ORIGINAL ARTICLE

A Test for White Noise in the Entropy-Complexity Plane

Eduarda T. C. Chagas[1] | Marcelo Queiroz[2,3] | Osvaldo A. Rosso[4] | Heitor S. Ramos[1] | Cristopher G. S. Freitas[2] | Leonardo V. Pereira[2] | Alejandro C. Frery[4]

[1]Departamento de Ciência da Computação, Universidade Federal de Minas, Brazil

[2]Laboratório de Computação Científica e Análise Numérica, Universidade Federal de Alagoas, Brazil

[3]Coordenação de Informática, Instituto Federal de Alagoas, Brazil

[4]Instituto de Física, Universidade Federal de Alagoas, Brazil

[5]School of Mathematics and Statistics, Victoria University of Wellington, New Zealand

Correspondence
Eduarda T. C. Chagas, Departamento de Ciência da Computação, Universidade Federal de Minas, Brazil
Email: eduarda.chagas@dcc.ufmg.br

Funding information
CNPq, Fapeal

The Bandt and Pompe methodology has been used successfully in the analysis of time series, by computing Information Theory descriptors of the histogram of ordinal patterns. Such descriptors lie in a 2D manifold: the Entropy-Complexity plane. This article proposes the first approach that uses the entropy-complexity plane to test the white noise hypothesis. Our test is based on true white noise sequences obtained from physical devices. The proposed methodology provides consistent results, managing to assess sequences of true random samples as random (adequate size), rejecting correlated sequences (sound power), and capturing the randomness of generators previously analyzed in the literature.

KEYWORDS
Random number generators, Bandt–Pompe approach, Entropy-complexity plane, Time Series, Information Theory

1 | INTRODUCTION

Time Series carry valuable information about the system which produces the data. Their analysis is usually based on two approaches (Cryer and Chan, 2008): in the (natural) time and transformed domains (for instance, frequency

Abbreviations: RNG, random number generator; PRNG, pseudorandom number generator; TWNRS, true white noise random sequence.

All authors contributed equally.

and wavelet). In the context of time-domain analysis, Bandt and Pompe (2002) proposed a new methodology. Such an approach is non-parametric and based on descriptors of Information Theory. The time series is transformed into ordinal patterns, with which a histogram is formed. The resulting distribution is less sensitive to outliers than the original data, and the histogram does not depend on any model and. Thus, the approach can be applied to a variety of situations.

The Bandt-Pompe methodology and its variants have been used successfully in the analysis of many types of dynamics, receiving so far more than 2.160 citations, according to the Web of Science[1]. We found works using this approach in several areas of scientific knowledge such as, for example: distinguishing noise from chaos (Rosso *et al.*, 2007); the study of electroencephalography signals using wavelet decomposition (Baravalle *et al.*, 2018a,b); description of El Niño/Southern Oscillation during the Holocene (Saco *et al.*, 2010); the characterization of household appliances through their energy consumption (Aquino *et al.*, 2017); detecting and quantifying stochastic and coherence resonance (Rosso and Masoller, 2009a,b); analysis and characterization of economic time series, e.g., stock market, sovereign bonds, credit rating, commodities, and cryptocurrencies (Zunino *et al.*, 2010, 2012a; Bariviera *et al.*, 2013, 2018; de Araujo *et al.*, 2019); online signature classification and verification (Rosso *et al.*, 2016). Schieber *et al.* (2016) verified the effect of attacks on complex networks by the displacement of points in the $H \times C$ plane. Aquino *et al.* (2015) described vehicles' behavior depending on the topology of cities, and Chagas *et al.* (2020) succeeded in expanding the use of such techniques for analyzing textured images corrupted by speckle noise. Bariviera *et al.* (2015) identified spurious interventions in the Libor market using the $H \times C$ plane representation. Echegoyen *et al.* (2020) were able to discriminate between individuals with mild cognitive impairment from those diagnosed with Alzheimer's disease using magnetoencephalography recordings.

With the Bandt and Pompe methodology, a time series is described by a point in a manifold of $\mathbb{R}^2$: the Entropy-Complexity plane $H \times C$. There are two well-known points in this plane: those of white noise and a completely deterministic sequence. Although the boundaries of the $H \times C$ plane are well-defined, its intrinsic topology's complete characterization is an open problem. In particular, the joint distribution of points in this plane under typical time series types, which would serve to build test statistics, is unknown.

The 2.160 citations received by the seminal paper appeared in 780 venues indexed by the Web of Science. Among them, the journals belong to 127 categories, spanning from Multidisciplinary Physics (24 % of the publications) to Zoology (only one of the of citing articles). There are 22 citing articles from journals that belong to the Statistics & Probability category. Five of these articles appeared in *Stochastic Environmental Research and Risk Assessment*, two in the *Journal of Time Series Analysis* and in *Theory and Applications of Time Series Analysis*, and each of the remaining ten appeared in a different journal. Most of these articles relate successful applications of the Bandt and Pompe methodology, except Sinn and Keller (2011) that obtained the sample entropy's properties under zero-mean Gaussian processes. It is also noteworthy that, in this category of publications, Abrams *et al.* (2013) provided a formal and more general proof of the structure of the boundary of the $H \times C$ manifold than that obtained by Martín *et al.* (2006). The lack of attention that the Bandt and Pompe approach has received by the Probability & Statistics community confirms that it is a fertile research avenue waiting to explore.

Several works have used deterministic and pseudorandom sequences to understand the properties of the points they produce in the $H \times C$ plane. Martín *et al.* (2006) analyzed the chaotic logistic map and discussed the boundaries of the $H \times C$ plane. De Micco *et al.* (2009) studied chaotic components in pseudorandom number generators. Ravetti *et al.* (2014) tackled the often hard problem of distinguishing chaos from noise. Zunino *et al.* (2012b) used a multi-scale approach to analyze the interplay between chaotic and stochastic dynamics.

With the knowledge of the expected variability of such points, according to the underlying dynamics, we can

[1] Checked on 23 October 2021

make hypothesis tests for a wide variety of models. Results in this direction can be found in the literature. Larrondo *et al.* (2006) showed that the Entropy-Complexity plane ($H \times C$) is a good indicator of Diehard tests' results on pseudorandom number generators. De Micco *et al.* (2008) assessed ways of improving pseudorandom sequences by their representation in this plane.

Motivated by previous works, in this paper, we advance the state-of-the-art providing the first test for white noise points in the $H \times C$ plane. In this proposal, the input is a sequence of true random observations generated by a physical-based procedure. We obtain the confidence regions by performing an orthogonal projection of the data onto the space of principal components, thus eliminating the restrictions imposed by the bounded space of the Entropy-Complexity plane. Our contributions can be summarized as follows:

- We provide the first contribution in constructing a test in the Entropy-Complexity Plane: we provide confidence regions and p-values.
- We evaluate this test's size by analyzing random sequences generated by physical procedures and pseudorandom generators (PRNGs).
- We verify the test's power contrasting correlated noise time series, and white noise series patched with a deterministic signal.

The rest of this paper is structured as follows. Section 2 introduces the elements of the study: Section 2.1 details the methodology, Section 2.2 describes the Entropy-Complexity plane, and Section 2.3 discusses applications of this approach. Section 2.1 describes our methodology: Section 3.1 gives the overall framework. Section 3.2 describes how we obtained true white noise random sequences from physical devices, Section 3.3 lists the parameters of relevance for the study, and Section 3.4 describes how we obtained the confidence intervals and p-values. Section 4 assesses the proposed test by verifying its size and power, and its application to well-known pseudorandom number generators. Section 5 concludes the paper with a discussion of these results.

2 | BANDT AND POMPE SYMBOLIZATION: A BACKGROUND

In our work, we consider the ordinal patterns formed from true white noise sequences (TWNS) using the Bandt and Pompe symbolization. Such sequences are then mapped in the two-dimensional plane of Information Theory descriptors, formed by the Permutation Entropy and the Statistical Complexity. We then obtain a test statistic, using confidence regions, that can discriminate among times series.

2.1 | The Bandt and Pompe Methodology

Let $\mathcal{X} \equiv \{x_t\}_{t=1}^{T}$ be a real-valued time series of length T, without ties. As stated by (Bandt and Pompe, 2002) in their seminal work:

> *"If the $\{x_t\}_{t=1}^{T}$ attain infinitely many values, it is common to replace them by a symbol sequence $\Pi \equiv \{\pi_j\}$ with finitely many symbols, and calculate source entropy from it".*

Also, as stressed by these authors,

> *"The corresponding symbol sequence must come naturally from the $\{x_t\}_{t=1}^T$ without former model assumptions".*

Let $\mathbb{A}_D$ (with $D \geq 2$ and $D \in \mathbb{N}$) be the symmetric group of order $D!$ formed by all possible permutation of order D, and the symbol component vector $\boldsymbol{\pi}^{(D)} = (\pi_1, \pi_2, \dots, \pi_D)$ so every element $\boldsymbol{\pi}^{(D)}$ is unique ($\pi_j \neq \pi_k \ \forall \ j \neq k$). Consider for the time series $\mathcal{X} \equiv \{x_t\}_{t=1}^T$ its time delay embedding representation, with embedding dimension $D \geq 2$ ($D \in \mathbb{N}$) and time delay $\tau \geq 1$ ($\tau \in \mathbb{N}$, also called "embedding time"):

$$\mathbf{X}_t^{(D,\tau)} = (x_t, x_{t+\tau}, \dots, x_{t+(D-1)\tau}), \tag{1}$$

for $t = 1, 2, \dots, N$ with $N = T - (D-1)\tau$. Then, the vector $\mathbf{X}_t^{(D,\tau)}$ can be mapped to a symbol $\boldsymbol{\pi}^{(D)} \in \mathbb{A}_D$. This mapping should be defined in a way that preserves the desired relation between the elements $x_t \in \mathbf{X}_t^{(D,\tau)}$, and all $t \in T$ that share this pattern (also called "motif") are mapped to the same $\boldsymbol{\pi}^{(D)}$. The two most frequent ways to define the mapping $\mathbf{X}^{(D,\tau)} \mapsto \boldsymbol{\pi}^{(D)}$ are:

a) ordering the ranks of $x_t \in \mathbf{X}^{(D,\tau)}$ in chronological order (*Rank Permutation*) or,
b) ordering the time indexes of $x_t \in \mathbf{X}^{(D,\tau)}$ (*Chronological Index Permutation*).

See details in the work by Traversaro *et al.* (2018). Without loss of generality, in the following, we will use the latter.

