overlapping ordinal symbolic mappings from numerical realizations of the logistic map with parameter $r = 4$ are studied.

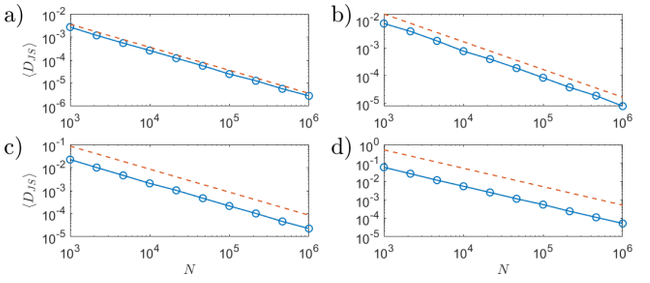


FIG. 10. The same as in Fig. 9 but when non-overlapping ordinal symbolic sequences are studied.

the importance of considering the forbidden ordinal patterns when setting Eq. (2).

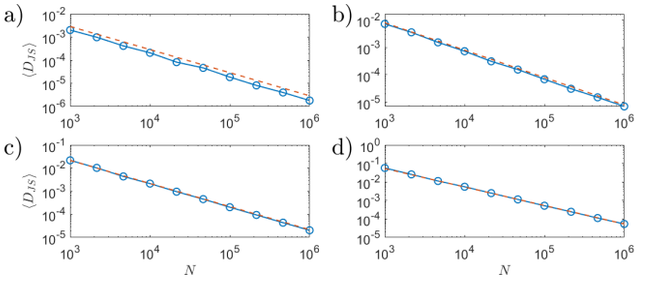


FIG. 11. The same as in Fig. 9 but now the analytical approximations, Eq. (2), were modified taking into consideration the existence of forbidden ordinal patterns. Thus, the number of symbols k for D between 3 and 6 are 5, 12, 31 and 75, respectively.

Finally, in Figs. 13 and 14, binary long-range correlated and Zipf i.i.d. symbolic sequences were tested, respectively. It should be stressed that they are not ordinal symbolic sequences and that they are just included for the sake of comparison. In the former case, sign series[28] obtained from fGn simulations with Hurst exponents $H \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ were generated while pseudorandom numbers that follow a Zipf distribution with k symbols were simulated in the latter case. The Zipf distribution follows a power law $1/f^{\alpha}$ with exponent $\alpha = 1$.

On the one hand, Fig. 13 confirms that the reference values differ significantly from the approximated theoretical expres-

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

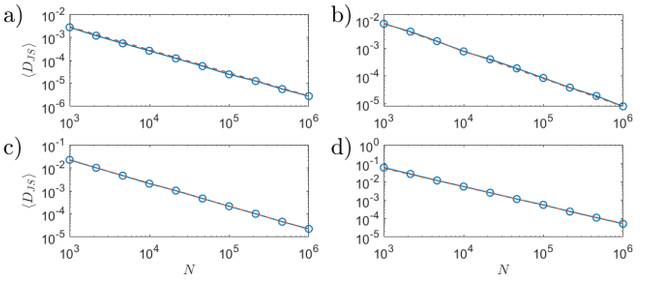


FIG. 12. The same as in Fig. 11 but when non-overlapping ordinal symbolic sequences are studied.

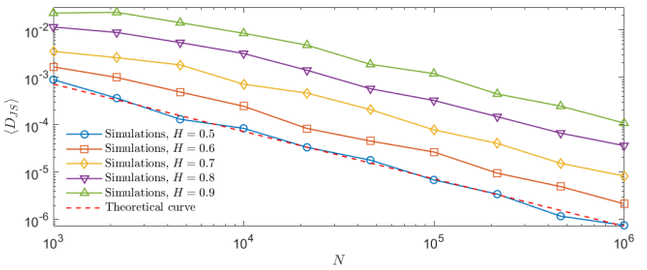


FIG. 13. Mean value of the JSd, $\langle D_{JS}\rangle$, for 100 independent correlated binary sequences with different Hurst exponent H of length N. JSd between the first half and the second half of these binary sequences has been estimated. These are not ordinal symbolic sequences and dashed line shows the analytical approximation, Eq. (2), with $k = 2$.

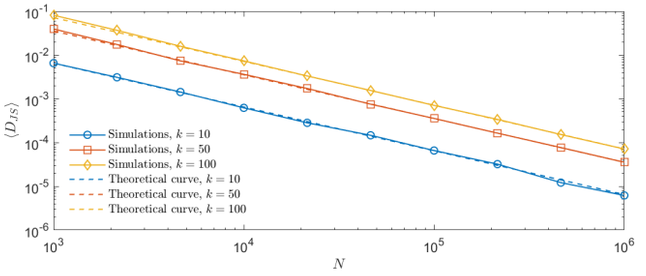


FIG. 14. The same as in Fig. 13 but when Zipf i.i.d. symbolic sequences with k symbols are tested. These are not ordinal symbolic sequences and the dashed lines show the analytical approximations, Eq. (2), adapted to the number of symbols considered in each case.

sion when symbolic sequences with long-range memory effects ($H > 0.5$) are considered. The same dependence with N is roughly observed but with a different constant factor. On the other hand, Fig. 14 illustrates that this approximation works well for pseudorandom symbolic sequences that follow other, non-Gaussian, amplitude distributions.

