



Recurrence plots of discrete-time Gaussian stochastic processes



Sofiane Ramdani^{a,*}, Frédéric Bouchara^{b,c}, Julien Lagarde^a, Annick Lesne^{d,e}

^a EuroMov, Montpellier University, Montpellier, France

^b CNRS UMR 7296, LSIS, ENSAM, Aix Marseille University, Marseille, France

^c CNRS UMR 7296, LSIS, Toulon University, La Garde, France

^d CNRS UMR 7600, Laboratoire de Physique Théorique de la Matière Condensée, UPMC-Paris 6, Sorbonne Universités, Paris, France

^e CNRS UMR 5535, Institut de Génétique Moléculaire de Montpellier, Montpellier University, Montpellier, France

HIGHLIGHTS

- We derive theoretical values of RP diagonal-based measures for Gaussian processes.
- We illustrate our results on AR(1) processes and fractional Gaussian noise.
- Our results provide a benchmark to improve estimation of the considered measures.

ARTICLE INFO

Article history:

Received 15 July 2015

Accepted 29 April 2016

Available online 9 May 2016

Communicated by T. Sauer

Keywords:

Recurrence plots

Discrete-time Gaussian processes

Correlation entropy

AR(1) process

Fractional Gaussian noise

ABSTRACT

We investigate the statistical properties of recurrence plots (RPs) of data generated by discrete-time stationary Gaussian random processes. We analytically derive the theoretical values of the probabilities of occurrence of recurrence points and consecutive recurrence points forming diagonals in the RP, with an embedding dimension equal to 1. These results allow us to obtain theoretical values of three measures: (i) the recurrence rate (*REC*) (ii) the percent determinism (*DET*) and (iii) RP-based estimation of the ε -entropy $\kappa(\varepsilon)$ in the sense of correlation entropy. We apply these results to two Gaussian processes, namely first order autoregressive processes and fractional Gaussian noise. For these processes, we simulate a number of realizations and compare the RP-based estimations of the three selected measures to their theoretical values. These comparisons provide useful information on the quality of the estimations, such as the minimum required data length and threshold radius used to construct the RP.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Recurrence plots (RPs) were originally introduced by Eckmann, Kamphorst and Ruelle [1] to extract information on dynamical systems from time series. The authors showed that such plots reveal subtle structures, directly related to the dynamics generating the analyzed data. Different chaotic systems were used to illustrate the results, including systems showing non-stationary time evolution but with adiabatically varying parameters.

Few years later, Zbilut and Webber [2,3] proposed to quantify the structures displayed by recurrence plots and introduced the so-called Recurrence Quantification Analysis (RQA). Their contribution motivated many works aiming to understand the relationships between the structure of recurrence plots and different

dynamical invariants. For instance, it has been shown that measures based on diagonal lines of the recurrence plot can be used to estimate correlation entropy (Renyi entropy of order 2) [4–7]. More recently, Robinson and Thiel proved a theorem which states that the recurrences of a dynamical system determine the topology of its attractor in phase space [8]. Detailed and extensive reviews on recurrence plots and RQA, including various applications, are available in [7,9].

Although RPs and RQA were first designed to investigate deterministic nonlinear dynamics, recurrences are also related to the correlations of the analyzed data, even if they are not generated by purely deterministic processes. Indeed, it has been already reported that correlated stochastic time series can produce long diagonals in their recurrence plots [7,10,11]. Note that an observation regarding correlations in recurrence plots was first made in the original paper by Eckmann, Kamphorst and Ruelle [1], who stated that “recurrence plot describes natural (but subtle) *time correlation information*”.

* Corresponding author.

E-mail address: sofiane.ramdani@umontpellier.fr (S. Ramdani).

Rohde et al. [12] performed a stochastic analysis of the features of unthresholded squared recurrence plots (USRPs) for stationary random processes. The authors showed that the expectation value of the squared distances used to define the USRP is dependent on the variance and the covariance of the process. To illustrate their findings, they performed numerical experiments on data generated by autoregressive models, harmonic processes and a Duffing system. In this work, generic Gaussian processes were not specifically addressed, despite their omnipresence, and the statistical properties of standard thresholded RPs were not studied. In a related paper [13], Zbilut and Marwan showed that recurrence quantification can be used to derive one form of the autocovariance function and that the recurrence-based autocorrelation is naturally able to integrate dynamics in higher-dimensional phase spaces.

In the context of RP-based analysis of stochastic processes, the seminal paper by Faure and Korn [4] presented numerical experiments performed on a first order autoregressive process, confirming in this case the link between the distribution of diagonal structures and the estimation of correlation entropy. In another work, Faure and Lesne [14] analyzed Markov chains by means of an extension of RPs to symbolic sequences. More recently, Sipers, Borm and Peeters [15] explored the information contained in unthresholded recurrence plots (URPs). They showed that the information that can be extracted from an URP of a trajectory generated from a scalar signal depends on the choice of the embedding parameters (dimension and time delay) used for phase space reconstruction. The authors also established that an URP always determines the power spectrum of the original signal and that for some optimal choices of the embedding parameters, it is possible to retrieve the signal itself up to a sign. In another recent work, Grendár, Majerová and Špitalský [16] focused on some RQA outputs (recurrence rate, determinism and mean diagonal line length) and derived, in the framework of ergodic processes, their strong laws of large numbers through the concept of correlation integral. In this theoretical contribution, the authors obtained generic results and applied them to i.i.d. processes, Markov chains and autoregressive processes. They also demonstrated that the evolution of the asymptotic determinism with respect to the embedding dimension is related to the order of the analyzed Markov process. From a computational point of view, Schultz et al. [17] showed that RQA measures based on diagonal lines such as determinism can be approximated with improved algorithmic complexity. Numerical experiments included data generated by autoregressive models, the logistic map and the Lorenz attractor.

In [18], Zou et al. investigated by means of recurrence networks (RN) a class of long-term correlated stochastic processes, namely fractional Brownian motions. The authors demonstrated that the non-stationarity of such processes can lead to spurious results of RN-based analysis [19]. More specifically, using the autocorrelation function and the false nearest-neighbors method, they showed that the choice of the embedding dimension is a crucial issue for non-stationary processes. The authors also explored fractional Gaussian noise and suggested that RN analysis can provide relevant results for these stationary stochastic processes if the intrinsic parameters are properly selected.

In the present work, we explore the statistical properties of the RPs of time series generated by discrete-time stationary and correlated Gaussian stochastic processes. More specifically, we derive the probability of occurrence of recurrence points of the RP and the probability of occurrence of diagonals of a given length when no embedding is used to construct the RP. These results allow us to estimate the theoretical values of three RP-based measures: the recurrence rate, percent determinism and the ε -entropy in the sense of correlation entropy. We illustrate these results for two stationary Gaussian random processes, namely the autoregressive AR(1) process and fractional Gaussian noise. The

numerical experiments performed on the simulations of these processes and the comparison of the theoretical values of the RP-based measures to their empirical estimations provide information on the minimal length of the data and input parameters to improve the reliability of these estimations.

2. Recurrence plots and recurrence quantification analysis

The construction of a RP is originally based on phase space reconstruction obtained through time delay embedding [20–22]. Suppose that (x_i) is a simulated or experimental time series of length N . The first step of the procedure consists in the construction of the $N_v = [N - (d - 1)\tau]$ time delay vectors defined by

$$\mathbf{x}_i = (x_i, x_{i+\tau}, x_{i+2\tau}, \dots, x_{i+(d-1)\tau})^T \quad (1)$$

where d is the embedding dimension and τ the time delay. Note that τ is an integer which corresponds to a physical time lag when multiplied by the sampling period in the case of real-world data.

The second step is to estimate the distances between all the couples of vectors \mathbf{x}_i and \mathbf{x}_j . Given a threshold ε , it is possible to define a bidimensional binary plot: the recurrence plot, in which the point (i, j) is represented by a “black” dot if the distance between \mathbf{x}_i and \mathbf{x}_j is smaller than ε . \mathbf{x}_i and \mathbf{x}_j are then called neighbors. The point (i, j) is called a recurrence point. In the opposite case, no dot is plotted (see [7,9] for more details).

This can be formalized by the following binary function

$$\omega_{i,j}(\varepsilon) = \Theta(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|) \quad (2)$$

for $(i, j) \in \{1, 2, \dots, N_v\}^2$, using the Heaviside function Θ and a norm $\|\cdot\|$ (in general, the Euclidean or maximum norms are used).