Consider, for instance, the time series $\mathcal{X} = (2.8, 2.2, 4.2, 5.8, 5.2, 5.5, 3.3, 4.7, 2.2, 1.5)$ depicted in Fig. 1 as a light blue line. Assume we are using patterns of length $D = 5$ with a unitary time lag $\tau = 1$. The code associated to $\mathbf{X}_3^{(5,1)} = (x_3, \dots, x_7) = (4.2, 5.8, 5.2, 5.5, 3.3)$, shown in red, is formed by the indexes in $\boldsymbol{\pi}^{(5)} = (1, 2, 3, 4, 5)$ which sort the elements of $\mathbf{X}_3^{(5,1)}$ in increasing order: 51342. With this, $\widetilde{\pi}^{(5)} = 51342$, and we increase the counting related to this motif in the histogram of all possible patterns of size $D = 5$.

The green line in Fig. 1 illustrates $\mathbf{X}_1^{(5,2)}$, i.e. the sequence of length $D = 5$ starting at x_1 with lag $\tau = 2$. In this case, $\mathbf{X}_1^{(5,2)} = (2.8, 4.2, 5.2, 3.3, 2.2)$, and the corresponding motif is $\widetilde{\pi}^{(5)} = 51423$.

After computing all the symbols, one obtains the histogram of proportions $\boldsymbol{h} = (h(j))_{1 \leq j \leq D!}$. Such histogram estimates the (unknown, in general) probability distribution function of these patterns. The next step into the characterization of the time series is computing descriptors from this histogram.

The first descriptor is a measure of the disorder of the system. The most frequently used feature for this is the Normalized Shannon entropy, defined as

$$H(\boldsymbol{h}) = -\frac{1}{\log D!} \sum_{j=1}^{D!} h(j) \log h(j), \tag{2}$$

with the convention that terms in the summation for which $h(j) = 0$ are null. This quantity is bounded in the unit interval. It is zero when $h(j) = 1$ for some j (and, thus, all other bins are zero), and one when $h(j) = 1/D!$ for every j (the uniform probability function).

Although very expressive, the Normalized Shannon Entropy is not able to describe all possible underlying dynamics. In particular, for intermediate values of H, there is a wide variety of situations worth characterizing. To this aim, López-Ruiz *et al.* (1995) proposed using the disequilibrium Q, a measure of how far $\boldsymbol{h}$ is from an equilibrium or noninformative distribution. They employed the Euclidean distance between $\boldsymbol{h}$ and the uniform probability function.

The Jensen-Shannon distance between $\boldsymbol{h}$ and the probability function $\boldsymbol{u} = (u(1), u(2), \dots, u(D!))$ stems as a

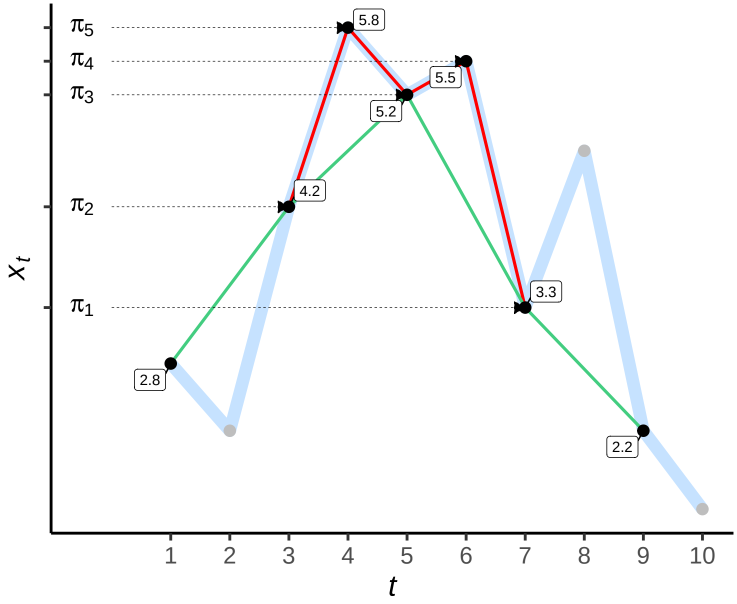


FIGURE 1 Illustration of the Bandt and Pompe coding

more adequate measure of such a distance:

$$Q'(\boldsymbol{h},\boldsymbol{u}) = \sum_{\ell=1}^{D!}\Big(h(\ell)\log\frac{h(\ell)}{u(\ell)} + u(\ell)\log\frac{u(\ell)}{h(\ell)}\Big). \tag{3}$$

This has been the preferred distance function since its proposal by Lamberti *et al.* (2004). Most works consider the uniform distribution $\boldsymbol{u} = (1/D!, 1/D!, \dots, 1/D!)$ as the reference model.

The quantity (3) is also called "disequilibrium." The normalized disequilibrium is $Q = Q'/\max\{Q'\}$, and we chose to work with the uniform probability distribution function as the equilibrium law. [Eduarda:] alguma referência (dentre as que já citamos) e algumas propriedades? Alguma simplificação quando u é a uniforme?

With this, they proposed $C = HQ$ as a measure of the Statistical Complexity of the underlying dynamics. A time series can then be mapped into a point in the $H \times C$ plane.

2.2 | The Entropy-Complexity Plane

The Entropy-Complexity plane is the set of all possible points (h, c) that can be produced by arbitrary time series analyzed with embedding dimension D that are mapped on histograms of $D!$ bins. The time delay is irrelevant, and we consider infinitely long series.

Let us consider two extreme cases:

Case I. Strictly monotonically increasing or decreasing series produce a single pattern, so the other $D! - 1$ bins of the histogram are zero. The entropy is zero, and the distance to the uniform distribution is maximal.

Therefore, the complexity is zero, and such series are mapped onto the point $(0, 0)$.

Case II. White noise produces a histogram of equal proportions $1/D!$ and maximal entropy. The distance to the equilibrium distribution is zero. Thus, such series are mapped onto the point $(1, 0)$.

Anteneodo and Plastino (1996) proved that, for a fixed value of entropy, there are two extreme values of complexity. Martín *et al.* (2006), using geometrical arguments on the space of configurations, found expressions for such boundaries. The lower boundary $C_{\min}$ is smooth, while the upper $C_{\max}$ is defined by $D! - 1$ pieces. The upper boundary converges to a smooth curve when $D \to \infty$.

Fig. 2 shows the boundaries of the $H \times C$ plane for the embedding dimensions $D = 3$ (red) $D = 4$ (green), and $D = 5$ (blue). The inset plot highlights the fine structure of the upper boundary inside the rectangle. The jagged structure of $C_{\max}$ increases the difficulty of finding distributions for the points in the $H \times C$ plane.

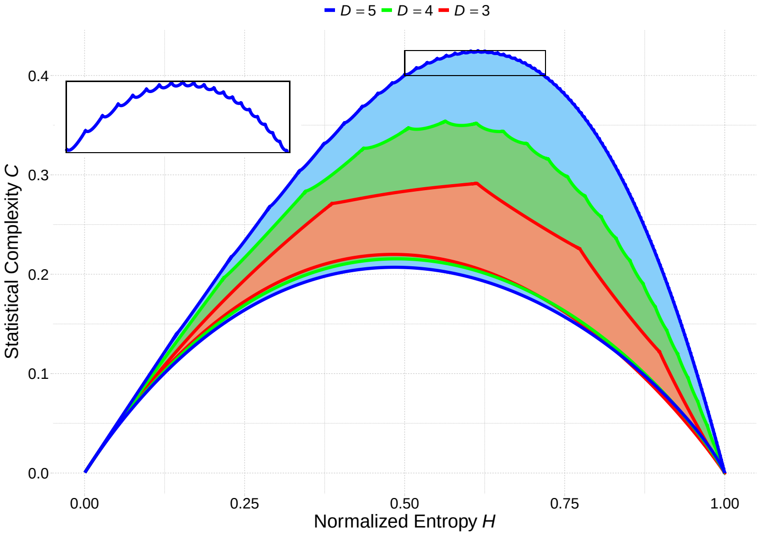


FIGURE 2 Boundaries of the $H \times C$ plane for dimension embeddings $D = 3, 4, 5$.

We illustrate the use of the Entropy-Complexity plane ($H \times C$) with the following time series:

- Colored k-noise, or f^{-k} noise: white ($k = 0$), $k = 1/2$, pink ($k = 1$), $k = 3/2$, red ($k = 2$), $k = 5/2$, and $k = 3$;
- Chaotic logistic series $x_t = r x_{t-1}(1 - x_{t-1})$, with $r = 3.6$ and 4;
- Deterministic series: monotonic increasing ($\log(x_t + 0.1)$, $x_t = \{1, 2, \dots, 10^4\}$) and periodic ($\sin(2x_t)\cos(2x_t)$, with $0 \leq x_t \leq 2\pi$ over ten thousand equally spaced points).

In all cases, we used $D = 6$ and $\tau = 1$. Fig. 3 shows nine of the histograms produced by these series using the Mersenne-Twister pseudorandom number generator; we omitted those corresponding to the deterministic series, as they produce one and two nonzero bins.

Fig. 4 shows the $H \times C$ plane with the bounds for $D = 6$, the time series, and the points they were mapped onto. The points due to f^{-k} noises appear joined by dotted segments. It is noticeable that deterministic patterns have more complexity than random ones. Also, points related to f^{-k} noises tend to clutter for $k < 1$, having the highest entropy values, as can be seen in Fig. 5.

Fig. 5 shows the rightmost lower corner of the $H \times C$ plane, emphasizing the location of the white ($k = 0$), $k = 1/2$, and pink ($k = 1$) noises.

Due to the infinitude of white noise sequences, although these sequences have the characteristic of presenting high values of entropy and low statistical complexity, their points will not necessarily be located in $(1, 0)$, but in a surrounding region. Our study's focus is to assess pure randomness by analyzing the empirical distribution of the points produced by true random sequences of finite size and obtaining regions of confidence in the $H \times C$ plane.