Visual inspection of Fig. 13 reveals that all curves become parallel in the asymptotic, large-N regime ($N \geq 10^4$). This parallel alignment suggests that long-range memory introduces a multiplicative constant factor, $C(H)$, into Eq. (2) without altering the scaling exponent. Moreover, the vertical shift grows with H, indicating that $C(H)$ is an increasing function of the Hurst exponent in the persistent regime ($H > 1/2$). This interesting finding paves the way to look for numerical generalization that explicitly accounts for memory effects. However, this analysis is beyond the scope of the present work, and we plan to systematically explore this extension in future research.

By contrasting the numerical results across the considered stochastic and chaotic systems—from Fig. 1 to Fig. 14—it can be concluded that empirical estimations systematically deviate from the theoretical baseline in highly structured ways. Specifically, an overestimation is observed when non-trivial temporal correlations or long-range memories are present (Figs. 7, 8 and 13) whereas an underestimation is characteristically associated with overlapping sampling schemes or the structural constraints of forbidden patterns (Figs. 1, 5, 9, 10 and 11) Consequently, preserving and monitoring the sign of this bias serves as a powerful diagnostic tool, yielding direct insights into the underlying physical mechanisms generating the data.

As a last remark, we note that some discrepancies, much less noticeable, between numerical results and the theoretical curves are observed for realizations with the shortest length ($N = 10^3$ data), and when the largest order ($D = 6$) is implemented for generating the ordinal symbolic sequences — see Figs. 3, 4, 7 and 8. These discrepancies can be attributed to finite-size effects. Some ordinal patterns do not appear, i.e., there are missing ordinal patterns[22], just because the time series is too short and the cardinality of the permutation set is too large. Being more precise in mathematical terms, the rule of thumb $N \gg D!$ is not accomplished in these particular instances.

IV. OVERLAPPING VERSUS NON-OVERLAPPING ORDINAL MAPPING SCHEMES: A MORE IN-DEPTH ANALYSIS

In this second part of our research, we try to understand why estimated values are systematically lower than those expected according to Eq. (2) when the overlapping ordinal scheme is implemented (see Figs. 1, 5 and 11). As this effect is more evident for smaller orders, in the following, we focus our analysis on $D = 3$ and $D = 4$. We estimate the relative frequencies of the associated ordinal patterns for 1,000 independent realizations of Gaussian white noise for both schemes to contrast them. It is very important to note here that the length of the symbolic sequences N is equal to 10^4 for both configurations and, consequently, the length of the original time series before being symbolized has to be D times larger in the non-overlapping approach. The mean and standard deviation (as error bars) of the empirical ordinal pattern probabilities (EOPP) for both configurations, overlapping and non-overlapping, and for both orders, $D = 3$ and $D = 4$, are depicted in Fig. 15. The ordinal patterns are arranged and indexed according to their lexicographical order, as shown in Fig. 2 of Ref. 29. Just for illustrative purposes, $\Pi_3 = \{\pi_1 = (1,2,3), \pi_2 = (1,3,2), \pi_3 = (2,1,3), \pi_4 = (2,3,1), \pi_5 = (3,1,2), \pi_6 = (3,2,1)\}$.

It can be concluded that the mean values for the EOPP are

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

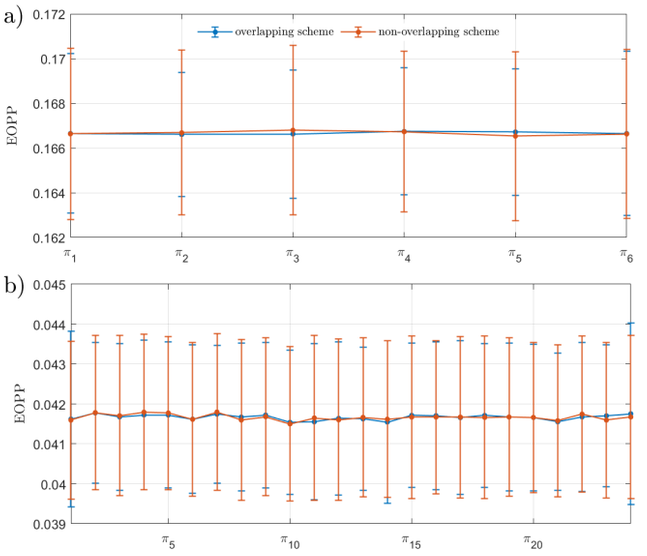


FIG. 15. Mean and standard deviation (as error bars) of the EOPP for 1,000 independent realizations of Gaussian white noise for both ordinal mapping configurations, overlapping and non-overlapping, and for orders a) $D = 3$ and b) $D = 4$. The length of the symbolic sequences N is equal to 10^4 for both schemes and the ordinal patterns are indexed according to their lexicographical order.