Several measures can be extracted from the structure of a RP [7,23]. We will focus here on three specific quantifications. The first one is the recurrence rate, denoted *REC*, which is simply obtained by counting the number of recurrence points and quantifying it as a rate with respect to the total number of available points on the RP [7,23]. The second commonly used measure is the fraction of recurrence points belonging to diagonal lines (i.e. segments parallel to the principal diagonal given by $i = j$) of length at least n . This second key measure, denoted *DET*, was introduced as the *percent determinism*. Denoting $J_k(\varepsilon)$ the number of diagonal lines of length exactly k , the *DET* measure for diagonals of length at least n can be defined by [7,23]

$$DET(\varepsilon, n) = \frac{\sum_{k=n}^{N_v} k J_k(\varepsilon)}{\sum_{k=1}^{N_v} k J_k(\varepsilon)}. \quad (3)$$

We should underline that *DET* is not a measure of determinism in the mathematical sense [10]. In the case of stochastic processes, it has been reported that *DET* is related to the correlations within the data [10].

Remark 1. In the present work, we numerically define the RQA measures of a stochastic process by constructing the RP of each individual sample path, performing the RQA, and then averaging the resulting quantities. This choice allows us to compare empirical estimations to theoretical values of RQA measures. It should be noted that potential alternative approaches could be the construction of a single RP defined as the average of the RPs of different realizations of the process (this would yield a matrix of recurrence frequencies) or constructing the RP from the average distances $\|\mathbf{x}_i - \mathbf{x}_j\|$.

Another important measure that is related to the diagonals within the RP is an entropy measure. Indeed, it has been shown that correlation entropy can be estimated from the RP through the distribution of the lengths of the diagonals [4–7,23]. Unlike the classical Grassberger and Procaccia approach, the RP-based estimation can be performed using only one embedding dimension [4,24]. More specifically, denoting $\kappa(\varepsilon)$ the corresponding ε -entropy, one can show that the number $\phi(\varepsilon, n)$ of diagonal lines of total length at least n asymptotically obeys the following scaling [4,24]

$$\phi(\varepsilon, n) \propto e^{-n\kappa(\varepsilon)}. \quad (4)$$

Although only approximate in a finite RP, this scaling relationship provides an estimation method of the ε -entropy, computed as the slope of the semi-log representation of $\phi(\varepsilon, n)$ as a function of the length n in its linear region. We denote $\kappa_\varepsilon(\varepsilon)$ this RP-based estimation. It a priori depends on the RP-size.

3. Statistical analysis of the RPs of discrete-time Gaussian processes

We consider a real-valued, discrete time, wide-sense stationary, centered Gaussian stochastic process \mathbf{x} described by a sequence of random variables $x_1, x_2, \dots, x_k, \dots$. In order to simplify notations, we denote the random variables and their realizations using the same letters and symbols.

Within this context, the time delay vector \mathbf{x}_i defined by Eq. (1) can be considered as a d -dimensional random vector. Its probability density function (pdf) is a multivariate Gaussian function [25–29] given by

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma^i|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{x}^T (\Sigma^i)^{-1} \mathbf{x}\right) \quad (5)$$

where $|\Sigma^i|$ is the determinant of the $(d \times d)$ matrix Σ^i , which is the covariance matrix of the process defined by

$$\Sigma_{k,l}^i = \langle x_{i+(k-1)\tau} x_{i+(l-1)\tau} \rangle \quad (6)$$

for $(k, l) \in \{1, 2, \dots, d\}^2$. The angular brackets $\langle \cdot \rangle$ denote expectation or ensemble averaging.

In order to analyze the statistical properties of the RP defined by Eq. (2), we consider the random vector $\mathbf{y}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$. This new d -dimensional Gaussian vector has the covariance matrix $\Delta^{i,j}$ defined by

$$\Delta_{k,l}^{i,j} = \langle (x_{i+(k-1)\tau} - x_{j+(k-1)\tau})(x_{i+(l-1)\tau} - x_{j+(l-1)\tau}) \rangle \quad (7)$$

for $(k, l) \in \{1, 2, \dots, d\}^2$. Note that the random vector $\mathbf{y}_{i,j}$ is Gaussian because it is composed of differences of components of a Gaussian process \mathbf{x} , which are by definition joint normal [27,29].

Hence, defining the matrix $\Sigma^{i,j}$ by the elements

$$\Sigma_{k,l}^{i,j} = \langle x_{i+(k-1)\tau} x_{j+(l-1)\tau} \rangle \quad (8)$$

we can rewrite the elements of $\Delta^{i,j}$

$$\Delta_{k,l}^{i,j} = \Sigma_{k,l}^i - \Sigma_{k,l}^{i,j} - \Sigma_{k,l}^{j,i} + \Sigma_{k,l}^j. \quad (9)$$

This last equation leads to the matrix relationship

$$\Delta^{i,j} = \Sigma^i - \Sigma^{i,j} - \Sigma^{j,i} + \Sigma^j. \quad (10)$$

The pdf associated to $\mathbf{y}_{i,j}$ is then given by

$$p_{i,j}(\mathbf{y}) = \frac{1}{(2\pi)^{d/2} |\Delta^{i,j}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{y}^T (\Delta^{i,j})^{-1} \mathbf{y}\right). \quad (11)$$

This pdf allows us to compute the probability $P_{i,j}$ of the occurrence of a recurrence point at the location (i, j) on the RP constructed with a d -dimensional embedding.

3.1. Computation of the probability $P_{i,j}$ of the occurrence of a recurrence point

The probability to have \mathbf{x}_j within a distance ε from \mathbf{x}_i in the reconstructed phase space can be obtained through the estimation of the d -dimensional integral defined by

$$P_{i,j}(\varepsilon) = \int_{\mathbb{D}(\varepsilon)} p_{i,j}(\mathbf{y}) d\mathbf{y} \quad (12)$$

where the domain $\mathbb{D}(\varepsilon)$ is defined by $\mathbb{D}(\varepsilon) = \{\mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\| \leq \varepsilon\}$.

In the general case ($d \geq 2$), this integral can be estimated numerically using the algorithms proposed by Genz [30] or Sheil and O'Muircheartaigh [31], depending on the norm used (maximum or Euclidean norm respectively).

If we consider the one-dimensional case ($d = 1$ and $\tau = 1$), the matrix $\Delta^{i,j}$ is reduced to a scalar $\alpha_{i,j}$ given by

$$\alpha_{i,j} = \langle x_i^2 \rangle - 2\langle x_i x_j \rangle + \langle x_j^2 \rangle. \quad (13)$$

Note that, in the case of a wide-sense stationary and centered process \mathbf{x} , we can express $\alpha_{i,j}$ in terms of its variance σ_x^2 and covariance function C_x :

$$\alpha_{i,j} = 2[\sigma_x^2 - C_x(i-j)]. \quad (14)$$

A similar result was obtained in [12] for the computation of the expected value of the squared distance between two variables of a stochastic process (for $d = 1$). This expression was also reported in [13] as the average of the squared distances between two samples of a time series (also in the case $d = 1$).

According to Eq. (12), in the one-dimensional case, the probability $P_{i,j}(\varepsilon)$ reads

$$P_{i,j}(\varepsilon) = \frac{1}{\sqrt{2\pi\alpha_{i,j}}} \int_{-\varepsilon}^{+\varepsilon} \exp\left(-\frac{y^2}{2\alpha_{i,j}}\right) dy. \quad (15)$$

Simple manipulations lead to an expression involving the error function

$$P_{i,j}(\varepsilon) = \operatorname{erf}\left(\frac{\varepsilon}{\sqrt{2\alpha_{i,j}}}\right) \quad (16)$$

which holds when $i \neq j$. Obviously, if $i = j$ we always have $P_{i,j}(\varepsilon) = 1$.

In the case of a centered white Gaussian noise with variance σ^2 , we have $\alpha_{i,j} = 2\sigma^2$ and we get $P_{i,j}(\varepsilon) = \operatorname{erf}\left(\frac{\varepsilon}{2\sigma}\right)$, for any location (i, j) on the RP. This specific result was reported by Thiel et al. [32] in a study dedicated to the influence of observational white noise on RQA outcomes.

As previously underlined, due to the stationarity of the considered Gaussian process, $\alpha_{i,j}$ only depends on the difference $|i - j|$ in the general case (see Eq. (14)). Thus, as justified by Eq. (16), this observation also holds for the probability $P_{i,j}(\varepsilon)$ through its dependence on the covariance of the process. It should be noted that it is possible to make the assumption that $P_{i,j}(\varepsilon)$ is approximately independent of (i, j) for $|i - j| \gg 1$, that is for points (i, j) that are not too close to the main diagonal. According to Eqs. (11) and (12), it amounts to assume that the covariance of the process is weakly dependent on the difference $|i - j|$, for $|i - j| \gg 1$. The numerical experiments will confirm this assumption for the simulated processes. When $|i - j| \gg 1$, the probability $P_{i,j}(\varepsilon)$ provides an estimation of the theoretical recurrence rate $REC_{th}(\varepsilon)$. The numerical experiments presented below will show that this result already holds for $|i - j|$ larger than a few tens. Note that in [12], a similar remark was made concerning the effect of the difference $|i - j|$ on the estimation of the expected value of the squared distances $\|\mathbf{x}_i - \mathbf{x}_j\|^2$ to analyze USRPs, using an Euclidean norm.