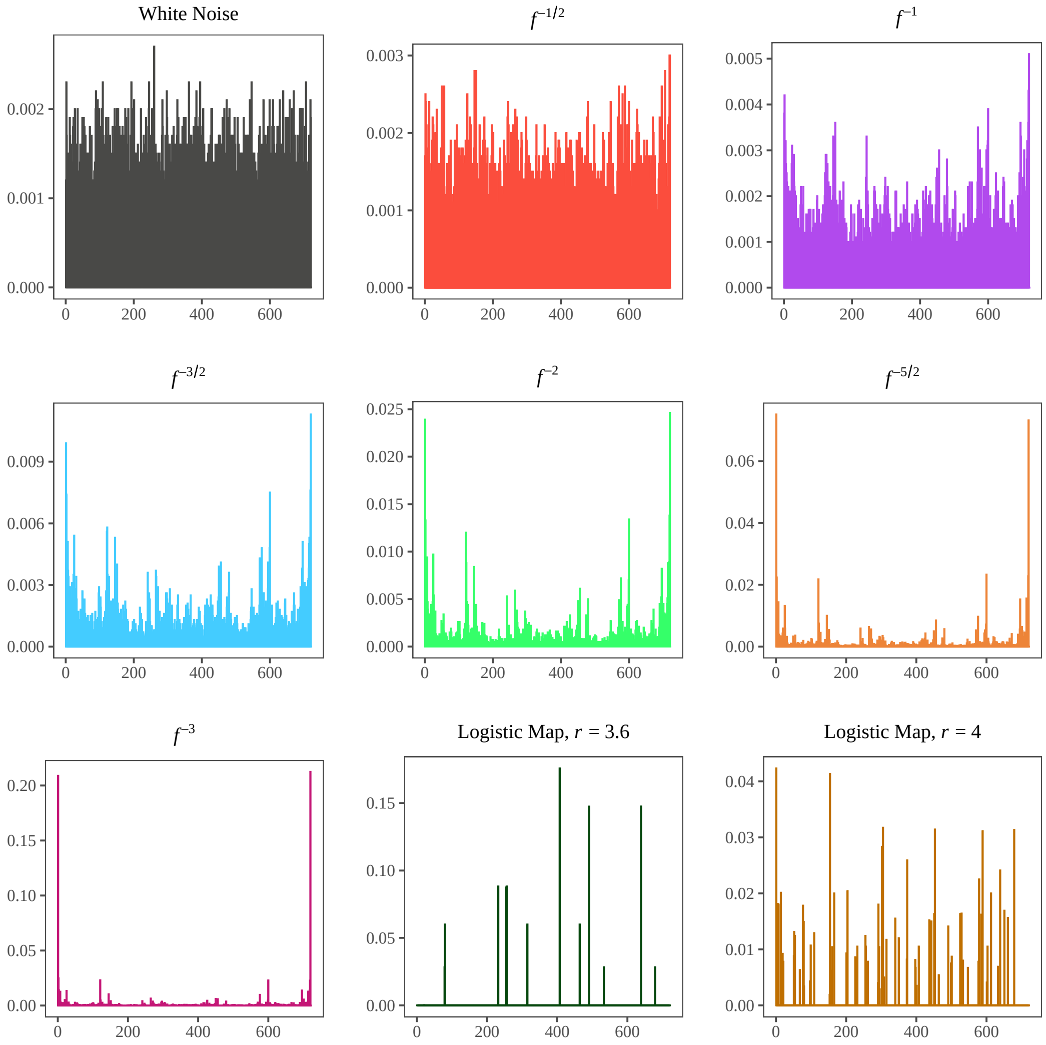


FIGURE 3 Patterns histograms of selected time series for dimension embedding $D = 6$, time delay $\tau = 1$, and sequence length $T = 1 \times 10^4$.

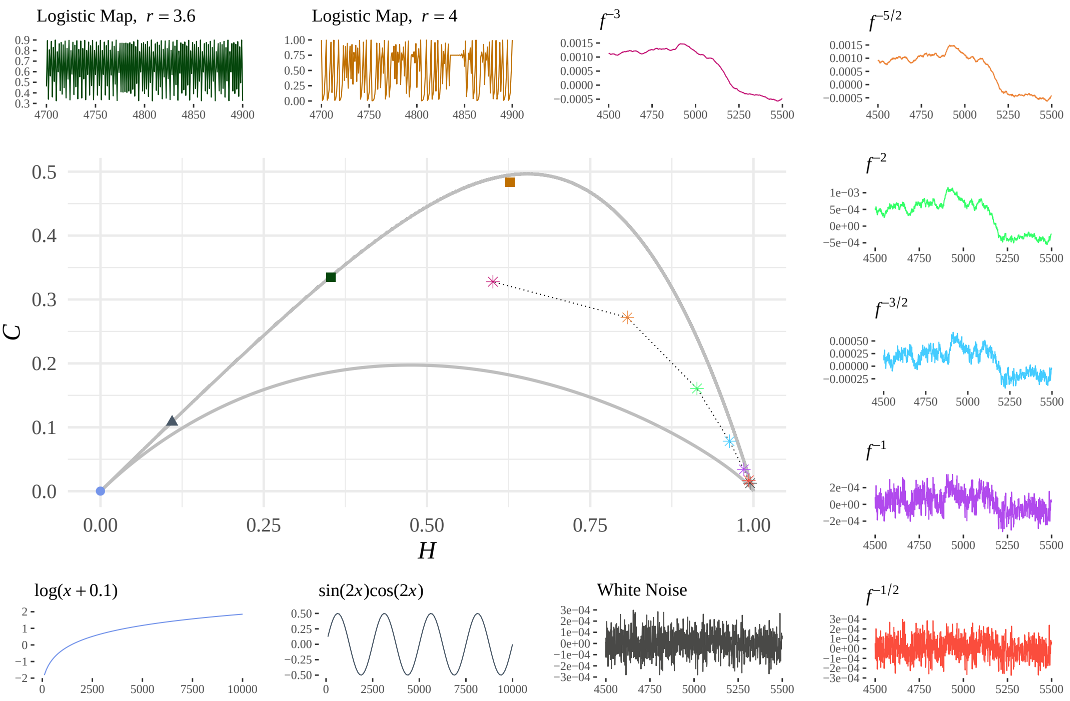


FIGURE 4 Eleven systems and their points in the $H \times C$ plane for dimension embedding $D = 6$, time delay $\tau = 1$, and sequence length $T = 1 \times 10^4$.

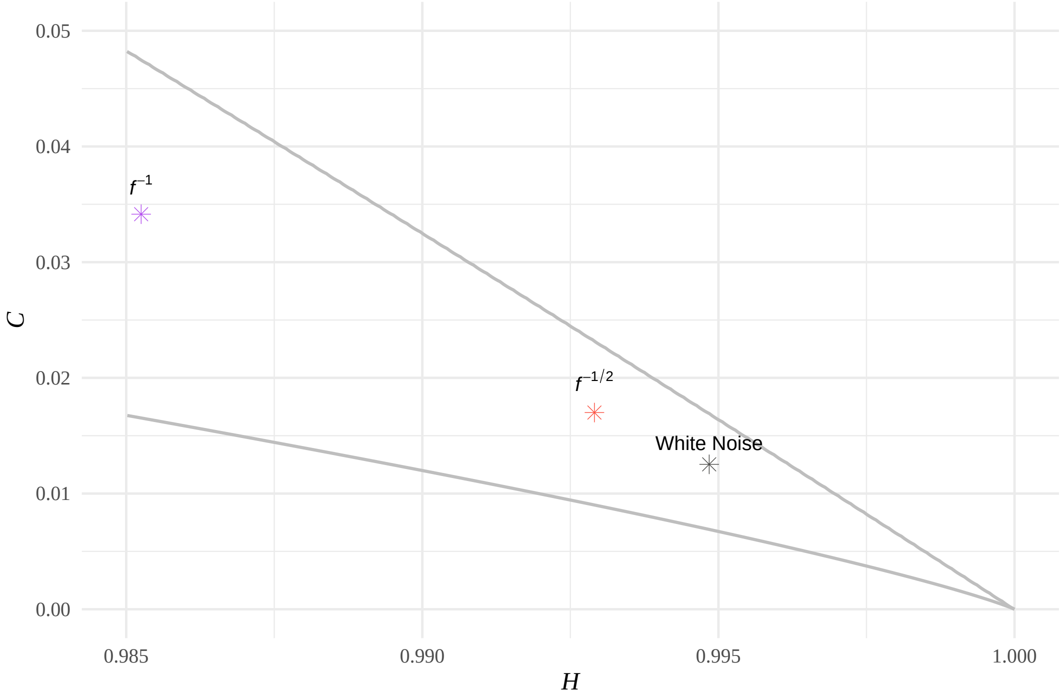


FIGURE 5 Representations of white noise, $f^{-1/2}$, and f^{-1} noise in the $H \times C$ plane for dimension embedding $D = 6$, time delay $\tau = 1$, and sequence length $T = 1 \times 10^4$.

2.3 | The Entropy-Complexity Plane in the Literature

Initial works on the characterization of white noises with permutation entropy arose from the need to discriminate between them and chaotic maps (Rosso *et al.*, 2013; Xiong *et al.*, 2020). Notice that the logistic map has been used for pseudorandom number generation; cf. Persohn and Povinelli (2012) and the references therein. It was found that the measure of statistical complexity was able to efficiently quantify the performance of pseudorandom number generators, expanding the possibilities of using Information Theory descriptors with the Bandt-Pompe symbolization (Larrondo *et al.*, 2013; González *et al.*, 2005).

Table 1 presents a summary of the main works in the literature that perform analysis of non-chaotic algorithmic generators, according to their features in the $H \times C$ plane. We also provide the length T and embedding dimension D of the time series under scrutiny. The following algorithmic generators were analyzed:

- Mother RNG, available in Marsaglia's website (MOT, Marsaglia, 1994);
- Multiple with carry RNG (MWC, Marsaglia, 1994);
- Combo RNG (COM, Marsaglia, 1994);
- Lehmer RNG (LEH, Payne *et al.*, 1969);
- Fractional Brownian motion (fBm) and fractional Gaussian noise (fGn); see Bardet *et al.* (2003);
- Colored noise with power spectrum f^{-k} with $k \geq 0$ (Larrondo, 2012);
- Linear Congruential Generator (LCG, Knuth, 1997).

TABLE 1 Results of the main works of white noise sequences analysis in the $H \times C$ plane.

Reference	PRNG	T	D	H	C	Considered random?
Larrondo *et al.* (2013)	MOT	NA	6	$\cong 0.9969$	$\cong 0$	no
González *et al.* (2005)	MWC	65536	NA	$\cong 1$	0.3	yes
	MOT	65536	NA	$\cong 1$	0.3	yes
	COM	65536	NA	$\cong 1$	0.05	yes
Larrondo *et al.* (2006)	LEH	5×10^6	5	$\cong 1$	10^{-4}	yes
	MOT	5×10^6	5	$\cong 1$	10^{-4}	yes
	MWC	5×10^6	5	$\cong 1$	10^{-4}	yes
Rosso *et al.* (2013)	LCG	1×10^7	6	0.997871	0.005101	no
Xiong *et al.* (2020)	fGn	2×10^{17}	6	$\cong 1$	$\cong 0$	yes
	f^{-k}	2×10^{17}	6	$\cong 1$	$\cong 0$	yes

None of these works provide p-values or hypothesis tests for their analysis. The focus of the articles is finding the set of descriptors that best discriminates chaos from noise. The authors make assessments about the randomness of sequence on ad hoc visual inspection of the point's location in the $H \times C$ plane. Our work fills such a gap for finite sequences of white noise and proposes a methodology that can be extended to any other situation.