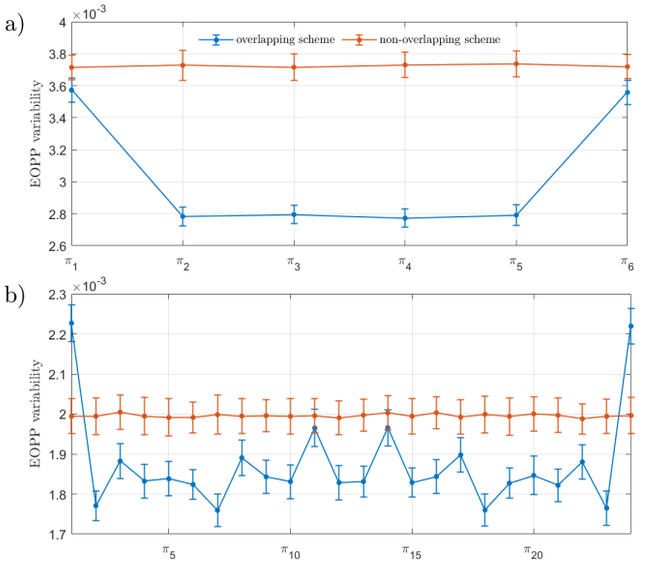


FIG. 16. Mean and standard deviation (as error bars) of the variability of the EOPP for 100 repetitions of the analysis for Gaussian white noises presented in Fig. 15. A comparison between the overlapping and non-overlapping ordinal mapping for orders a) $D = 3$ and b) $D = 4$ is included. The length of the symbolic sequences N is equal to 10^4 for both schemes and the ordinal patterns are indexed according to their lexicographical order.

similar, close to 1/6 for $D = 3$ and to 1/24 for $D = 4$, as expected, for both schemes. However, the standard deviations do not coincide: most of them are lower in the overlapping case, with the only exception of the ones obtained for π_1 and π_{24} for $D = 4$. Moreover, in the overlapping implementation, the dispersions are not homogeneous, with larger values for the monotone ordinal patterns than for the non-monotone ones. A similar but less noticeable result (not shown here for the sake of space and brevity) is observed for larger orders D ($D = 5$ and $D = 6$). We repeated the same analysis 100 times, estimating then 100 independent values for the standard deviation of the empirical probability for each ordinal pattern. Figure 16 shows the mean and standard deviation of the variability of the EOPP for $D = 3$ and $D = 4$ for both configurations. Significant differences are observed: while variability estimations are similar for all ordinal patterns for the non-overlapping setup, they clearly depend on the ordinal pattern when an overlapping scheme is implemented, with monotone ordinal patterns (π_1 and π_6 for $D = 3$, and π_1 and π_{24} for $D = 4$) always having the largest values.

To better contrast the observed differences for the EOPP as a function of the order D, we estimate the global variability Σ_D as the mean of the standard deviation (SD) mean values obtained for the relative frequencies of the ordinal patterns

$$\Sigma_D = \frac{\sum_{i=1}^{D!} \langle \mathrm{SD}(EOPP) \rangle}{D!}, \tag{4}$$

for the two different schemes. The results obtained are shown in Fig. 17. The global variability Σ_D is smaller in the overlapping implementation, especially for small orders D ($D = 3$ and $D = 4$). This reduced dispersion leads to the lower than expected values estimated for the $\langle D_{JS} \rangle$ in the previous overlapping analyses (Figs. 1, 5 and 11). Moreover, this bias decreases as D increases, in line with the findings for Σ_D.

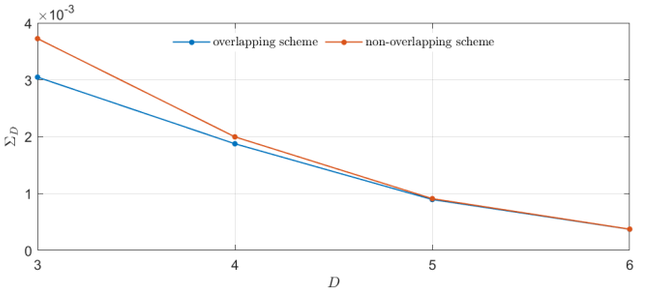


FIG. 17. The global variability Σ_D, Eq. (4), as a function of the order D when numerical realizations of Gaussian white noise are analyzed.

Interestingly, this bias is not observed when the ordinary Brownian motion was tested (Fig. 3). We have performed the same analysis trying to understand the reason behind this curious result, and 1,000 independent symbolic sequences of length N equal to 10^4 data points obtained from ordinary Brownian motion realizations were tested for both configurations, overlapping and non-overlapping, and for orders $D = 3$ and $D = 4$. Figures 18 and 19 show the mean and standard deviation (as error bars) of the EOPP and their variability, respectively, while Fig. 20 shows the results obtained for the global variability (Eq. (4)). In this case, Σ_D is similar for both schemes for each order D. Essentially, we found that

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

the higher variability of strictly monotone ordinal patterns in the overlapping setup compensates for the lower variability of non-monotone ones. We consider this finding particularly useful, especially from a practical perspective, since the null hypothesis can be checked by implementing an overlapping scheme on much smaller time series and without the presence of a significant bias.