Concerning the theoretical recurrence rate $REC_{th}(\varepsilon)$, it may be noted that the corresponding measure in the context of recurrence networks is the so-called *edge density* [33–35].

3.2. Computation of the probability $P_{i,j}^n$ of the occurrence of a diagonal of length n starting from point (i, j)

Here, we consider the probability $P_{i,j}^n(\varepsilon)$ to have n consecutive recurrence points forming a diagonal starting from point (i, j) on the recurrence plot constructed without embedding ($d = 1$) from data generated by the discrete-time Gaussian process \mathbf{x} .

In the one-dimensional case, this probability corresponds to the joint events defined by the inequalities $|x_i - x_j| \leq \varepsilon$, $|x_{i+1} - x_{j+1}| \leq \varepsilon$, \dots , $|x_{i+n-1} - x_{j+n-1}| \leq \varepsilon$. Hence, if we define a new random vector $\mathbf{z}_{i,j}^n$ by

$$\mathbf{z}_{i,j}^n = (x_i - x_j, x_{i+1} - x_{j+1}, \dots, x_{i+n-1} - x_{j+n-1})^T \quad (17)$$

it is straightforward to see that $P_{i,j}^n(\varepsilon)$ is the probability to have $\|\mathbf{z}_{i,j}^n\|_\infty \leq \varepsilon$, where $\|\cdot\|_\infty$ is the maximum norm.

The covariance matrix $\Omega^{i,j}$ of the n -dimensional random Gaussian vector $\mathbf{z}_{i,j}^n$ is defined by

$$\Omega_{k,l}^{i,j} = \langle (x_{i+k-1} - x_{j+k-1})(x_{i+l-1} - x_{j+l-1}) \rangle \quad (18)$$

for $(k, l) \in \{1, 2, \dots, n\}^2$.

This covariance matrix can be decomposed as done for $\Delta^{i,j}$ in the previous subsection. Indeed, if we denote for $(k, l) \in \{1, 2, \dots, n\}^2$

$$\Lambda_{k,l}^{i,j} = \langle x_{i+k-1} x_{i+l-1} \rangle \quad (19)$$

and

$$\Lambda_{k,l}^{i,j} = \langle x_{i+k-1} x_{j+l-1} \rangle \quad (20)$$

we can write the matrix relationship

$$\Omega^{i,j} = \Lambda^i - \Lambda^{i,j} - \Lambda^{i,j} + \Lambda^j. \quad (21)$$

Hence, the pdf associated to the random vector $\mathbf{z}_{i,j}^n$ is given by

$$p_{i,j}^n(\mathbf{z}) = \frac{1}{(2\pi)^{n/2} |\Omega^{i,j}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{z}^T (\Omega^{i,j})^{-1} \mathbf{z}\right). \quad (22)$$

Finally, the probability $P_{i,j}^n(\varepsilon)$ to have n consecutive recurrence points starting from a point (i, j) on the RP can be written

$$P_{i,j}^n(\varepsilon) = \int_{\mathbb{M}(\varepsilon)} p_{i,j}^n(\mathbf{z}) d\mathbf{z} \quad (23)$$

with $\mathbb{M}(\varepsilon) = \{\mathbf{z} : \mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\|_\infty \leq \varepsilon\}$.

If we set $n = 1$, we obtain the case of a single recurrence point as we have $P_{i,j}^{n=1}(\varepsilon) = P_{i,j}(\varepsilon)$. For $n \geq 2$, $P_{i,j}^n(\varepsilon)$ can be estimated numerically using the approach described by Genz [30]. The computation is based on a Cholesky decomposition of the matrix $\Omega^{i,j}$ and a classical Monte-Carlo estimation method. The details of this approach are described in the [Appendix](#).

In the case of a centered white Gaussian noise with variance σ^2 , it is clear that $P_{i,j}^n(\varepsilon) = \left[\text{erf}\left(\frac{\varepsilon}{2\sigma}\right)\right]^n$, for $i \neq j$. This result was also reported in [32].

3.3. Computation of the probability $Q_{i,j}^n$ of the occurrence of a diagonal of length exactly n

The computation of this probability is motivated by the DET measure and RP-based entropy estimations, which will be detailed below. The probability $Q_{i,j}^n(\varepsilon)$ to have exactly n consecutive recurrence points forming a diagonal of length exactly n , starting from a point (i, j) on the one-dimensional embedding recurrence plot, can be deduced from the knowledge of $P_{i,j}^n(\varepsilon)$.

In the generic case and without any assumption on the nature of the pdf of the stochastic process, it reads as follows

$$Q_{i,j}^n(\varepsilon) = [P_{i,j}^n(\varepsilon) - P_{i,j}^{n+1}(\varepsilon)] - [P_{i-1,j-1}^{n+1}(\varepsilon) - P_{i-1,j-1}^{n+2}(\varepsilon)] \quad (24)$$

where $P_{i,j}^n(\varepsilon)$ is given by Eq. (23).

In the right-hand side of Eq. (24), the first term ensures that the diagonal is not longer than n beyond (i, j) , and removing the second term ensures that it actually starts in (i, j) . By removing the brackets, one can also see that the probability $P_{i,j}^{n+1}(\varepsilon)$ is subtracted from $P_{i,j}^n(\varepsilon)$ to ensure that the point $(i+n, j+n)$ is not a recurrence point. Then, the probability $P_{i-1,j-1}^{n+1}(\varepsilon)$ is subtracted so that the point $(i-1, j-1)$ is also a non-recurrence point. Finally, the probability $P_{i-1,j-1}^{n+2}(\varepsilon)$ is added because it is associated to an event, which is included in the two events corresponding to the probabilities $P_{i,j}^{n+1}(\varepsilon)$ and $P_{i-1,j-1}^{n+1}(\varepsilon)$. Note that $Q_{i,j}^n(\varepsilon)$ is the probability that (i, j) is the starting point of a diagonal line of length exactly n , that is, precluding the possibility of a backward diagonal fragment starting from (i, j) . The space of possible states is binary (“is the starting point” or “is not the starting point”). Presumably, this quantity decreases when n increases.

In the case of a centered white Gaussian noise with variance σ^2 , we have $P_{i,j}^n(\varepsilon) = \left[\text{erf}\left(\frac{\varepsilon}{2\sigma}\right)\right]^n$ and the expression of the probability $Q_{i,j}^n(\varepsilon)$ given by Eq. (24) is simplified and can be written

$$Q_{i,j}^n(\varepsilon) = \left[\text{erf}\left(\frac{\varepsilon}{2\sigma}\right)\right]^n \left[1 - \text{erf}\left(\frac{\varepsilon}{2\sigma}\right)\right]^2. \quad (25)$$

This result confirms the findings reported in [32] for the analysis of the RP of white Gaussian noise.

3.4. Computation of the probability $R_{i,j}^n$ that (i, j) is the starting point of a diagonal of length at least n and the theoretical ε -entropy $\kappa(\varepsilon)$

As previously described in Section 2, the RP-based estimation of the ε -entropy $\kappa_\varepsilon(\varepsilon)$ (in the sense of correlation entropy) is obtained through the distribution of the diagonal of length at least n . Here, we derive the probability $R_{i,j}^n(\varepsilon)$ to find a diagonal of length at least n starting from point (i, j) .

Since $Q_{i,j}^k(\varepsilon)$ is the probability to have exactly k consecutive recurrence points forming a diagonal of length exactly k , starting from a point (i, j) , we can write

$$R_{i,j}^n(\varepsilon) = \sum_{k \geq n} Q_{i,j}^k(\varepsilon). \quad (26)$$

Here again, the space of possible states is binary. This quantity obviously decreases when n increases (besides, whatever the behavior of $Q_{i,j}^k(\varepsilon)$). It is to note that $R_{i,j}^{n=1}(\varepsilon)$ is not the probability that (i, j) is recurrence point, due to the condition of being the starting point of a line, which amounts to the condition that $(i-1, j-1)$ is not a recurrence point.

The probability $Q_{i,j}^n(\varepsilon)$ being calculated through the probability $P_{i,j}^n(\varepsilon)$ (see Eq. (24)), the computation of $R_{i,j}^n(\varepsilon)$ can be simplified by making an assumption, which is similar to the assumption made in Section 3.1 for the computation of the probability $P_{i,j}(\varepsilon)$. The assumption is that $P_{i,j}^n(\varepsilon)$ is approximately independent of (i, j) for $|i-j| \gg 1$, which corresponds to locations on the RP that are not too close to the main diagonal. It amounts to assume that the divergence of trajectories (here originating from (i, j)) is statistically homogeneous in the phase space. $P_{i,j}^n(\varepsilon)$ is the probability that (i, j) is the origin of a forward diagonal fragment of length n , without any assumption on the points $(i-1, j-1)$, $(i-2, j-2)$, \dots , at the previous time steps. In other words, the complete diagonal line to which (i, j) belongs could be longer than n .