3 | PROPOSED METHOD

In this section, we formalize the task of building confidence regions in the Entropy-Complexity manifold. Then, we present our proposal to change space through the algorithm of the principal components analysis. Our goal is to find a latent space representative of the data without the restrictions of the $H \times C$ plane's boundaries. Through this new representation of the data, we calculate empirical regions with different levels of confidence. Finally, after calculating these regions, we build a test statistic that determines the probability that a given sequence belongs to the distribution of the points provided.

3.1 | Overall Framework

Our methodology consists of the following steps:

1. Observe a large number of true random white noise sequences (TRWNS).
2. Map each TRWNS onto a point in the $H \times C$ plane.
3. Obtain the principal components of these points.
4. Compute enclosing boxes in the principal components space.
5. Transform the coordinates of these boxes back to the $H \times C$ plane.

3.2 | True Random Numbers

Random numbers are used in many fields, from gambling to cryptography, to guarantee a secure, realistic, or unpredictable behavior. Pseudorandom results can be achieved by software in a deterministic way, but some applications need actual random numbers (despite the somewhat elusive nature of actual randomness). Randomness can be observed in unpredictable real-world phenomena like cathodic radiation or atmospheric noise.

In this study, we used two sources of true random numbers, both from the observation and measurement of physical phenomena. The first uses vacuum states. The setup consists of an ordinary laser source to generate a local oscillator (LO), a half-wave plate, a polarizing beamsplitter (BPS), and two balanced detectors working together adding or subtracting the photocurrents results in a quadrature measurement of the LO or vacuum state. The distribution of the vacuum state is binned into 2^n equal parts (bins of the same size), assigning a fixed bit combination of length n to each sample point in a given bin (Gabriel *et al.*, 2010). The second one employs atmospheric noise captured by a cheap radio receiver with no filter for unwanted static sounds caused by atmospheric noise. This generator was implemented over a distributed setup with radios located at different geographical locations sending random bits to a cloud server that processes data and hosts the values. The history of this service and other information can be found at Haahr (1998–2018).

We used 54×10^6 B words from each physical generator, which approximately amounts 200 MB of data.

3.3 | Parameters Settings and Dataset

We conducted an ablation study to identify the influence of the parameters T, D, and τ in the construction of empirical confidence regions. We verified that the results involving the time delay parameter variation did not show significant differences in repeated experiments; therefore, in the sequel, we did no consider τ as a determining factor. On the other hand, we found two relevant variables: the length of the sequence and the embedding dimension. We, thus, employed the following factors:

- Sequence length $T \in \mathcal{T} = \{1 \times 10^3, 5 \times 10^4\}$,
- Embedding dimension $D \in \mathcal{D} = \{3, 4, 5, 6\}$.

and kept $\tau = 1$, which is the most frequently used option. The values of D are within the range recommended in the literature (Bandt and Pompe, 2002).

Using this parametric space, we analyzed the different degrees of information captured by the ordinal patterns formed. For the construction of the confidence regions presented, we used:

- A set of 104 596 points in the $H \times C$ plane, corresponding to sequences of length $T = 1000$, for each value of $D \in \mathcal{D}$, and
- a set of 2093 points in the $H \times C$ plane, corresponding to sequences of length $T = 50000$, for each value of $D \in \mathcal{D}$.

We used the `R` platform (R Core Team, 2020, v. 4.0.3) for data generation and analyses, and the `ggplot2` library (Wickham, 2009) for generating the plots.

3.4 | Empirical Confidence Regions and p-values

Our first approaches to analyzing sequences of points in the $H \times C$ plane produced by TRWNS verified that they, and usual transformations, are far from bivariate Gaussian and generalized Hyperbolic distributions (Schmidt *et al.*, 2006). Different types of regression models of C explained by H did not produce acceptable results. Thus, we adopted a non-parametric approach and made an empirical analysis of the data obtained from physical sources for using them as our reference in the search for confidence regions and p-values.

Let $\underset{\sim}{x} = (x_1, x_2, \dots, x_N)$ be N times series of length T, and define an embedding dimension D. In the sequel, whenever possible, we will omit T and D. For each $n = 1, 2, \dots, N$, the time series x_n is mapped onto the point (h_n, c_n) in the $H \times C$ plane, thus $\underset{\sim}{hc} = \big((h_1, c_n), (h_2, c_2), \dots, (h_N, c_N)\big)$ are the points that correspond to the N time series. Fig. 6a illustrates this step. We will obtain confidence regions and p-values from $\underset{\sim}{hc}$.

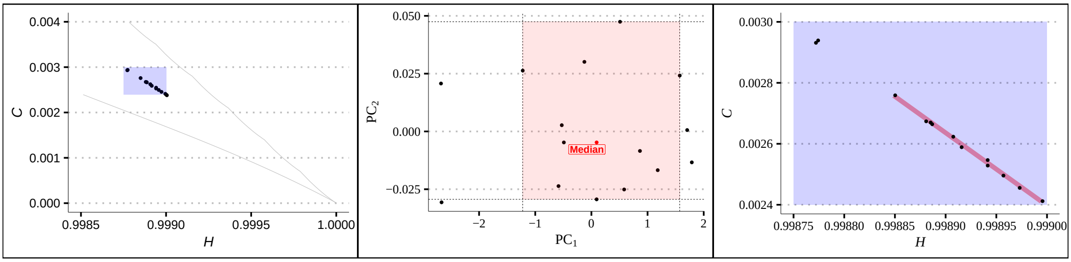
The first step consists in finding and applying the principal components transformation to $\underset{\sim}{hc}$. With this, we obtain the set of uncorrelated points $\underset{\sim}{uv} = \big((u_1, v_1), (u_2, v_2), \dots, (u_N, v_N)\big)$, in which u_n and v_n are the first and second principal components of h_n and v_n, respectively. This projection allows us to obtain a "central" point of the data set, around which we will build a rectangular box containing $100\,(1-\alpha)\%$ of the observations. Such box is a variation of the bagplot (Rousseeuw *et al.*, 1999). Notice that finding the smallest box that encloses k out of N points is difficult; cf. the work by Chan and Har-Peled (2020).

For simplicity, and without loss of generality, assume N is odd.

1. Find the indexes that sort the values of the first principal component $\boldsymbol{u} = (u_1, u_2, \dots, u_N)$ in ascending order: $\boldsymbol{r} = (r_1, r_2, \dots, r_N)$, i.e., u_{r_1} is the minimum value, and u_{r_N} is the maximum value.
2. Find the point (u, v) whose first principal component is the median: $(u_{r_{(N+1)/2}}, \cdot)$. Apply the inverse principal components transformation, and obtain $\boldsymbol{P}' = (h', v')$. Call the corresponding time series "emblematic time series."
3. Find the point (u, v) whose first principal component is the quantile $\alpha/2$: $(u_{r_{[N\alpha/2]}}, \cdot)$.
4. Find the point (u, v) whose first principal component is the quantile $1-\alpha/2$: $(u_{r_{[N(1-\alpha/2)]}}, \cdot)$.
5. The values $u_{r_{[N\alpha/2]}}$ and $u_{r_{[N(1-\alpha/2)]}}$ are the rightmost and leftmost bounds of the box, respectively.

6. The bottom bound of the box is the smallest second principal component value whose first principal component is at least $u_{r_{[N\alpha/2]}}$; denote this values $v_{\min}$.
7. The top bound of the box is the largest second principal value whose first principal component is at most $u_{r_{[N(1-\alpha/2)]}}$; denote this value $v_{\max}$.
8. The corners of the box are $(u_{r_{[N\alpha/2]}}, v_{\min})$, $(u_{r_{[N\alpha/2]}}, v_{\max})$, $(u_{r_{[N(1-\alpha/2)]}}, v_{\min})$ and $(u_{r_{[N(1-\alpha/2)]}}, v_{\max})$.
9. Apply the inverse principal components transformation to these corners obtaining $\boldsymbol{P}_1 = (h_{v_1}, c_{v_1})$, $\boldsymbol{P}_2 = (h_{v_2}, h_{v_2})$, $\boldsymbol{P}_3 = (h_{v_3}, c_{v_3})$ and $\boldsymbol{P}_4 = (h_{v_4}, c_{v_4})$.

Fig. 6 illustrates these steps. Fig. 6a shows the points produced by TRWNS in the $H \times C$ plane. The blue box includes a certain percentage of points, with sides parallel to the H and C axes. The area in the $H \times C$ plane overestimates the desired proportion and may include "unacceptable" points. Fig. 6b shows the previous points projected onto the principal components space (steps 2 to 7). The red box includes the same percentage of desired points, with axes parallel to the first and second principal components. We highlighted in red the point whose first principal component is the median of the observed values. Fig. 6c shows the result of projecting back the red box from the principal components space to the $H \times C$ plane (step 9). The comparison of the red and blue boxes shows that the area has been reduced, thus improving the test's power.



(a) Mapping true white noise random sequences onto the $H \times C$ plane.

(b) Transformation of the points in the $H \times C$ plane by Principal Components, and determination of minimal boxes.

(c) Inverse transformation from the Principal Components plane to the $H \times C$ plane.

FIGURE 6 Outline of the methodology used for the construction of the confidence regions.

Algorithm 1 provides details on how we obtain the confidence regions, defined by a set of points $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_2, \boldsymbol{P}_4$, for each $D \in \mathcal{D}$, each $T \in \mathcal{T}$, and each significance level α. We also obtain the "emblematic point" $\boldsymbol{P}'$, a kind of median point in the $H \times C$ plane for each situation.

These confidence regions obtained provide a powerful tool to make binary assessments about the adequacy of a given time series $\boldsymbol{x}$ to the null hypothesis $\mathcal{H}_0$ that it is white noise. More generally, we are interested in obtaining the p-value of $\boldsymbol{x}$ under $\mathcal{H}_0$. We present a procedure to obtain an approximate p-value based on the evidence collected to build the confidence regions.