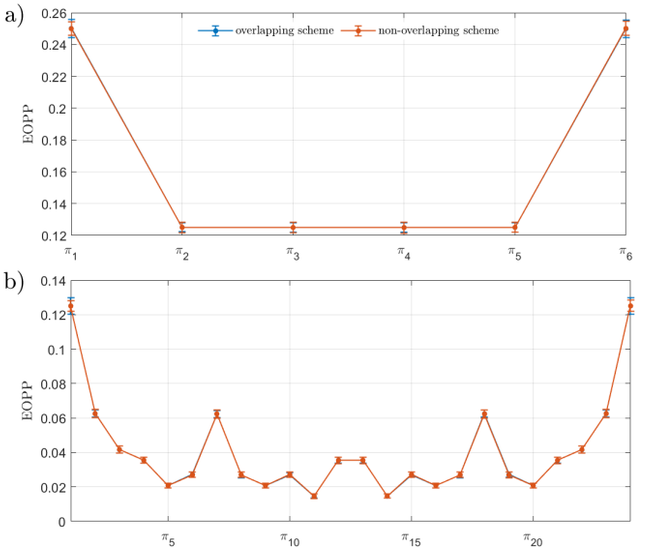


FIG. 18. The same as in Fig. 15 but for 1,000 independent realizations of ordinary Brownian motion.

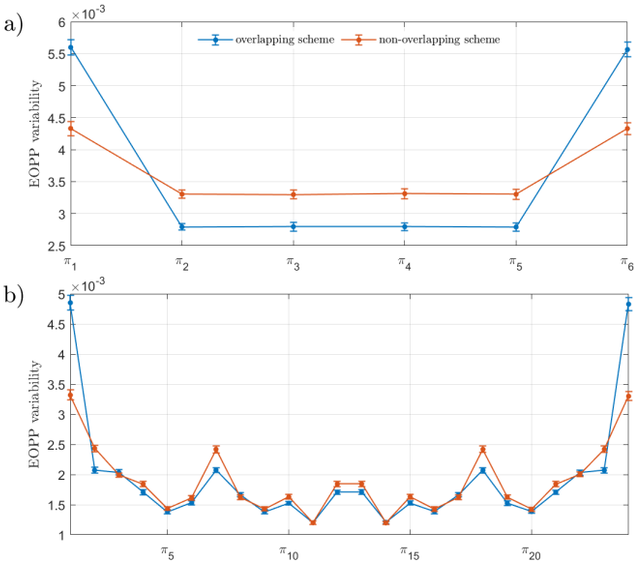


FIG. 19. The same as in Fig. 16 but here the variability of the EOPP for ordinary Brownian motions are analyzed.

Taking advantage of this fact, in the following, we check the robustness of this approach to identify the presence of weak long-range correlations, i.e. Hurst exponent H slightly larger than 0.5 that confirm dependencies between distant elements and, consequently, non-randomness, when time series of reduced length L are available. With that goal in mind, 1,000 independent numerical realizations of fGn and fBm were generated with $L = 1,000$ data points and Hurst exponents $H \in \{0.50, 0.52, \ldots, 0.7\}$. In Figs. 21 and 22, $\langle D_{JS} \rangle$ is plotted as a function of H when persistent dynamics are contrasted against fully random counterparts. References are white noises and ordinary Brownian motions in the fGn and fBm analyses, respectively. Both schemes, overlapping and non-overlapping, with orders $D = 3$ and $D = 4$ were implemented. After comparing the results obtained, it is concluded that applying the pJSd directly on the process (fBm) by using an overlapping scheme seems to be the optimal framework since the best discrimination is achieved. In particular, for $D = 4$, fBm with Hurst exponent $H = 0.6$ is robustly distinguished from the ordinary Brownian motion with $H = 0.5$, which represents the random reference in this case.

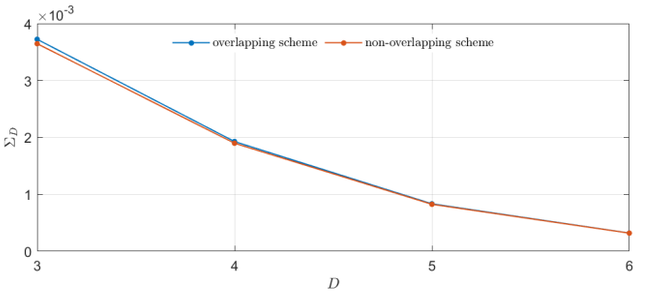


FIG. 20. The same as in Fig. 17 but when numerical realizations of ordinary Brownian motion are analyzed.