Practically and for the processes included in the numerical experiments, we will see in the next section that this assumption always holds for $|i-j|$ larger than a few tens.

Going back to the result of Eq. (24), and assuming that $P_{i,j}^n(\varepsilon) = P^n(\varepsilon)$ is independent of (i, j) , we get a simplified expression of the probability $Q^n(\varepsilon)$ of the occurrence of a diagonal of length exactly n

$$Q^n(\varepsilon) \simeq P^n(\varepsilon) - 2P^{n+1}(\varepsilon) + P^{n+2}(\varepsilon). \quad (27)$$

Note that a similar relationship holds also for the so-called cumulative and noncumulative distributions of diagonal lines in the context of chaotic dynamics [36]. Starting from this result, we can write an approximation $R^n(\varepsilon)$ of $R_{i,j}^n(\varepsilon)$, for which we also assume an independence with respect to (i, j) . The probability $R^n(\varepsilon)$ of the occurrence of a diagonal of length at least n starting from a given point, can be obtained by cumulatively summing for $k \geq n$ the probabilities $Q^k(\varepsilon)$ to have diagonals of length exactly k starting from this given point

$$R^n(\varepsilon) = \sum_{k \geq n} Q^k(\varepsilon). \quad (28)$$

Then, by replacing $Q^k(\varepsilon)$ according to Eq. (27), we obtain

$$R^n(\varepsilon) \simeq \sum_{k \geq n} P^k(\varepsilon) - 2P^{k+1}(\varepsilon) + P^{k+2}(\varepsilon). \quad (29)$$

It is then straightforward to see that an approximation of $R^n(\varepsilon)$ reads

$$R^n(\varepsilon) \simeq P^n(\varepsilon) - P^{n+1}(\varepsilon). \quad (30)$$

Theoretically, as it is the case for $\phi(\varepsilon, n)$, $R^n(\varepsilon)$ has an exponential scaling with respect to the length n of the diagonal lines [14]. We may note that an equivalent relationship can be derived for the quantities corresponding to number of diagonals in the case of RPs of symbolic sequences [14,24]. Thus, $R^n(\varepsilon)$ provides a theoretical approach for the computation of the ε -entropy $\kappa(\varepsilon)$ through its semi-log representation as a function of n , as we have

$$R^n(\varepsilon) \propto e^{-n\kappa(\varepsilon)}. \quad (31)$$

This scaling relationship provides a way to compute the ε -entropy $\kappa(\varepsilon)$ from the analytically derived expression of $R^n(\varepsilon)$. In Section 4, the value of $\kappa(\varepsilon)$ obtained by this theoretical approach will be compared for two Gaussian processes to the RP-based estimation $\kappa_e(\varepsilon)$ described in Section 2. In both cases, the result is approximate due to the estimation of the exponent from a limited number of values of n . In the theoretical approach, an additional approximation is involved in the analytical computation of $R^n(\varepsilon)$ (see Eqs. (29) and (30)). In the RP-based case, additional sampling effects are involved due to the finite-size of the RP (i.e. finite length of the observation) and finite number of realizations.

3.5. Computation of the theoretical percent determinism DET_{th}

The theoretical percent determinism $DET_{th}(\varepsilon, n)$ can be derived from the probability $Q^n(\varepsilon)$ of the occurrence of a diagonal of length exactly n . Here, we estimate $Q^n(\varepsilon)$ from the probability $Q_{i,j}^n(\varepsilon)$ given by Eq. (24), by assuming once again that this quantity is independent from (i, j) when $|i - j|$ is large enough.

According to the definition of the percent determinism given by Eq. (3) and after a normalization of the numerator and denominator of the right-hand side of this equation, we can write

$$DET_{th}(\varepsilon, n) = \frac{\sum_{k \geq n} kQ^k(\varepsilon)}{\sum_{k \geq 1} kQ^k(\varepsilon)}. \quad (32)$$

In the following section, the theoretical values $DET_{th}(\varepsilon, n)$ will be compared to the estimated percents determinism obtained from simulated series of the considered Gaussian processes.

4. Numerical experiments

In this section, we take as a basis the analytical expressions derived above for the recurrence rate, the percent determinism and the ε -entropy to investigate the practical use and quality of the RP-based estimates of these quantities. In particular, we explore the sampling effects (finite duration N), taking as a benchmark two Gaussian processes: the AR(1) process and fractional Gaussian noise (fGn).

4.1. General numerical procedure

For each of the considered processes, AR(1) and fGn, we numerically estimate the quantities $P_{i,j}(\varepsilon)$, $P_{i,j}^n(\varepsilon)$, $Q_{i,j}^n(\varepsilon)$ involved in the computation of the theoretical recurrence rate $REC_{th}(\varepsilon)$, percent determinism $DET_{th}(\varepsilon, n)$ and ε -entropy $\kappa(\varepsilon)$. For each process, we analyze two cases corresponding to different values of the process parameters. All the theoretical probabilities involved in the estimations of $REC_{th}(\varepsilon)$, $DET_{th}(\varepsilon, n)$ and $\kappa(\varepsilon)$ were computed for $|i - j| = 100$. The choice of this value is justified by a preliminary study of the dependence of $P_{i,j}(\varepsilon)$ with respect to $|i - j|$ for different values of the threshold ε , as well as a similar study for $P_{i,j}^n(\varepsilon)$ with $n = 2$ and 4.

Then, we compare the theoretical recurrence rates $REC_{th}(\varepsilon)$ to the statistics of recurrence rates numerically obtained from the simulated processes by using 50 realizations of 3 different lengths: $N = 1000, 500$ and 250 time points, and ε values ranging from 0.2 to 1.2, with a step of 0.2. As mentioned in Remark 1, the RP-based estimations are obtained in the same way for all RQA measures, by constructing the RP of each individual sample path and then averaging the resulting quantities over the sample paths.

The theoretical percent determinism $DET_{th}(\varepsilon, n)$ is then compared for $n = 2, 3$ and 4 to its empirical counterpart for each process, using the same numbers of realizations, data lengths and ε values as those considered for the recurrence rate. For the computation of $DET_{th}(\varepsilon, n)$ from Eq. (32), we set the maximal value of k to 10 ($k \leq 10$), which was an appropriate choice for the range of explored ε values (0.2 to 1.2).

Finally, according to the results of Section 3.4, we compare the RP-based estimation $\kappa_e(\varepsilon)$ to theoretical value $\kappa(\varepsilon)$ derived from the analytically computed expression of $R^n(\varepsilon)$, for ε values ranging from 0.1 to 2.0 with a step of 0.1. The computation of the theoretical ε -entropy $\kappa(\varepsilon)$ was performed through the semi-log representation of $R^n(\varepsilon)$ (see Eq. (30)) in its linear region. The RP-based estimations $\kappa_e(\varepsilon)$ were obtained for 50 realizations of AR(1) and fGn processes, using 4 different lengths $N = 500, 250, 125$ and 75. In this case, we used the semi-log representation of the RP-based histogram $\phi(\varepsilon, n)$ (see Eq. (4)) in its linear region.

To compute RQA measures of the simulated data, we used the Cross Recurrence Plot Toolbox developed by N. Marwan [37]. No Theiler window [22,7] was used for the computations performed on these data.

4.2. First order autoregressive process AR(1)

We consider the centered and stationary Gaussian stochastic process AR(1) [38] defined by

$$x_i = \varphi x_{i-1} + \eta_i \quad (33)$$

where $|\varphi| < 1$ and η is a zero-mean white Gaussian noise with variance σ^2 . Note that the AR(1) process can be seen as the discrete-time counterpart to the Ornstein–Uhlenbeck process.

The variance of this process is given by

$$\langle x_i^2 \rangle = \frac{\sigma^2}{1 - \varphi^2}. \quad (34)$$

The covariance of the process is given by

$$\langle x_i x_j \rangle = \frac{\sigma^2}{1 - \varphi^2} \varphi^{|i-j|}. \quad (35)$$

Thus, we can express the quantity $\alpha_{i,j}$ given by Eq. (14) as follows

$$\alpha_{i,j} = 2 \frac{\sigma^2}{1 - \varphi^2} (1 - \varphi^{|i-j|}). \quad (36)$$

The probability $P_{i,j}(\varepsilon)$ of the occurrence of a recurrence point can then be obtained according to Eq. (16)

$$P_{i,j}(\varepsilon) = \text{erf} \left(\frac{\varepsilon}{2\sigma} \sqrt{\frac{1 - \varphi^2}{1 - \varphi^{|i-j|}}} \right). \quad (37)$$

According to the results of Section 3.2, the probability $P_{i,j}^n(\varepsilon)$ to have a diagonal of n consecutive recurrence points starting from a point (i, j) can be obtained through the integration of the n -variate Gaussian pdf $p_{i,j}^n(\mathbf{z})$, given by Eq. (22), over the domain $\mathbb{M}(\varepsilon) = \{\mathbf{z} : \mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\|_\infty \leq \varepsilon\}$. We recall that the associated covariance matrix $\Omega^{i,j}$ is defined by $\Omega^{i,j} = \Lambda^i - \Lambda^{i,j} - \Lambda^{j,i} + \Lambda^j$ (see Eq. (21)). For an AR(1) process with covariance given by Eq. (35), we obtain

$$\Lambda_{k,l}^i = \langle x_{i+k-1} x_{i+l-1} \rangle = \frac{\sigma^2}{1 - \varphi^2} \varphi^{|k-l|} \quad (38)$$

which is independent of i , and

$$\Lambda_{k,l}^{i,j} = \langle x_{i+k-1} x_{j+l-1} \rangle = \frac{\sigma^2}{1 - \varphi^2} \varphi^{|i-j+k-l|} \quad (39)$$

for $(k, l) \in \{1, 2, \dots, n\}^2$.