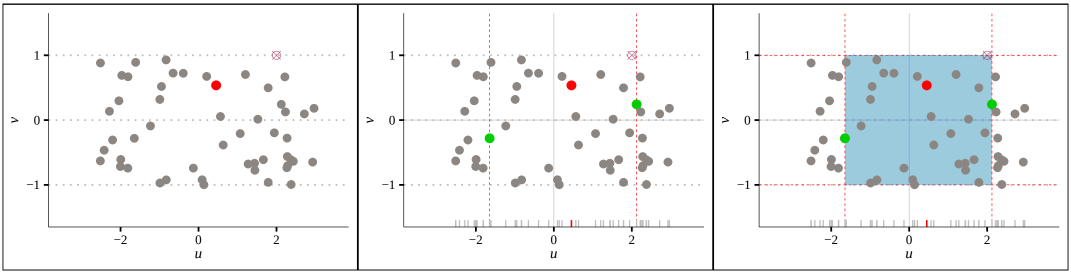
The procedure operates on the principal components space and consists of measuring the closeness between the "emblematic point" and the observed point. We are given a time series $\boldsymbol{x}$ of size T, and we want its p-value when contrasted with TWNRS of the same size at embedding dimension D. We use N TWNRS of size T, compute their points in the $H \times C$ plane, and project them to the corresponding principal components space. We then do the same with $\boldsymbol{x}$, and obtain a new point (u_x, v_x). The closer $\boldsymbol{x}$ is to the emblematic time series, the larger its p-value. Assume that the emblematic time series is represented by (u, v) in the principal components space. We measure this closeness by building a box around (u_x, v_x) that contains (u, v); assume that $u_x > u$, then:

1. the right side of the box is the smallest u_j which is larger that u_x; assume it corresponds to the quantile η_u of $\underset{\sim}{u} = (u_1, u_2, \ldots, u_N)$. By definition, $\eta_u \geq 1/2$.
2. the left side of the box is the $1 - \eta_u$ quantile of $\underset{\sim}{u}$.
3. the top side of the box is the smallest v_j which is larger that v_x; assume it corresponds to the quantile η_v of $\underset{\sim}{v} = (v_1, v_2, \ldots, v_N)$. By definition, $\eta_v \geq 1/2$.
4. the bottom side of the box is the $1 - \eta_v$ quantile of $\underset{\sim}{v}$.

Fig. 7 illustrates theses steps.

The definition of the box for the case $u_x < u$ follows naturally, and is described in Algorithm 2. With this approach, we obtain the smallest box that (i) contains the new point, and (ii) is defined by observed points from TRWNS.

Such boxes are less prone to distortions in this space since the distribution of the points becomes less asymmetric than in the $H \times C$ plane; cf. Fig. 8. Algorithm 2 shows the details.



(a) Points produced by TRWNSs in the space of principal components.

(b) Points whose first principal component is the median (red) and quantiles of order η_u and $1 - \eta_u$ (green).

(c) The p-value of the new point is the proportion of points outside the box.

FIGURE 7 Outline of the methodology used to calculate the p-value. The new point is denoted as a crossed circle.

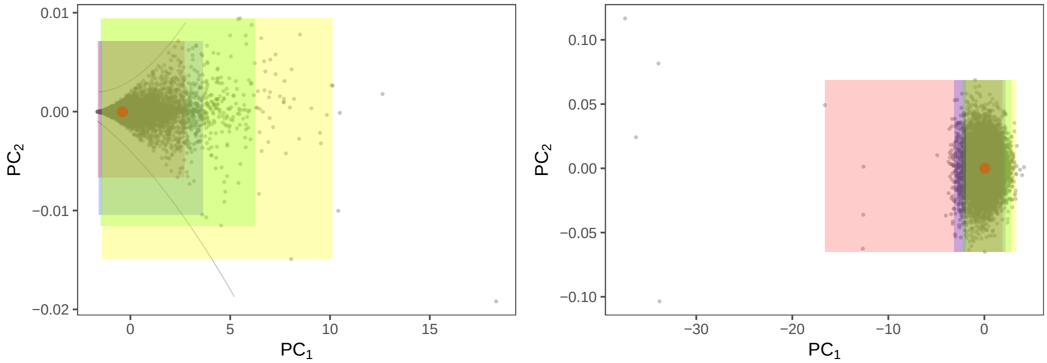
4 | RESULTS

4.1 | Empirical confidence regions

Tables 2 and 3 list the coordinates in the $H \times C$ plane of the emblematic point $\boldsymbol{P}'$ and of the four points $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4$ that define the confidence regions at 90 %, 95 %, 99 %, and 99.9 %, for $D = 3, 4, 5, 6$ and $N = 1000, 50000,$. The points are presented counterclockwise, starting with the one with the largest complexity.

Fig. 8 shows the results for $T = 50000$ and $D = 3, 6$ in the new principal components space, along with the quantiles of order 90 %, 95 %, 99 %, and 99.9 %. We also show the projection of the $H \times C$ plane boundaries in this space, as well the median of each data set, the latter being represented as red dots.

The confidence regions exceed the $H \times C$ boundaries, but this issue does not compromise the test's size since no



(a) Points in the Principal Components plane for for $D = 3$

(b) Points in the Principal Components plane for for $D = 6$

FIGURE 8 Representation of true random white noise sequences of length $T = 50000$ in the PCA space for $D = 3$ and $D = 6$, and the quantiles of 90 %, 95 %, 99 %, and 99.9 %.

points can be observed outside such boundaries.

Fig. 8 also shows that the data are not evenly distributed among the axes of the first principal component. They tend to concentrate close to the point that corresponds to $(1, 0)$ in the $H \times C$ plane. As we use order statistics to define the confidence regions, this issue is also of little relevance for our results. Moreover, Fig. 9 shows that such asymmetry diminishes when the embedding dimension D increases.

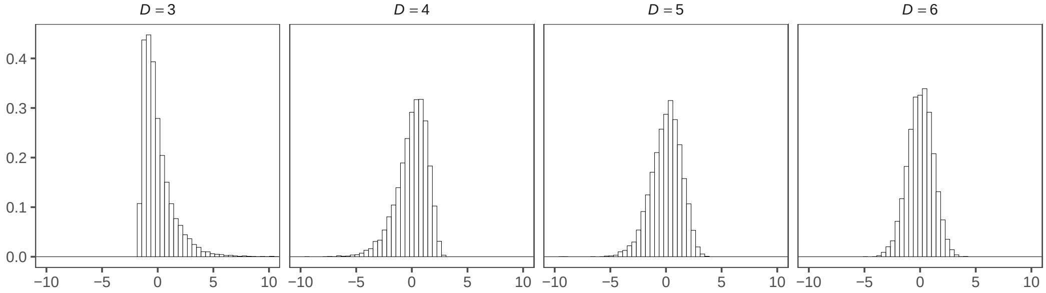


FIGURE 9 Histograms of the first principal component for $D = 3, 4, 5, 6$

4.2 | Test Size

We assessed the size of the test by contrasting 100 new TWNRS for each situation of $D = 3, 4, 5, 6$ and of $\alpha = 0.01, 0.05$. Table 4 and Fig. 10 show the results.

On the one hand, long series ($T = 50000$) present a good size for every embedding dimension. On the other hand, short series ($T = 1000$) exhibit only one situation with a noticeable divergence between the expected and the observed size: the test rejects 13 % of the 100 series when $D = 6$. In contrast, we expected 1 % of rejection. This might be because, in this case, the condition $D! \ll T$ is not respected. Notice that the wrongly rejected TWNRS are all close to the point $(1, 0)$.

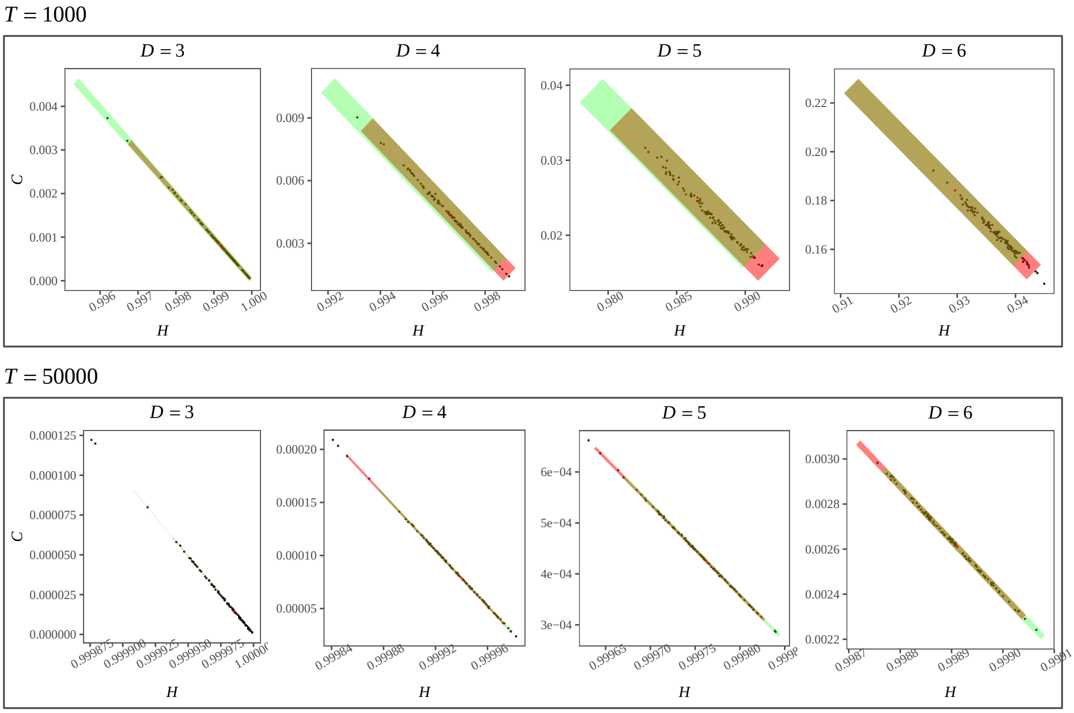


FIGURE 10 New 100 TRWNS and confidence regions.

We may then conclude that the test has good empirical size, provided $D! \ll T$, a condition that does not hold for $D = 6$ and $T = 1000$.

4.3 | Test Power

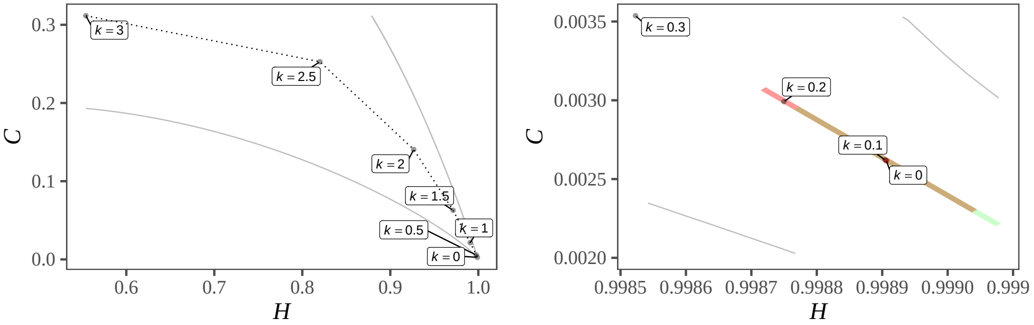
4.3.1 | H_1: Correlated Noise

We assessed the power of the test by contrasting time series with different correlation structure (under the f^{-k} model) in the $H \times C$ plane. Several studies in the literature have used this approach for identifying and characterizing randomness.