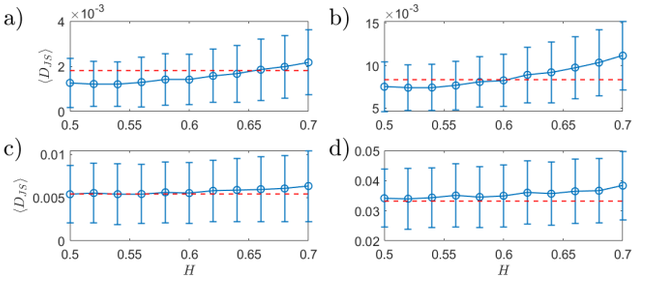


FIG. 21. Mean and standard deviation (as error bars) of the pJSd, $\langle D_{JS} \rangle$, between numerical realizations of fGn with Hurst exponent $H \in \{0.50, 0.52, \ldots, 0.7\}$ and their fully random references, i.e., white noises of the same length L. A total of 1,000 independent numerical realizations with $L = 1,000$ data points were generated for each H. Results obtained for overlapping (top plots a) and b)) and non-overlapping schemes (bottom plots c) and d)), and $D = 3$ (left plots a) and c)) and $D = 4$ (right plots b) and d)) are included. Dashed lines show the analytical value obtained from Eq. (2) adapted to each case.

We consider this finding relevant enough because the identification of subtle long-range temporal correlations from small time series represents a crucial challenge in several scientific fields. It is clear that this information is essential for modeling and forecasting purposes. For example, the presence of this memory effect in financial time series confirms

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

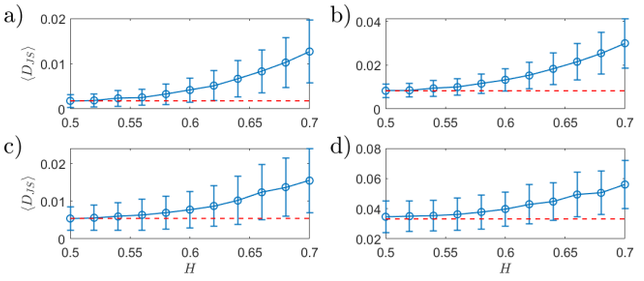


FIG. 22. The same as in Fig. 21 but for fBm. In this case the fully random references are ordinary Brownian motion realizations.

deviations from the Efficiency Market Hypothesis (EMH) that is ideally assumed for them[30]. Identifying periods of higher or lower efficiency over time can help to understand which factors govern the performance of stock market dynamics. Next, we will validate the practical utility of this approach by applying it to a significant financial time series: the Bitcoin hourly price. Finally, it is worth noting that the improved performance of ordinal analysis on the process (fBm) compared to noise (fGn) to detect long-range dependencies aligns with previous findings[31].

V. A FINANCIAL REAL-WORLD ILLUSTRATIVE APPLICATION

To demonstrate the practical utility of the previously introduced ordinal approach, we apply it to characterize the informational efficiency of Bitcoin (BTC). Historical hourly price data were retrieved from the CryptoCompare API via the v2/histohour endpoint[32]. The resulting dataset, spanning from January 1, 2018, to November 11, 2024, consists of 60,145 observations of hourly closing prices. The length and granularity of this time series provide the high-frequency resolution necessary to investigate the complex dynamical phenomena examined in this study. Moreover, the analysis was conducted directly on prices taking into consideration that the ordinary Brownian motion is a widely accepted basic mathematical model for financial price data that satisfy the EMH[33].

A sliding window approach has been implemented with rolling windows of size $L = 1,000$ data points and a step equal to one data point. This dynamical analysis offers the possibility to unveil the presence of hidden temporal correlations and, consequently, signatures of market inefficiencies, during specific time periods within the financial record.

Figure 23 shows the mean and standard deviation (as error bar) of the pJSd of each rolling window with respect to 100 independent numerical realizations of the ordinary Brownian motion, which is the efficiency benchmark in this case, for $D = 4$ and an overlapping scheme. We have confirmed that this is the optimal setup since results for $D = 3$ and/or a non-overlapping scheme (not shown for the sake of space) are not so clear. The results obtained for each year were included separately for better visualization. The dates detailed on the x axis correspond to the last date of each rolling window. It can be concluded that the distance to the null hypothesis evolves dynamically, with certain time periods in which significant statistical deviations from the EMH are robustly identified. The number of rolling windows in which the null hypothesis can be rejected decreases with time. This trend toward greater consistency with the EMH can be interpreted as a proof of the evolution of the BTC market to maturity, which is in line with the findings of previous investigations[34–36]. Although linking inefficient time periods with historical events could yield valuable insights, this analysis should be done carefully, and it is beyond the scope of the present work. We believe that the included results confirm the robustness of the proposed approach and successfully illustrate its implementation in a practical context.

VI. CONCLUSIONS

Discriminating whether two time series originate from the same system and/or have the same dynamical properties is essential for a better understanding and classification of them. The pJSd has recently been proposed for such a purpose. In order to reach reliable conclusions within this framework, it is necessary to characterize the robustness and finite-size bias of its estimated values. In this work, we try to fill this gap on the basis of heuristic analyses, and some useful conclusions have been obtained for a robust implementation of the pJSd in practical applications. Our more important findings are summarized below.