Using these two expressions and for fixed parameters φ and σ and a given diagonal length n , it is possible, through Eq. (23), to compute $P_{i,j}^n(\varepsilon)$ for different values of the absolute difference $|i-j|$. We will show below that $P_{i,j}^n(\varepsilon)$ is almost independent of $|i-j|$ for $|i-j|$ large enough. After selecting a sufficiently large value for $|i-j|$, we can compute $R^n(\varepsilon)$ and $\kappa(\varepsilon)$ from the results given by Eqs. (30) and (31). Using expression (32), we can also obtain the value of $DET_{th}(\varepsilon, n)$. These computations and the comparison to the RP-based estimations were performed for AR(1) processes with $\varphi = 0.1$ and $\varphi = 0.9$ (and $\sigma = 1$, see Eq. (33)).

4.3. Fractional Gaussian noise (fGn)

Fractional Gaussian noise (fGn) is often defined as the increment process of fractional Brownian motion (fBm) [39,40], which is a Gaussian and self-similar process with stationary increments. fGn is a centered, stationary and Gaussian process. We consider here the discrete-time case, for which the autocovariance function is given by

$$\gamma(k) = \frac{\sigma^2}{2} (|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \quad (40)$$

where H is the Hurst exponent ($0 < H < 1$) and σ^2 is the variance of the process. We will set $\sigma^2 = 1$ for the simulations.

To simulate sample paths of fGn, we used the classical approach based on the Cholesky decomposition of the covariance matrix [40], which (i, j) entry is given by $C_f(i-j) = \gamma(|i-j|)$.

From Eq. (14), the quantity $\alpha_{i,j}$ reads

$$\alpha_{i,j} = 2(1 - C_f(i-j)). \quad (41)$$

For fGn, the probability $P_{i,j}(\varepsilon)$ can thus be obtained from Eq. (16)

$$P_{i,j}(\varepsilon) = \text{erf} \left(\frac{\varepsilon}{2\sqrt{1 - C_f(i-j)}} \right) \quad (42)$$

In the case of fGn, the covariance matrix $\Omega^{i,j}$ involved in the computation of $P_{i,j}^n(\varepsilon)$ is also defined by $\Omega^{i,j} = \Lambda^i - \Lambda^{i,j} - \Lambda^{j,i} + \Lambda^j$ (see Eq. (21)), where

$$\Lambda_{k,l}^i = \langle x_{i+k-1} x_{i+l-1} \rangle = C_f(k-l) \quad (43)$$

which is independent of i , and

$$\Lambda_{k,l}^{i,j} = \langle x_{i+k-1} x_{j+l-1} \rangle = C_f(i-j+k-l) \quad (44)$$

for $(k, l) \in \{1, 2, \dots, n\}^2$.

As for the AR(1) process, the numerical results will show that $P_{i,j}^n(\varepsilon)$ is almost independent of $|i-j|$ for $|i-j|$ large enough, allowing the analytic computation of the theoretical values of $\kappa(\varepsilon)$ and $DET_{th}(\varepsilon, n)$. These numerical experiments and the comparison with the RP-based estimations were performed for fGn with Hurst exponents $H = 0.3$ and $H = 0.7$.

Remark 2. Due to the stochastic nature of the algorithm used to compute multivariate normal integrals (see Appendix), it was relevant to quantify the variability of the theoretical values of $P_{i,j}^n(\varepsilon)$ given by Eq. (23) in order to evaluate the quality of these computations. The statistics of $P_{i,j}^n(\varepsilon)$ were computed over 50 runs of the algorithm, for $|i-j| = 100$ and for both AR(1) and fGn processes. We observed that the relative variability of $P_{i,j}^n(\varepsilon)$ (coefficient of variation CV_p) was dependent on both ε and n . The coefficient of variation CV_p was found to be increasing with ε and also when n was increased from 2 to 4. The maximal value of CV_p over ε values ranging from 0.1 to 2.0 (with a step of 0.1) was obtained in the case of AR(1) process with $\varphi = 0.9$ and $n = 4$. Its value was $CV_p = 3.6517 \times 10^{-4}$. This result ensured an efficient estimation of the theoretical values of $P_{i,j}^n(\varepsilon)$ (for $|i-j| = 100$), from its analytical expression obtained in Section 3.2 (Eq. (23)).

4.4. Dependence of $P_{i,j}$ and $P_{i,j}^n$ on $|i-j|$

We show here that $P_{i,j}(\varepsilon)$ and $P_{i,j}^n(\varepsilon)$ are very weakly dependent on $|i-j|$ when $|i-j|$ is large enough. The evolution of $P_{i,j}(\varepsilon)$ for $|i-j|$ increasing from 0 to 100 is depicted in Fig. 1 for 5 different values of ε . In all cases, $P_{i,j}(\varepsilon)$ starts from the maximal value 1 for $i = j$ and then is stabilized for $|i-j|$ large enough.

Figs. 2 and 3 show the equivalent result for probability $P_{i,j}^n(\varepsilon)$, respectively for $n = 2$ and $n = 4$. These curves confirm that these probabilities also are very weakly dependent on $|i-j|$ when this difference is larger than few tens. These observations justify the choice $|i-j| = 100$ reported in Section 4.1.

4.5. Comparison of the theoretical and empirical recurrence rates

We present here the comparison between the statistics of recurrence rates REC computed from the RPs of 50 simulated series of each process and the theoretical recurrence rates $REC_{th}(\varepsilon)$, which are computed from $P_{i,j}(\varepsilon)$ for $|i-j| = 100$. Figs. 4 and 5 show these results for 3 different lengths of the simulated series: $N = 1000, 500$ and 250 .

The results depicted in Figs. 4 and 5 show a very good accuracy of the RP-based estimation of the recurrence rates REC for AR(1) process with $\varphi = 0.1$ and both fGn processes, for all lengths of the simulated series. A moderate increase of the variability of REC is observed when the length N of the series is reduced, especially for the largest values of the threshold ε .

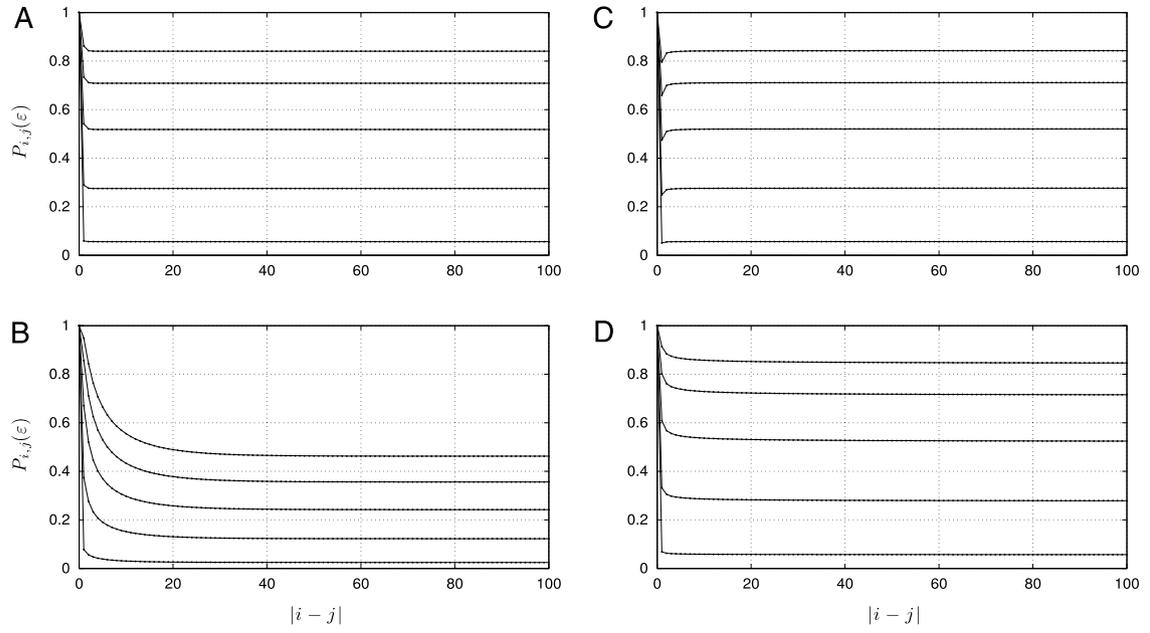


Fig. 1. The theoretical probabilities $P_{i,j}(\epsilon)$ as a function of $|i-j|$ for $\epsilon = 0.1, 0.5, 1.0, 1.5$ and 2.0 , from lower to upper curves. (A) AR(1) with $\phi = 0.1$, (B) AR(1) with $\phi = 0.9$, (C) fGn with $H = 0.3$, (D) fGn with $H = 0.7$.