Our study's basis is the emblematic time series for each length T and dimension embedding D. Recall that the emblematic time was chosen as the most representative of the data set. We use these series, transform them into f^{-k} correlated noise, and verify the new point's location in the $H \times C$ plane.

As we can observe in the plane, as the correlation between the observations increases, that is, $k > 0$, the randomness decreases, and the entropy presented decreases, informing the loss of its stochastic characteristic.

Fig. 11a shows the overall effect of transforming the emblematic time series into f^{-k} correlated noise, with $k = 1/2, 1, 3/2, 2, 5/2, 3$. At this scale, the emblematic time series $k = 0$ and the one with $k = 1/2$ appear overlapped.



(a) Points in the $H \times C$ plane of the emblematic white noise ($k = 0$) and its transformations to become f^{-k} correlated noise with $k = 1/2, 1, 3/2, 2, 5/2, 3$.

(b) Points of the emblematic white noise ($k = 0$), and its f^{-k} correlated noise versions, with $k = 1/10, 1/5, 3/10$ along with the confidence regions for white noise.

FIGURE 11 Analysis of the test power with correlated f^{-k} noise.

As the correlation increases with k, the randomness decreases, causing a drop in the entropy; the series become progressively more predictable.

Fig. 11b is a zoom close to the $(1, 0)$ point, along with the confidence regions for the white noise. We see that $k = 0$ and $k = 0.1$ are inside the 95 % confidence region, and $k = 0.2$ is inside the 99 % box. Notice that the time series with $k = 3/10$ is outside the confidence regions and does not pass the randomness test. The same holds for all $k > 3/10$.

4.3.2 | H_1: Patch of Deterministic Function

Consider the times series of length T comprised of $T - [c100]$ white noise observations $x_1, x_2, \ldots, x_{T-[c100]}$ followed by the signal $y_{T-[c100]+1}, y_{T-[c100]+2}, \ldots, y_T$, in which $0 \leq c \ll 1$ is the percentage of contamination. We will verify the behavior of our test for the case $T = 1000$, and y_t an increasing function on t. Fig. 12 shows the effect of such a contamination on the white noise time series whose point in the $H \times C$ plane appears in black. The Figure also shows the confidence region at 99 %.

The original point is displaced, at all levels of contamination, to a region with smaller entropy and larger complexity. When the percentage of contamination is below approximately 2 %, the test will not reject the null hypothesis at the 99 %; in fact, the p-value is of around 1.5 % for $c = 0, 0.01, 0.02$. The p-value falls to orders of 10^{-6} when the contamination is of 3 % and above,

It is worth noticing that a test based solely on the entropy would have less power. In this case, the complexity component enables the test to reject the null hypothesis at small contamination percentages.

4.4 | Revisiting the White Noise Hypothesis in the Literature

In this section, we compare the performance of our test with that of previous analyses that employ the Entropy-Complexity plane. To this aim, we produced 100 sequences of length $T = 5 \times 10^4$ for each generator and computed the p-value for each $D = \{3, 4, 5, 6\}$. Previous results are shown in Table 1, and ours are in Table 5. We grouped our

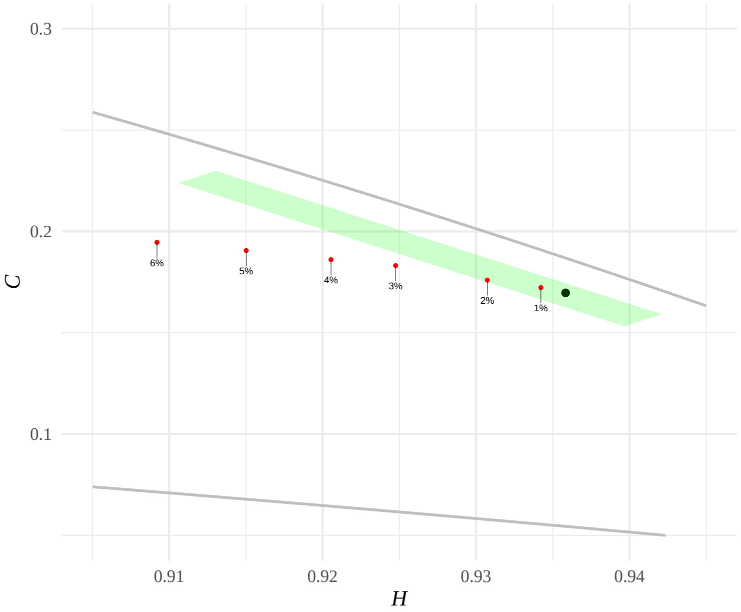


FIGURE 12 Effect of patching an increasing function at different levels to a white noise sequence.

results in those that rejected (R) the null hypothesis and those that did not reject it (NR).

Comparing Tables 1 and 5, we see that our test captures adequately the random dynamics of the sequences produced by most of the analyzed generators.

It is noteworthy that the generator Combo RNG sequences only pass our white noise test for $D = 3$. In higher embedding dimensions, as we consider longer words, the sequences produced by this generator are not labeled as white noise.

5 | CONCLUSIONS

We presented and evaluated the first test for white noise in the Entropy-Complexity plane. Our proposal is based on two stages: (1) building nonparametric empirical confidence regions in the principal components space and mapping these boxes back to the $H \times C$ plane. (2) computing an approximate p-value for a given sample by comparing it with points produced by true white noise random sequences (TWNRS). We obtained the TWNRS with data from physical devices.

Our test has a good size, mostly with long TWNRS. We also determined the power of our test for the alternative hypothesis of correlated f^{-k} noise and found that it rejects the null hypothesis ($k = 0$) for $k > 3/10$.

Although our work focuses on the study of short sequences, we were able to capture the random behavior of well-known pseudorandom number generators already analyzed in the literature. With this, we verified the adequacy of our technique as it is capable of detecting correlation structures.

6 | REPRODUCIBILITY AND REPLICABILITY

Following the recommendations provided by Frery *et al.* (2020), we make the text, source code, and data used in this study available at the *Confidence-Regions* repository `https://github.com/EduardaChagas/ConfidenceRegions`.

7 | ACKNOWLEDGEMENTS

This work was partially funded by the Coordination for the Improvement of Higher Education Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq).

references

Abrams, A., Babson, E., Landau, H., Landau, Z. and Pommersheim, J. (2013) Distributions of order patterns of interval maps. *Combinatorics Probability & Computing*, **22**, 319–341.

Anteneodo, C. and Plastino, A. R. (1996) Some features of the López-Ruiz-Mancini-Calbet (LMC) statistical measure of complexity. *Physics Letters A*, **223**, 348–354.

Aquino, A. L. L., Cavalcante, T. S. G., Almeida, E. S., Frery, A. C. and Rosso, O. A. (2015) Characterization of vehicle behavior with information theory. *The European Physical Journal B: Condensed Matter and Complex Systems*, **85**, 257–269. URL `http://dx.doi.org/10.1140/epjb/e2015-60384-x`.

Aquino, A. L. L., Ramos, H. S., Frery, A. C., Viana, L. P., Cavalcante, T. S. G. and Rosso, O. A. (2017) Characterization of electric load with information theory quantifiers. *Physica A*, **465**, 277–284.

Bandt, C. and Pompe, B. (2002) Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, **88**, 174102–1–174102–4.

Baravalle, R., Rosso, O. A. and Montani, F. (2018a) Discriminating imagined and non-imagined tasks in the motor cortex area: Entropy-complexity plane with a wavelet decomposition. *Physica A: Statistical Mechanics and its Applications*, **511**, 27–39.

Baravalle, R., Rosso, O. A. and Montani, F. (2018b) Rhythmic activities of the brain: Quantifying the high complexity of beta and gamma oscillations during visuomotor tasks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **28**, 075513.

Bardet, J.-M., Lang, G., Oppenheim, G., Philippe, A. and Taqqu, M. S. (2003) Generators of long-range dependent processes: a survey. *Theory and applications of long-range dependence*, 579–623.

Bariviera, A. F., Guercio, M. B., Martinez, L. B. and Rosso, O. A. (2015) The (in)visible hand in the Libor market: an information theory approach. *The European Physical Journal B*, **88**, 208.

Bariviera, A. F., Zunino, L., Guercio, M. B., Martinez, L. B. and Rosso, O. A. (2013) Efficiency and credit ratings: a permutation-information-theory analysis. *Journal of Statistical Mechanics: Theory and Experiment*, **2013**, P08007.

Bariviera, A. F., Zunino, L. and Rosso, O. A. (2018) An analysis of high-frequency cryptocurrencies prices dynamics using permutation-information-theory quantifiers. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **28**, 075511.

Chagas, E., Frery, A. C., Rosso, O. and S.Ramos, H. (2020) Analysis and classification of sar textures using information theory. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1–1. URL `http://dx.doi.org/10.1109/JSTARS.2020.3031918`.

Chan, T. M. and Har-Peled, S. (2020) Smallest k-enclosing rectangle revisited. *Discrete & Computational Geometry*. URL `http://dx.doi.org/10.1007/s00454-020-00239-3`.

Cryer, J. D. and Chan, K.-S. (2008) *Time Series Analysis With Applications in R.* Springer, 2 edn.

de Araujo, F. H. A., Bejan, L., Rosso, O. A. and Stosic, T. (2019) Permutation entropy and statistical complexity analysis of Brazilian agricultural commodities. *Entropy*, **21**, 1220.

De Micco, L., González, C. M., Larrondo, H. A., Martin, M. T., Plastino, A. and Rosso, O. A. (2008) Randomizing nonlinear maps via symbolic dynamics. *Physica A: Statistical Mechanics and its Applications*, **387**, 3373–3383. URL http://dx.doi.org/10.1016/j.physa.2008.02.037.

De Micco, L., Larrondo, H. A., Plastino, A. and Rosso, O. A. (2009) Quantifiers for randomness of chaotic pseudo-random number generators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367**, 3281–3296. URL http://dx.doi.org/10.1098/rsta.2009.0075.

Echegoyen, I., López-Sanz, D., Martínez, J. H., Maestú, F. and Buldú, J. M. (2020) Permutation entropy and statistical complexity in mild cognitive impairment and alzheimer's disease: An analysis based on frequency bands. *Entropy*, **22**, 116.

Frery, A. C., Gomez, L. and Medeiros, A. C. (2020) A badging system for reproducibility and replicability in remote sensing research. *IEEE Journal of Selected Topics on Applied Earth Observations and Remote Sensing*, **13**, 4988–4995.