We have compared the performances of the overlapping and non-overlapping ordinal schemes when mapping the original time. It is clear that the overlapping approach introduces spurious temporal correlations, that unavoidably generate a bias with the analytical expression. However, we have found that this bias is reduced by increasing the number of symbolic elements, i.e., by using larger orders D. The observed discrepancies are more noticeable in the case of fBm with long-range correlations. We conjecture that the mapped ordinal symbolic sequences obtained from this fractal stochastic model are not i.i.d, keeping some of the underlying memory. On the other hand, there is an excellent agreement between numerical results and analytical approximation in the case of ordinary Brownian motion (memoryless) in both, overlapping and non-overlapping, setups.

When dealing with chaotic systems, it is necessary to take into account the presence of forbidden ordinal patterns associated with them. The number of admissible ordinal motifs should be considered when setting the number of symbols k in Eq. (2). We confirmed that for symbolic sequences with long-range correlations, the estimated numerical values are significantly higher than the analytical approximation. Conversely, agreement is excellent for pseudorandom sequences following non-Gaussian amplitude distributions.

In-depth investigation confirms that the remarkable agreement between the numerical results and the analytical approximation for ordinary Brownian motion is attributable to EOPP variability. Based on this finding, we then confirm that applying an overlapping ordinal scheme directly to the process

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

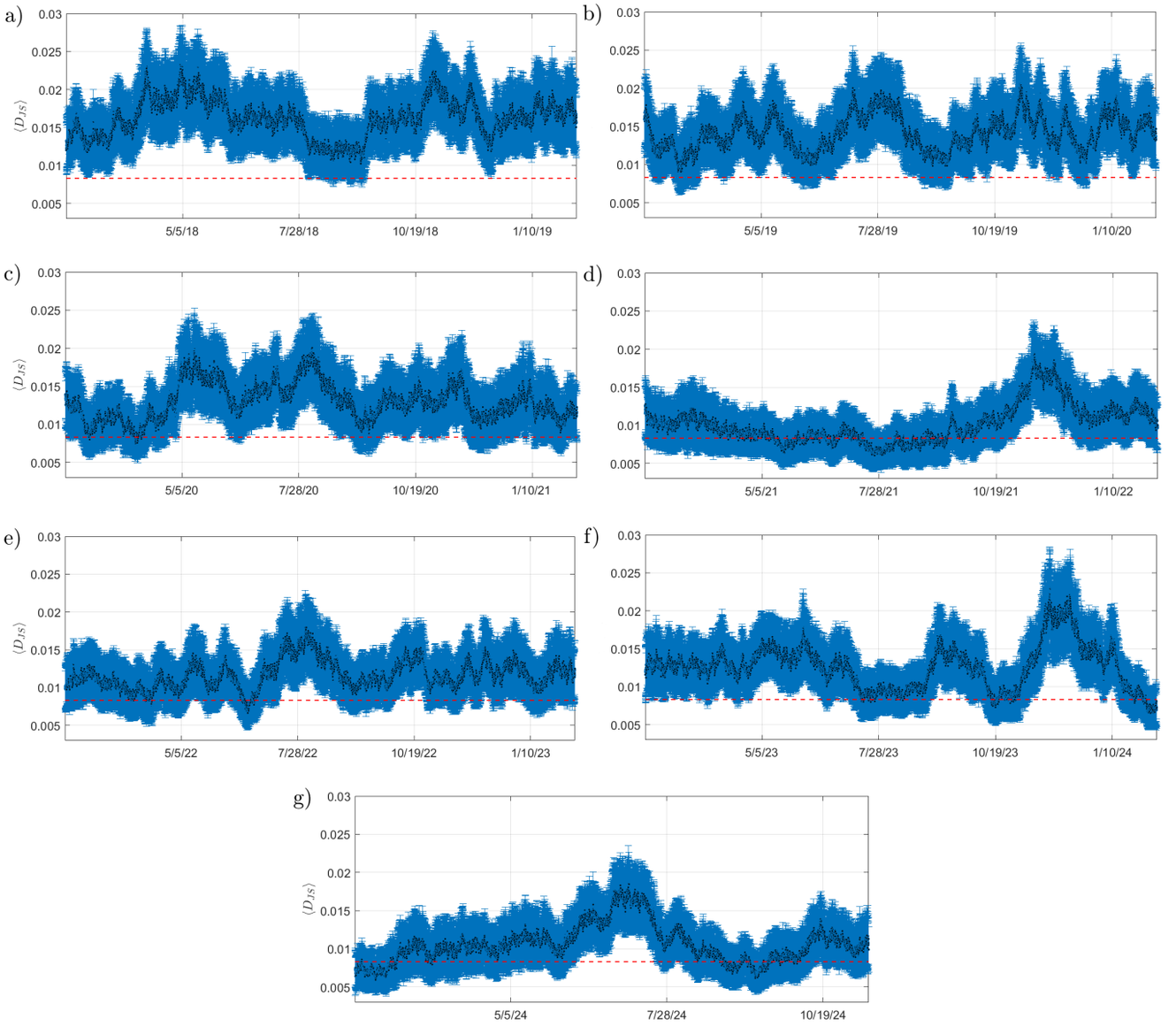


FIG. 23. Dynamical analysis of BTC hourly prices from January 2018 to November 2024. Mean and standard deviation (as error bar) of the pJSd of each rolling window with respect to 100 independent numerical realizations of the ordinary Brownian motion estimated by using $D = 4$ and an overlapping scheme are plotted for years: a) 2018, b) 2019, c) 2020, d) 2021, e) 2022, f) 2023, and g) 2024. Windows of size $L = 1{,}000$ data points and maximum overlapping between windows, i.e. step equal to one data point, have been implemented. Dashed line shows the analytical value obtained from Eq. (2) adapted to this case, i.e., k = 4! and $N = 2 \times (L - D + 1) = 1994$.