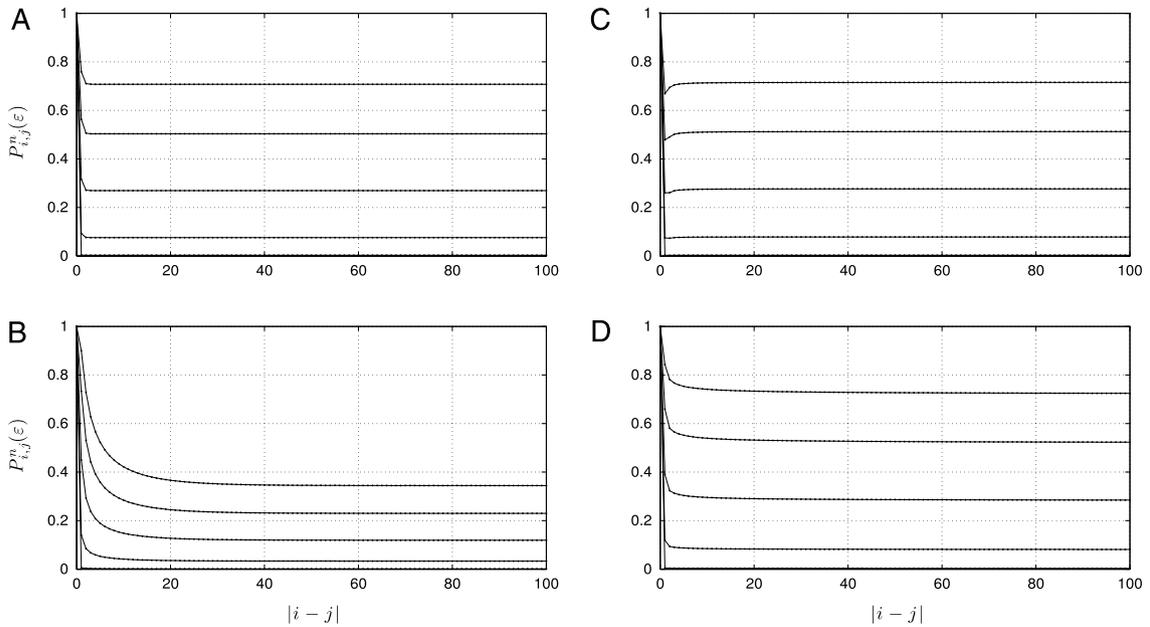


Fig. 2. The theoretical probabilities $P_{i,j}^n(\epsilon)$ as a function of $|i-j|$ for $n = 2$ and $\epsilon = 0.1, 0.5, 1.0, 1.5$ and 2.0 , from lower to upper curves. (A) AR(1) with $\phi = 0.1$, (B) AR(1) with $\phi = 0.9$, (C) fGn with $H = 0.3$, (D) fGn with $H = 0.7$. For $\epsilon = 0.1$, the curves are very close to zero when $|i-j| \geq 2$.

In the case of AR(1) process with $\phi = 0.9$ (Fig. 4, panels (D), (E) and (F)), the REC values are lower than in other cases and the increase of the variability is more important with the decrease of N . In addition, we observe a bias of the estimation for $N = 500$ and 250 , which increases with the threshold ϵ . In these cases, there is a systematic overestimation of REC.

4.6. Comparison of the theoretical and empirical percent determinism

In this subsection, we show the results of the comparison between the statistics of percent determinism DET estimated over 50 simulated series of each process and the theoretical percent

determinism $DET_{th}(\epsilon, n)$, which are computed from $P_{i,j}^n(\epsilon)$ for $|i-j| = 100$. Figs. 6 and 7 show these results for $n = 2, 3, 4$ and for 3 different lengths of the simulated series: $N = 1000, 500$ and 250 .

In all cases, we globally observe an expected increase of the variability of the estimation of percent determinism when the length of the series is decreased. The results depicted in panels (A), (B) and (C) of Fig. 6 indicate that, for the AR(1) process with $\phi = 0.1$, the estimation of DET is accurate for diagonal lines of length at least n for all the considered values of n ($n = 2, 3$ and 4), for the three considered lengths $N = 1000, 500$ and 250 points, and for $\epsilon \geq 0.4$. For $\epsilon = 0.2$, the quality of the estimation is still good for $N = 1000$ and 500 (see panels (A), (B) of Fig. 6). However, for this lowest value of ϵ and when $N = 250$ (panel (C) of Fig. 6), an overestimation is observed and it increases when n is increased.

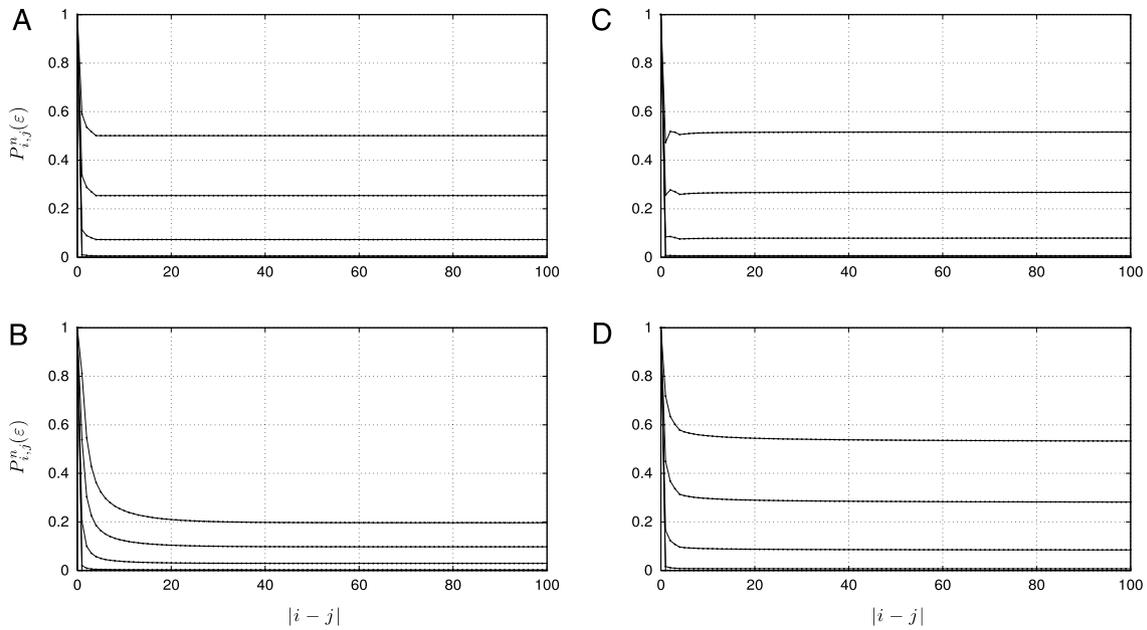


Fig. 3. Same as Fig. 2 for $n = 4$. For $\varepsilon = 0.1$ and $\varepsilon = 0.5$, the curves are very close to zero when $|i - j| \geq 2$.