Gabriel, C., Wittmann, C., Sych, D., Dong, R., Mauerer, W., Andersen, U. L., Marquardt, C. and Leuchs, G. (2010) A generator for unique quantum random numbers based on vacuum states. *Nature Photonics*, **4**, 711–715. URL http://dx.doi.org/10.1038/NPHOTON.2010.197.

González, C., Larrondo, H. and Rosso, O. (2005) Statistical complexity measure of pseudorandom bit generators. *Physica A: Statistical Mechanics and its Applications*, **354**, 281–300.

Haahr, M. (1998–2018) RANDOM.ORG: true random number service. https://www.random.org. Accessed: 2018-06-01.

Knuth, D. (1997) Sorting and searching, vol. 3 of the art of computer programming, section 6.2. 2.

Lamberti, P. W., Martín, M. T., Plastino, A. and Rosso, O. A. (2004) Intensive entropic non-triviality measure. *Physica A: Statistical Mechanics and its Applications*, **334**, 119–131. URL http://www.sciencedirect.com/science/article/pii/S0378437103010963.

Larrondo, H. (2012) Matlab program: noisefk. m. http://www.mathworks.com/matlabcentral/fileexchange/35381.

Larrondo, H. A., Martín, M. T., González, C. M., Plastino, A. and Rosso, O. A. (2006) Random number generators and causality. *Physics Letters A*, **352**, 421–425. URL http://www.sciencedirect.com/science/article/pii/S0375960105018232.

Larrondo, H. A., Micco, L. D., Gonzalez, C. M., Plastino, A. and Rosso, O. A. (2013) Statistical complexity of chaotic pseudo-random number generators. In *Concepts and Recent Advances in Generalized Information Measures and Statistics* (eds. A. M. Kowalski, R. D. Rossignoli and E. M. F. Curado), 283–308. Bentham Science Publishers.

López-Ruiz, R., Mancini, H. and Calbet, X. (1995) A statistical measure of complexity. *Physics Letters A*, **209**, 321–326. URL http://www.sciencedirect.com/science/article/pii/0375960195008675.

Marsaglia, G. (1994) Yet another RNG. *Posted to the Electronic Billboard Sci. Stat. Math, August*, **1**.

Martín, M. T., Plastino, A. and Rosso, O. A. (2006) Generalized statistical complexity measures: Geometrical and analytical properties. *Physica A: Statistical Mechanics and its Applications*, **369**, 439–462.

Payne, W., Rabung, J. R. and Bogyo, T. (1969) Coding the Lehmer pseudo-random number generator. *Communications of the ACM*, **12**, 85–86.

Persohn, K. J. and Povinelli, R. J. (2012) Analyzing logistic map pseudorandom number generators for periodicity induced by finite precision floating-point representation. *Chaos, Solitons & Fractals*, **45**, 238–245.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Ravetti, M. G., Carpi, L. C., Gonçalves, B. A., Frery, A. C. and Rosso, O. A. (2014) Distinguishing noise from chaos: objective versus subjective criteria using Horizontal Visibility Graph. *PLOS One*, **9**, 1–15.

Rosso, O., Larrondo, H., Martin, M., Plastino, A. and Fuentes, M. (2007) Distinguishing noise from chaos. *Physical Review Letters*, **99**, 154102.

Rosso, O. and Masoller, C. (2009a) Detecting and quantifying temporal correlations in stochastic resonance via information theory measures. *The European Physical Journal B*, **69**, 37–43.

Rosso, O. A. and Masoller, C. (2009b) Detecting and quantifying stochastic and coherence resonances via information-theory complexity measurements. *Physical Review E*, **79**, 040106.

Rosso, O. A., Olivares, F., Zunino, L., De Micco, L., Aquino, A. L., Plastino, A. and Larrondo, H. A. (2013) Characterization of chaotic maps using the permutation bandt-pompe probability distribution. *The European Physical Journal B*, **86**, 116.

Rosso, O. A., Ospina, R. and Frery, A. C. (2016) Classification and verification of handwritten signatures with time causal information theory quantifiers. *PLoS ONE*, **11**, e0166868.

Rousseeuw, P. J., Ruts, I. and Tukey, J. W. (1999) The bagplot: A bivariate boxplot. *The American Statistician*, **53**, 382.

Saco, P. M., Carpi, L. C., Figliola, A., Serrano, E. and Rosso, O. A. (2010) Entropy analysis of the dynamics of el niño/southern oscillation during the holocene. *Physica A: Statistical Mechanics and its Applications*, **389**, 5022–5027.

Schieber, T. A., Carpi, L., Frery, A. C., Rosso, O. A., Pardalos, P. M. and Ravetti, M. G. (2016) Information theory perspective on network robustness. *Physics Letters A*, **380**, 359–364.

Schmidt, R., Hrycej, T. and Stützle, E. (2006) Multivariate distribution models with generalized hyperbolic margins. *Computational Statistics & Data Analysis*, **50**, 2065–2096.

Sinn, M. and Keller, K. (2011) Estimation of ordinal pattern probabilities in Gaussian processes with stationary increments. *Computational Statistics & Data Analysis*, **55**, 1781–1790.

Traversaro, F., Redelico, F., Risk, M., Frery, A. and Rosso, O. (2018) Bandt-pompe symbolization dynamics for time series with tied values: A data-driven approach. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **28**, 075502.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Xiong, H., Shang, P., He, J. and Zhang, Y. (2020) Complexity and information measures in planar characterization of chaos and noise. *Nonlinear Dynamics*, **100**, 1673–1687.

Zunino, L., Bariviera, A. F., Guercio, M. B., Martinez, L. B. and Rosso, O. A. (2012a) On the efficiency of sovereign bond markets. *Physica A: Statistical Mechanics and its Applications*, **391**, 4342–4349.

Zunino, L., Soriano, M. C. and Rosso, O. A. (2012b) Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach. *Phys. Rev. E*, **86**, 046210. URL `http://link.aps.org/doi/10.1103/PhysRevE.86.046210`.

Zunino, L., Zanin, M., Tabak, B. M., Pérez, D. G. and Rosso, O. A. (2010) Complexity-entropy causality plane: A useful approach to quantify the stock market inefficiency. *Physica A: Statistical Mechanics and its Applications*, **389**, 1891–1901.

```
input  : A data base of true random values
input  : The desired values of embedding dimension $\mathcal{D}$, sequence length $\mathcal{T}$, and confidence levels $\mathcal{A}$
output : Confidence regions as points in the $H \times C$ plane
for each $D \in \mathcal{D}$ do
    for each $T \in \mathcal{T}$ do
        for each $n = 1, 2, \dots, N$ do
            build the time series $p_n$ with unused values from the data base;
            compute the point $(h_n, c_n)$ in the $H \times C$ plane that corresponds to $p_n$;
        end
        obtain $\text{PC}(D, T)$, the principal components transformation based on the points
          $(h_1, c_1), (h_2, c_2), \dots, (h_N, c_N)$, and its inverse $\text{PC}^{-1}(D, T)$;
        apply $\text{PC}(D, T)$ to the points $(h_1, c_1), (h_2, c_2), \dots, (h_N, c_N)$, and obtain $(u_1, v_1), (u_2, v_2), \dots, (u_N, v_N)$;
        find the indexes $\boldsymbol{r} = (r_1, r_2, \dots, r_N)$ that sort the values of the first principal component
          $\boldsymbol{u} = (u_1, u_2, \dots, u_N)$ in ascending order;
        find the point $(u, v)$ whose first principal component is the median: $(u_{r_{(N+1)/2}}, \cdot)$;
        apply the inverse principal components transformation $\text{PC}^{-1}(D, T)$ to $(u, v)$, and obtain
          $\boldsymbol{P}' = (h', v')$; call the corresponding time series “emblematic time series”;
        return $\boldsymbol{P}'$;
        for each confidence level $\alpha \in \mathcal{A}$ do
            find the point $(u, v)$ whose first principal component is the quantile $\alpha/2$: $(u_{r_{[N\alpha/2]}}, \cdot)$;
            find the point $(u, v)$ whose first principal component is the quantile $1 - \alpha/2$: $(u_{r_{[N(1-\alpha/2)]}}, \cdot)$;
            the values $u_{r_{[N\alpha/2]}}$ and $u_{r_{[N(1-\alpha/2)]}}$ are the rightmost and leftmost bounds of the box, respectively;
            the bottom bound of the box is the smallest second principal component value whose first
              principal component is at least $u_{r_{[N\alpha/2]}}$; denote this value $v_{\min}$;
            the top bound of the box is the largest second principal value whose first principal component is
              at most $u_{r_{[N(1-\alpha/2)]}}$; denote this value $v_{\max}$;
            the corners of the box are $(u_{r_{[N\alpha/2]}}, v_{\min})$, $(u_{r_{[N\alpha/2]}}, v_{\max})$, $(u_{r_{[N(1-\alpha/2)]}}, v_{\min})$ and
              $(u_{r_{[N(1-\alpha/2)]}}, v_{\max})$;
            apply the inverse principal components transformation $\text{PC}^{-1}(D, T)$ to these corners obtaining
              $\boldsymbol{P}_1 = (h_{v_1}, c_{v_1})$, $\boldsymbol{P}_2 = (h_{v_2}, h_{v_2})$, $\boldsymbol{P}_3 = (h_{v_3}, c_{v_3})$ and $\boldsymbol{P}_4 = (h_{v_4}, c_{v_4})$;
            return $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_3$, $\boldsymbol{P}_4$;
        end
    end
end
```