(fBm) is the best alternative for unveiling the presence of subtle long-range temporal correlations from small time series. We demonstrate the practical utility of this framework by applying it to evaluate the efficiency of a real-world financial time series.

The behavior observed for binary long-range correlated sequences (Fig. 13) suggests that the presence of memory modifies the analytical equation obtained for uncorrelated symbolic sequences by introducing a multiplicative constant factor, while maintaining the same dependence on the symbolic sequence length N. Comprehensive numerical and analytical investigations will be pursued in future work to formally validate this hypothesis.

ACKNOWLEDGMENTS

L.Z. and P.R.S. acknowledge the financial support from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Appendix: Stochastic and chaotic systems

The different stochastic and chaotic systems considered along this work are detailed below.

- fractional Gaussian noise (fGn): proposed by Mandelbrot and Van Ness[37], it is a family of stationary Gaussian stochastic processes characterized by the Hurst exponent H, a real parameter strictly bounded within the interval $(0,1)$. When $H = 1/2$, the process reduces to the standard, uncorrelated Gaussian white noise. Conversely, the correlations are negative (anti-persistence) for $H < 1/2$ and positive (persistence) for $H > 1/2$. Moreover, these correlations are summable in the former case and not summable in the latter one. Due to this divergence property, fGn with $H \in (1/2,1)$ is often referred to as long-range correlated. fGn has a power spectrum of the form $S(f) \sim 1/f^{\beta}$ with the exponent $\beta = 2H-1$ ($-1 < \beta < 1$).

- fractional Brownian motion (fBm): it is a family of stochastic processes that are Gaussian, self-similar, and endowed with stationary increments. Actually, fGn is the stationary increment process of the non-stationary fBm. The same Hurst parameter is shared by them, and when $H = 1/2$ the classical Brownian motion is recovered from the fBm family. Because fBm is non-stationary, its traditional Fourier transform does not exist. However, a generalized power spectrum $S(f) \sim 1/f^{\beta}$ with exponent $\beta = 2H+1$ ($1 < \beta < 3$) can be associated with it. It is worth remarking that both stochastic processes, fGn and fBm, have been widely accepted to represent and analyze real-world data that exhibits fractal properties and power-law correlations[24].

- Logistic map: defined by the quadratic recurrence equation $x_{n+1} = r x_n(1-x_n)$, it represents a prototypical example of how simple nonlinear equations can generate complex phenomena[38]. In particular, a chaotic dynamics with Lyapunov exponent $\lambda = \ln 2$ is obtained when the bifurcation parameter $r = 4$.