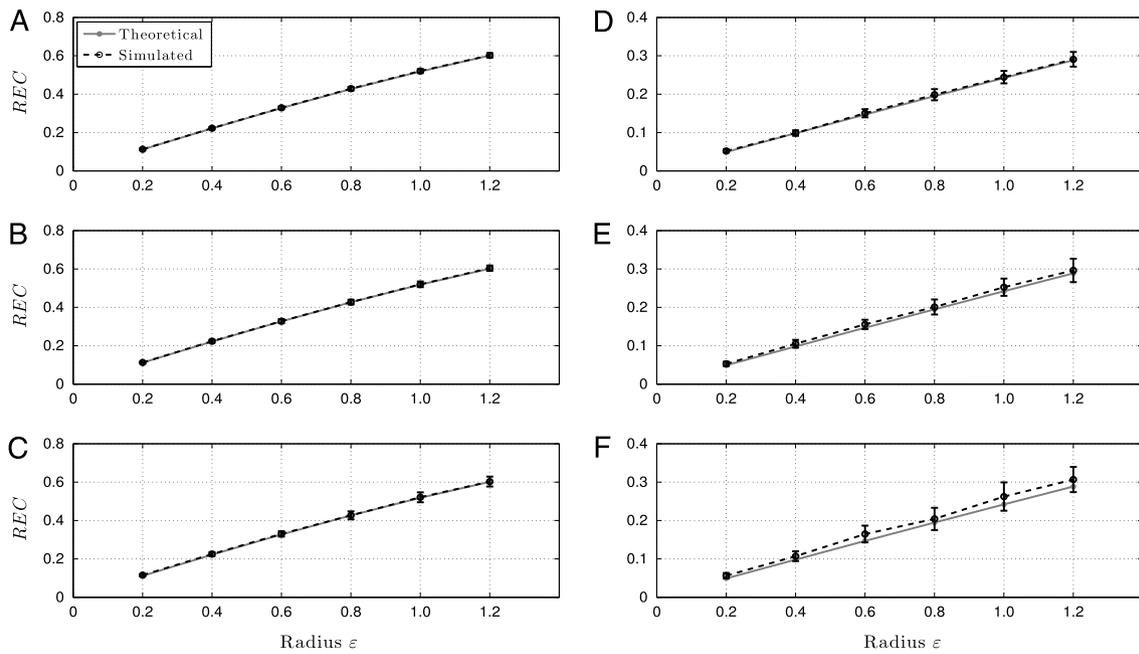


Fig. 4. The theoretical recurrence rates $REC_{th}(\varepsilon)$ compared to the mean recurrence rates obtained from 50 simulated series of the AR(1) processes. The error bars indicate standard deviations. (A) AR(1) with $\varphi = 0.1$, and $N = 1000$, (B) $N = 500$, (C) $N = 250$, (D) AR(1) with $\varphi = 0.9$, and $N = 1000$, (E) $N = 500$, (F) AR(1) $N = 250$.

In the case of AR(1) process with $\varphi = 0.9$, the results shown in panel (D) of Fig. 6 corresponding to $N = 1000$ indicate a good estimation of DET for $\varepsilon \geq 0.4$ for $n = 2, 3$ and 4 . For $\varepsilon = 0.2$, a moderate overestimation is observed and is more pronounced when n is increased. In the case $N = 500$ (see panel (E)), a similar behavior is observed but with better estimations when $\varepsilon \geq 0.6$. For $N = 250$ (see panel (F)), the estimation of DET is accurate for $n = 2$ and $\varepsilon \geq 0.6$, but a variable bias is noted for $n = 3$ and $n = 4$ for most ε values. This bias is reduced when ε is increased.

Fig. 7 shows the results obtained for fGn processes. For the two cases $H = 0.3$ (panels (A), (B) and (C)) and $H = 0.7$ (panels (D), (E) and (F)), the results are qualitatively similar. The estimations are very good in almost all cases, except for $N = 250$. In this case, the overestimation is mainly observed for $\varepsilon = 0.2$.

4.7. Comparison of the theoretical and empirical ε -entropies

We here perform the comparison of the ε -entropy $\kappa_\varepsilon(\varepsilon)$, estimated from the empirical histogram of diagonal lines, with its theoretical counterpart.

These quantities are derived respectively through the number $\phi(\varepsilon, n)$ of diagonal lines of total lengths equal or larger than n and the semi-log representations of the theoretical probability $R^n(\varepsilon)$. The former is estimated from the RPs of simulated sample paths of the considered stochastic processes. Note that we adopt here the same procedure as for the recurrence rate and percent determinism (see Remark 1), that is, we generate the RPs and histograms $\phi(\varepsilon, n)$ of each sample path, then we estimate the individual $\kappa_\varepsilon(\varepsilon)$ and then calculate their statistics over the sample paths.

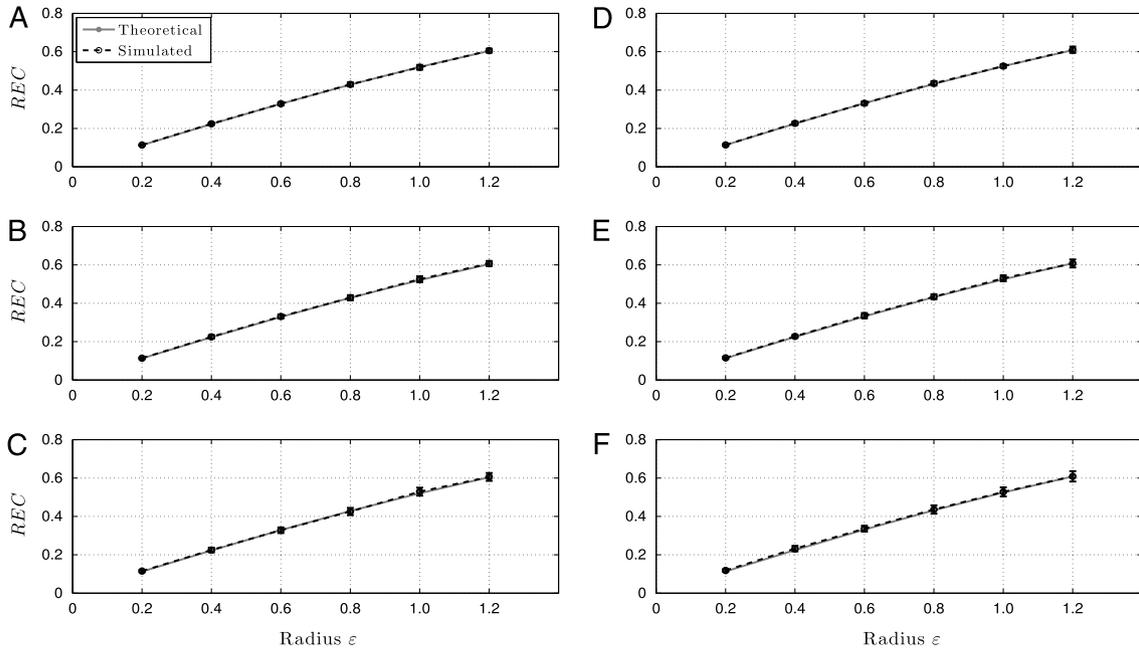


Fig. 5. REC values for fGn process (same configuration as Fig. 4). (A, B, C) with $H = 0.3$ and (D, E, F) with $H = 0.7$.

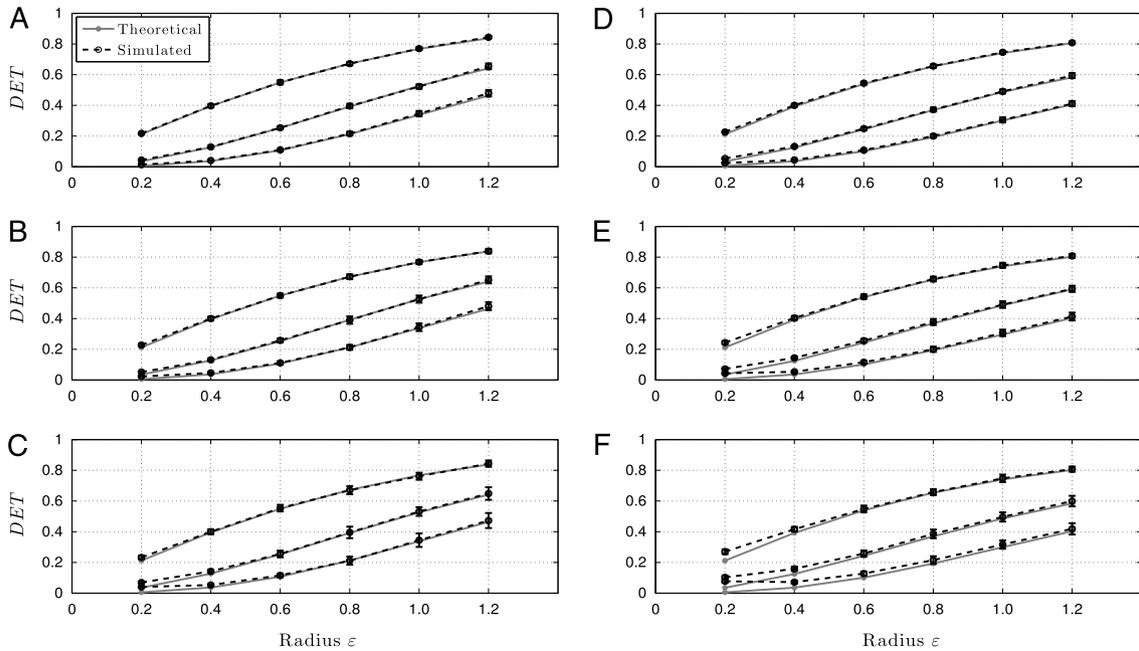


Fig. 6. The theoretical percents determinism $DET_{th}(\epsilon, n)$ compared to the mean percents determinism obtained from 50 simulated series of the AR(1) processes. In each panel, the 3 curves respectively correspond to $n = 2, n = 3$ and $n = 4$, from upper to lower curve. The error bars indicate standard deviations. (A) AR(1) with $\varphi = 0.1$, and $N = 1000$, (B) $N = 500$, (C) $N = 250$, (D) AR(1) with $\varphi = 0.9$, and $N = 1000$, (E) $N = 500$, (F) AR(1) $N = 250$.