Algorithm 1: Determination of confidence regions and emblematic time series

```
input  : The sequence $\boldsymbol{x}$ of length $T$ to be contrasted to the null hypothesis $\mathcal{H}_0$ that it is adherent to white
         noise
input  : The embedding dimension $D$
input  : $N$ points $(h_1, c_1), (h_2, c_2), \dots, (h_N, c_N)$ of true white noise series of length $T$; the principal
         components tranformation PC$(D, T)$ induced by these points; the points projected onto the $H \times C$
         plane: $(u_1, v_1), (u_2, v_2), \dots, (u_N, v_N)$
output : An approximate $p$-value
find the point $(u, v)$ whose first principal component is the median: $(u_{r_{(N+1)/2}}, \cdot)$;
find the point $(h, c)$ of the sequence $\boldsymbol{x}$;
find the projection $(u_x, v_x)$ of $(h, c)$ onto the principal components space using PC$(D, T)$;
if $u_x > u$ then
    $u_{r_{[N(1-\alpha/2)]}}$ is defined as the smallest element larger than $u_x$;
    $u_{r_{[N\alpha/2]}} \leftarrow 2u - u_{r_{[N(1-\alpha/2)]}}$;
else
    if $u_x < u$ then
        $u_{r_{[N\alpha/2]}}$ is the largest minor element of $u_x$;
        $u_{r_{[N(1-\alpha/2)]}} \leftarrow 2u - u_{r_{[N\alpha/2]}}$;
    else
        $u_{r_{[N\alpha/2]}}$ and $u_{r_{[N(1-\alpha/2)]}}$ is equal to $u$, the median point of the first principal component;
    end
end
obtain the maximum values of the second component whose values of the first principal component are at
  least $u_{r_{[N\alpha/2]}}$ and at most $u_{r_{[N(1-\alpha/2)]}}$ and denote it $v_{\max}$;
obtain the minimum values of the second component whose values of the first principal component are at
  least $u_{r_{[N\alpha/2]}}$ and at most $u_{r_{[N(1-\alpha/2)]}}$ and denote it $v_{\min}$;
the corners of the box $b_\alpha(h, c)$ are $(u_{r_{[N\alpha/2]}}, v_{\min})$, $(u_{r_{[N\alpha/2]}}, v_{\max})$, $(u_{r_{[N(1-\alpha/2)]}}, v_{\min})$ and $(u_{r_{[N(1-\alpha/2)]}}, v_{\max})$;
count $n_{\boldsymbol{x}}$, the number of points out of the $N$ points which belong to $b_\alpha(h, c)$;
return $1 - n_{\boldsymbol{x}}/N$
```

Algorithm 2: Determination of the p-value of the sequence $\boldsymbol{x}$ under $\mathcal{H}_0$

TABLE 2 Coordinates in the $H \times C$ plane of the emblematic series and the points that define the confidence regions at 90 %, 95 %, 99 %, and 99.9 % for $D = 3, 4, 5, 6$ and $N = 1000$

		$N = 1000$			
D	Point	90 %	95 %	99 %	99.9 %
3	$\boldsymbol{P'}$	(0.9992089, 0.0007800)			
	$\boldsymbol{P}_1$	(0.9973334, 0.0025601)	(0.9967311, 0.0031343)	(0.9953009, 0.0045054)	(0.9931825, 0.0065387)
	$\boldsymbol{P}_2$	(0.9974047, 0.0026304)	(0.9968219, 0.0032238)	(0.9954349, 0.0046375)	(0.9933704, 0.006724)
	$\boldsymbol{P}_3$	(0.9999497, 0)	(0.9999398, 0)	(0.9999203, 0)	(0.9998925, 0)
	$\boldsymbol{P}_4$	$(1, 5.17 \times 10^{-5})$	$(1, 6.12 \times 10^{-5})$	$(1, 8.45 \times 10^{-5})$	(1, 0.0001104)
4	$\boldsymbol{P'}$	(0.9967032, 0.0043297)			
	$\boldsymbol{P}_1$	(0.994364, 0.0081246)	(0.9937138, 0.0089796)	(0., 99225750.0108947)	(0.9902578, 0.0135243)
	$\boldsymbol{P}_2$	(0.9939234, 0.0075452)	(0.9932534, 0.0083741)	(0.9917308, 0.0102022)	(0.9897312, 0.0128318)
	$\boldsymbol{P}_3$	(0.9994791, 0.0013982)	(0.9991609, 0.0018166)	(0.9987924, 0.0023012)	(0.9985727, 0.0025901)
	$\boldsymbol{P}_4$	(0.9990385, 0.0008188)	(0.9987005, 0.0012111)	(0.9982658, 0.0016087)	(0.9980461, 0.0018976)
5	$\boldsymbol{P'}$	(0.9864873, 0.0245632)			
	$\boldsymbol{P}_1$	(0.9811818, 0.0321294)	(0.9801289, 0.0340045)	(0.977917, 0.0377295)	(0.9753326, 0.0425299)
	$\boldsymbol{P}_2$	(0.9827429, 0.0350291)	(0.9817117, 0.0369446)	(0.9796031, 0.0408613)	(0.9770187, 0.0456617)
	$\boldsymbol{P}_3$	(0.9919707, 0.0120896)	(0.9909376, 0.0139279)	(0.9898161, 0.0156277)	(0.9892599, 0.0166608)
	$\boldsymbol{P}_4$	(0.9935319, 0.0149893)	(0.9925204, 0.016868)	(0.9915021, 0.0187595)	(0.9909459, 0.0197926)
6	$\boldsymbol{P'}$	(0, 9296429, 0.1841438)			
	$\boldsymbol{P}_1$	(0.9121895, 0.2201993)	(0.9105951, 0.2239294)	(0.9105951, 0.2239294)	(0.9077672, 0.2305874)
	$\boldsymbol{P}_2$	(0.9146048, 0.2260776)	(0.9130413, 0.2298829)	(0.9130413, 0.2298829)	(0.9102595, 0.2366531)
	$\boldsymbol{P}_3$	(0.9443868, 0.1418373)	(0.9419202, 0.1476904)	(0.9396577, 0.1531967)	(0.9383611, 0.1561279)
	$\boldsymbol{P}_4$	(0.9468021, 0.1477156)	(0.9443663, 0.1536439)	(0.9421039, 0.1591502)	(0.9408534, 0.1621937)

TABLE 3 Coordinates in the $H \times C$ plane of the emblematic series and the points that define the confidence regions at 90 %, 95 %, 99 %, and 99.9 % for $D = 3, 4, 5, 6$ and $N = 50000$

		$N = 50000$			
D	Point	90 %	95 %	99 %	99.9 %
3	$\boldsymbol{P}'$	$(0.9999853, 1.45 \times 10^{-5})$			
	$\boldsymbol{P}_1$	$(0.9999489, 5.06 \times 10^{-5})$	$(0.9999384, 6.11 \times 10^{-5})$	$(0.9999079, 9.11 \times 10^{-5})$	$(0.9998625, 0.0001361)$
	$\boldsymbol{P}_2$	$(0.9999487, 5.04 \times 10^{-5})$	$(0.9999382, 6.09 \times 10^{-5})$	$(0.9999077, 9.09 \times 10^{-5})$	$(0.9998622, 0.0001358)$
	$\boldsymbol{P}_3$	$(0.9999998, 4 \times 10^{-7})$	$(0.9999994, 9 \times 10^{-7})$	$(0.9999982, 2 \times 10^{-6})$	$(0.9999973, 3 \times 10^{-6})$
	$\boldsymbol{P}_4$	$(0.9999996, 2 \times 10^{-7})$	$(0.9999991, 7 \times 10^{-7})$	$(0.999998, 1.8 \times 10^{-6})$	$(0.999997, 2.7 \times 10^{-6})$
4	$\boldsymbol{P}'$	$(0.9999394, 7.94 \times 10^{-5})$			
	$\boldsymbol{P}_1$	$(0.9999684, 3.98 \times 10^{-5})$	$(0.9999725, 3.44 \times 10^{-5})$	$(0.9999783, 2.68 \times 10^{-5})$	$(0.9999833, 2.02 \times 10^{-5})$
	$\boldsymbol{P}_2$	$(0.9999696, 4.13 \times 10^{-5})$	$(0.9999737, 3.6 \times 10^{-5})$	$(0.9999795, 2.83 \times 10^{-5})$	$(0.9999845, 2.18 \times 10^{-5})$
	$\boldsymbol{P}_3$	$(0.9998075, 0.0002508)$	$(0.9998506, 0.0001942)$	$(0.9998756, 0.0001615)$	$(0.9998889, 0.000144)$
	$\boldsymbol{P}_4$	$(0.9998087, 0.0002524)$	$(0.9998518, 0.0001958)$	$(0.9998768, 0.000163)$	$(0.9998901, 0.0001456)$
5	$\boldsymbol{P}'$	$(0.9997616, 0.0004264)$			
	$\boldsymbol{P}_1$	$(0.9998172, 0.0003232)$	$(0.9998259, 0.0003075)$	$(0.9998428, 0.0002774)$	$(0.9998573, 0.0002517)$
	$\boldsymbol{P}_2$	$(0.9998194, 0.0003273)$	$(0.9998282, 0.0003116)$	$(0.999845, 0.0002814)$	$(0.9998593, 0.0002553)$
	$\boldsymbol{P}_3$	$(0.9994812, 0.0009246)$	$(0.9996371, 0.0006455)$	$(0.9996703, 0.0005862)$	$(0.9996884, 0.000554)$
	$\boldsymbol{P}_4$	$(0.9994834, 0.0009286)$	$(0.9996394, 0.0006495)$	$(0.9996725, 0.0005901)$	$(0.9996904, 0.0005576)$
6	$\boldsymbol{P}'$	$(0.9989108, 0.0026093)$			
	$\boldsymbol{P}_1$	$(0.9990169, 0.002336)$	$(0.9990368, 0.002288)$	$(0.9990736, 0.0021997)$	$(0.9991069, 0.0021197)$
	$\boldsymbol{P}_2$	$(0.9990249, 0.0023554)$	$(0.9990449, 0.0023074)$	$(0.9990817, 0.0022191)$	$(0.999115, 0.0021392)$
	$\boldsymbol{P}_3$	$(0.9978983, 0.0050219)$	$(0.998714, 0.0030633)$	$(0.998765, 0.0029407)$	$(0.9987884, 0.0028845)$
	$\boldsymbol{P}_4$	$(0.9979064, 0.0050413)$	$(0.998722, 0.0030827)$	$(0.9987731, 0.0029601)$	$(0.9987965, 0.0029039)$

TABLE 4 Empirical sizes of the test

N	D	95 %	99 %
1000	3	0.98	1.00
	4	0.98	0.96
	5	1.00	0.94
	6	0.97	0.87
50000	3	0.97	0.96
	4	0.94	0.95
	5	0.97	0.96
	6	0.98	0.99

TABLE 5 Results of the sequences generated by the main PRNGs in the literature. The sequences have length $T = 5 \times 10^4$.

Algorithm	D	p-value	HC-PCA
MOT	3	0.305	NR
	4	0.572	NR
	5	0.455	NR
	6	0.508	NR
MWC	3	0.501	NR
	4	0.477	NR
	5	0.496	NR
	6	0.496	NR
COM	3	0.123	NR
	4	0.002	R
	5	1.11×10^{-16}	R
	6	1.11×10^{-16}	R
LEH	3	0.531	NR
	4	0.515	NR
LEH	5	0.495	NR
	6	0.501	NR
fGn	3	0.521	NR
	4	0.519	NR
	5	0.498	NR
	6	0.470	NR
f^{-k}	3	0.482	NR
	4	0.520	NR
	5	0.513	NR
	6	0.508	NR
LCG	3	0.009	R
	4	1.11×10^{-16}	R
	5	1.11×10^{-16}	R
	6	1.11×10^{-16}	R