[1] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," Phys. Rev. Lett. **88**, 174102 (2002).
[2] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: A review," Entropy **14**, 1553–1577 (2012).
[3] J. M. Amigó, K. Keller, and J. Kurths, "Recent progress in symbolic dynamics and permutation complexity—Ten years of permutation entropy," Eur. Phys. J. Spec. Top. **222**, 241–247 (2013).
[4] J. M. Amigó, K. Keller, and V. A. Unakafova, "Ordinal symbolic analysis and its application to biomedical recordings," Phil. Trans. R. Soc. A **373**, 20140091 (2015).
[5] M. Zanin and F. Olivares, "Ordinal patterns-based methodologies for distinguishing chaos from noise in discrete time series," Commun. Phys. **4**, 190 (2021).
[6] I. Leyva, J. H. Martínez, C. Masoller, O. A. Rosso, and M. Zanin, "20 years of ordinal patterns: Perspectives and challenges," Europhysics Letters **138**, 31001 (2022).
[7] J. M. Amigó and O. A. Rosso, "Ordinal methods: Concepts, applications, new developments, and challenges—In memory of Karsten Keller (1961–2022)," Chaos **33**, 080401 (2023).
[8] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska, "Surrogate data for hypothesis testing of physical systems," Physics Reports **748**, 1–60 (2018).
[9] J. Lin, "Divergence measures based on the Shannon entropy," IEEE Trans. Inf. Theory **37**, 145–151 (1991).
[10] 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 **65**, 041905 (2002).
[11] L. Zunino, F. Olivares, H. V. Ribeiro, and O. A. Rosso, "Permutation Jensen-Shannon distance: A versatile and fast symbolic tool for complex time-series analysis," Phys. Rev. E **105**, 045310 (2022).
[12] A. C.-C. Yang, S.-S. Hseu, H.-W. Yien, A. L. Goldberger, and C.-K. Peng, "Linguistic analysis of the human heartbeat using frequency and rank order statistics," Phys. Rev. Lett. **90**, 108103 (2003).
[13] D. M. Mateos, L. E. Riveaud, and P. W. Lamberti, "Detecting dynamical changes in time series by using the Jensen Shannon divergence," Chaos **27**, 083118 (2017).
[14] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, "Compositional segmentation and long-range fractal correlations in DNA sequences," Phys. Rev. E **53**, 5181 (1996).
[15] J. Sparacino, P. W. Lamberti, and C. M. Arizmendi, "Shock detection in dynamics of single-headed motor proteins KIF1A via Jensen-Shannon divergence," Phys. Rev. E **84**, 041907 (2011).
[16] M. Gerlach, F. Font-Clos, and E. G. Altmann, "Similarity of symbol frequency distributions with heavy tails," Phys. Rev. X **6**, 021009 (2016).
[17] F. Olivares and L. Zunino, "Multiscale dynamics under the lens of permutation entropy," Physica A: Statistical Mechanics and its Applications **559**, 125081 (2020).
[18] L. Zunino, M. C. Soriano, and O. A. Rosso, "Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach," Phys. Rev. E **86**, 046210 (2012).
[19] L. Zunino, F. Olivares, F. Scholkmann, and O. A. Rosso, "Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions," Physics Letters A **381**, 1883–1892 (2017).
[20] D. J. Little and D. M. Kane, "Variance of permutation entropy and the influence of ordinal pattern selection," Phys. Rev. E **95**, 052126 (2017).
[21] K. Sakellariou, T. Stemler, and M. Small, "Markov modeling via ordinal partitions: An alternative paradigm for network-based time-series analysis," Phys. Rev. E **100**, 062307 (2019).
[22] L. C. Carpi, P. M. Saco, and O. Rosso, "Missing ordinal patterns in correlated noises," Physica A: Statistical Mechanics and its Applications **389**, 2020–2029 (2010).
[23] J. M. Amigó, S. Zambrano, and M. A. F. Sanjuán, "True and false forbidden patterns in deterministic and random dynamics," Europhysics Letters **79**, 50001 (2007).
[24] J. Beran, *Statistics for long memory processes*, Monographs on statistics and applied probability; 61 (Chapman & Hall, New York, 1994).
[25] G. Rangarajan and M. Ding, "Integrated approach to the assessment of long range correlation in time series data," Phys. Rev. E **61**, 4991–5001 (2000).
[26] J. Gao, J. Hu, W.-W. Tung, Y. Cao, N. Sarshar, and V. P. Roychowdhury, "Assessment of long-range correlation in time series: How to avoid pitfalls," Phys. Rev. E **73**, 016117 (2006).
[27] O. A. Rosso, F. Olivares, L. Zunino, L. D. Micco, A. L. L. Aquino, A. Plastino, and H. A. Larrondo, "Characterization of chaotic maps using the permutation Bandt-Pompe probability distribution," The European Physical Journal B **86**, 116 (2013).
[28] M. Gómez-Extremera, P. Carpena, P. C. Ivanov, and P. A. Bernaola-Galván, "Magnitude and sign of long-range correlated time series: Decomposition and surrogate signal generation," Phys. Rev. E **93**, 042201 (2016).
[29] U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths, and N. Wessel, "Classifying cardiac biosignals using ordinal pattern statistics and symbolic

ACCEPTED MANUSCRIPT
Chaos
An Interdisciplinary Journal
of Nonlinear Science
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0341345

dynamics," Comput. Biol. Med. **42**, 319–327 (2012).
[30]E. F. Fama, "Efficient capital markets: A review of theory and empirical work," The Journal of Finance **25**, 383–417 (1970).
[31]D. DeFord and K. Moore, "Random walk null models for time series data," Entropy **19**, 615 (2017).
[32]CryptoCompare, "Cryptocurrency historical time series data via API (v2/histohour)," `https://min-api.cryptocompare.com/data/v2/histohour` (2026), accessed: 2024-11-26.
[33]C. Bandt, "Order patterns, their variation and change points in financial time series and Brownian motion," Statistical Papers **61**, 1565–1588 (2020).
[34]A. Urquhart, "The inefficiency of Bitcoin," Economics Letters **148**, 80–82 (2016).
[35]S. Drożdż, R. Gębarowski, L. Minati, P. Oświęcimka, and M. Wątorek, "Bitcoin market route to maturity? Evidence from return fluctuations, temporal correlations and multiscaling effects," Chaos: An Interdisciplinary Journal of Nonlinear Science **28**, 071101 (2018).
[36]T. Takaishi, "Time-varying properties of asymmetric volatility and multifractality in Bitcoin," PLOS ONE **16**, 1–21 (2021).
[37]B. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian motions, fractional noises and applications," SIAM Rev. **10**, 422–437 (1968).
[38]R. M. May, "Simple mathematical models with very complicated dynamics," Nature **261**, 459–467 (1976).