Figs. 8–11 illustrate the RP-based estimations of $\kappa_\epsilon(\epsilon)$ from 50 simulated sample paths of length 500 for two AR(1) processes ($\varphi = 0.1$ and $\varphi = 0.9$) and two fGn processes ($H = 0.3$ and $H = 0.7$), for $\epsilon = 1.0$.

The comparison of the theoretical ϵ -entropy $\kappa(\epsilon)$ and the mean of the RP-based estimations $\kappa_\epsilon(\epsilon)$ is depicted in Fig. 12 for the AR(1) process for both $\varphi = 0.1$ (panels (A) to (D)) and $\varphi = 0.9$ (panels (E) to (H)) and for ϵ ranging from 0.1 to 2.0 with a step of 0.1. The statistics are computed over 50 simulated series of decreasing lengths $N = 500, 250, 125$, and 75 points, from upper to lower panels. The first qualitative observation is that the shape of the ϵ -entropy curve is coherent with the expected

behavior for a discrete-time Gaussian stochastic process [41,42,4]. We also observe that, for a fixed value of the threshold radius ϵ , the variability generally increases when the length N of the series is reduced, especially for small values of ϵ . In addition, for a given length, this variability is reduced when ϵ is increased. In the case $\varphi = 0.1$, the results indicate an accurate RP-based estimation of the ϵ -entropy for all lengths except for $N = 75$, but only for the two smallest values of $\epsilon, 0.1$ and 0.2 (see panel (D) of Fig. 12). When $\varphi = 0.9$, we observe a deviation from the theoretical values for $N = 125$ and $N = 75$, also for the smallest values of $\epsilon, 0.1$ and 0.2 (see panels (G) and (H) of Fig. 12). In these cases, the ϵ -entropy is underestimated when using the RP-based estimation method.

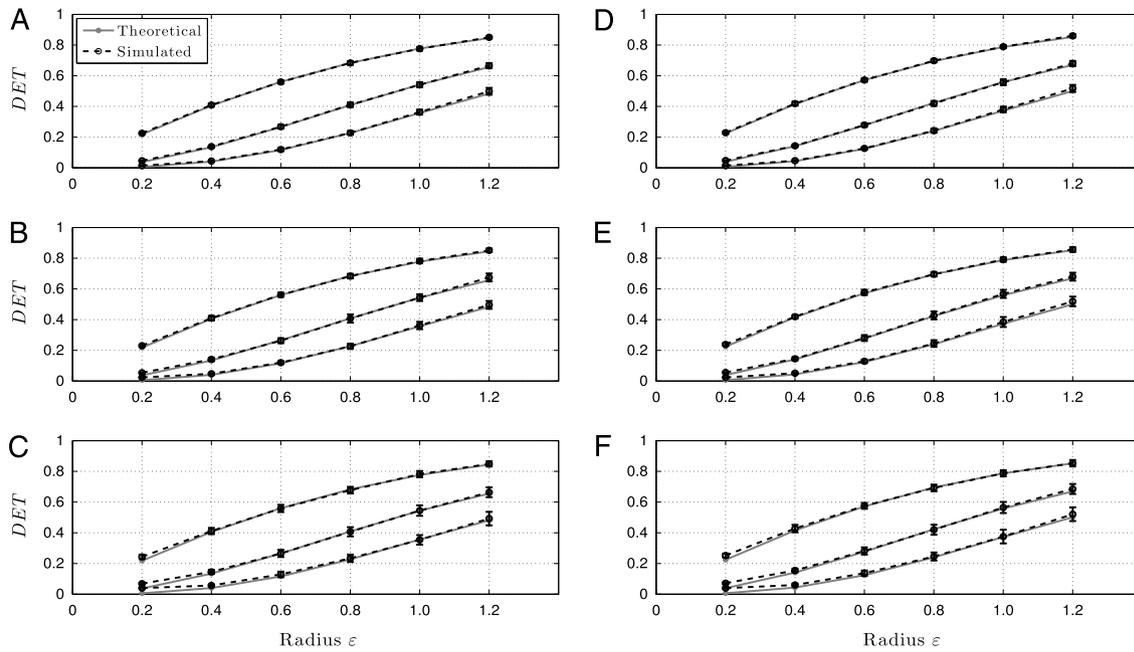


Fig. 7. DET values for fGn process (same configuration as Fig. 6). (A, B, C) with $H = 0.3$ and (D, E, F) with $H = 0.7$.

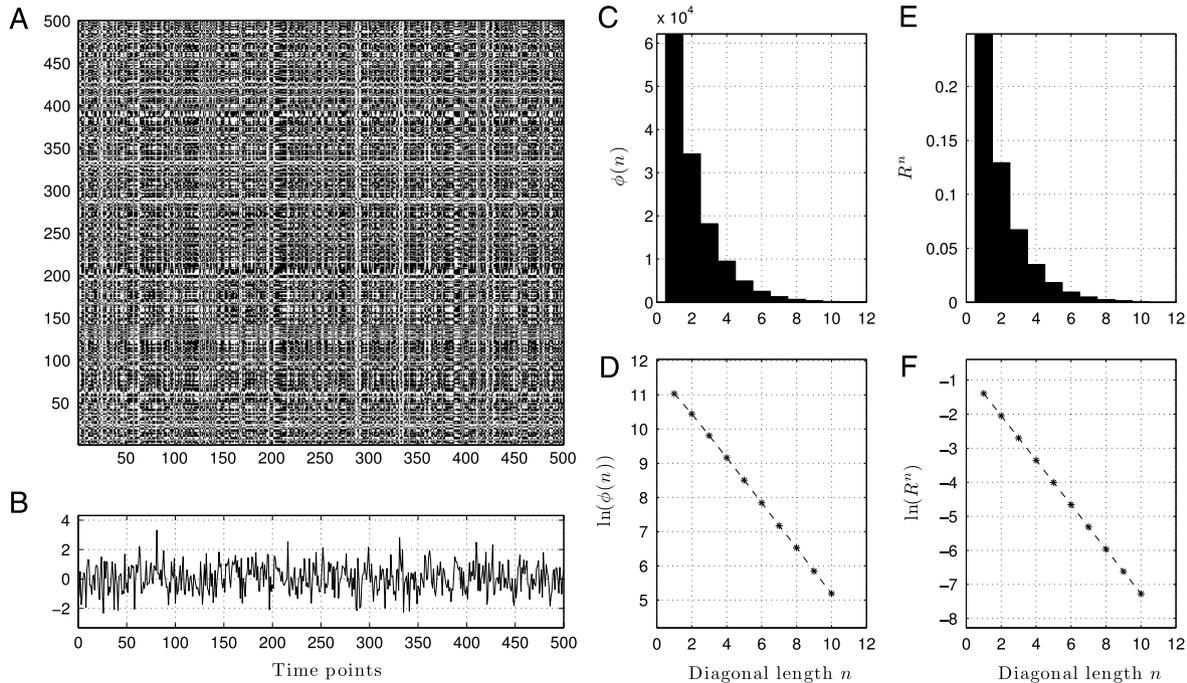


Fig. 8. RP (A) of a 500-points sample path (B) of AR(1) process with $\varphi = 0.1$, without embedding and for a threshold radius $\varepsilon = 1.0$. Panel (C) depicts the number $\phi(\varepsilon, n)$ of diagonal lines of total length equal or larger than n . In panel (D), the semi-log representation of $\phi(\varepsilon, n)$ is shown as a function of n , for the ten first values of n . Panel (E) shows the corresponding theoretical probability $R^n(\varepsilon)$ of the occurrence of a diagonal of length at least n obtained for the AR(1) process (for 10 values of n). Panel (F) depicts the semi-log of $R^n(\varepsilon)$ as a function of n . The slopes obtained from the two semi-log representations provide the empirical and theoretical values of the correlation entropy for $\varepsilon = 1.0$, respectively $\kappa_\varepsilon(\varepsilon) = 0.6485$ and $\kappa(\varepsilon) = 0.6536$.

For the fGn processes, the results depicted in Fig. 13 indicate similar results but the underestimation of the ε -entropy is only observed when the length of the series is $N = 75$.

5. Discussion and conclusion

The purpose of this study was to investigate the statistical properties of the recurrence plots of data generated by discrete-time stationary and Gaussian stochastic processes as a function of the size N of the RP and the resolution ε used to construct it.

More specifically, we focused on specific measures widely used in the literature (see [7,23] and references therein), including the recurrence rate, the percent determinism and the RP-based ε -entropy (in the sense of correlation entropy). We analytically showed that, in the one-dimensional case (without embedding), it is possible to estimate the theoretical probabilities of occurrence of a recurrence point and occurrence of a diagonal of a given length n . These results allowed us to compute the theoretical values of the three RP-based measures of interest. The choice of one-dimensional embedding, which implies a unit time delay,

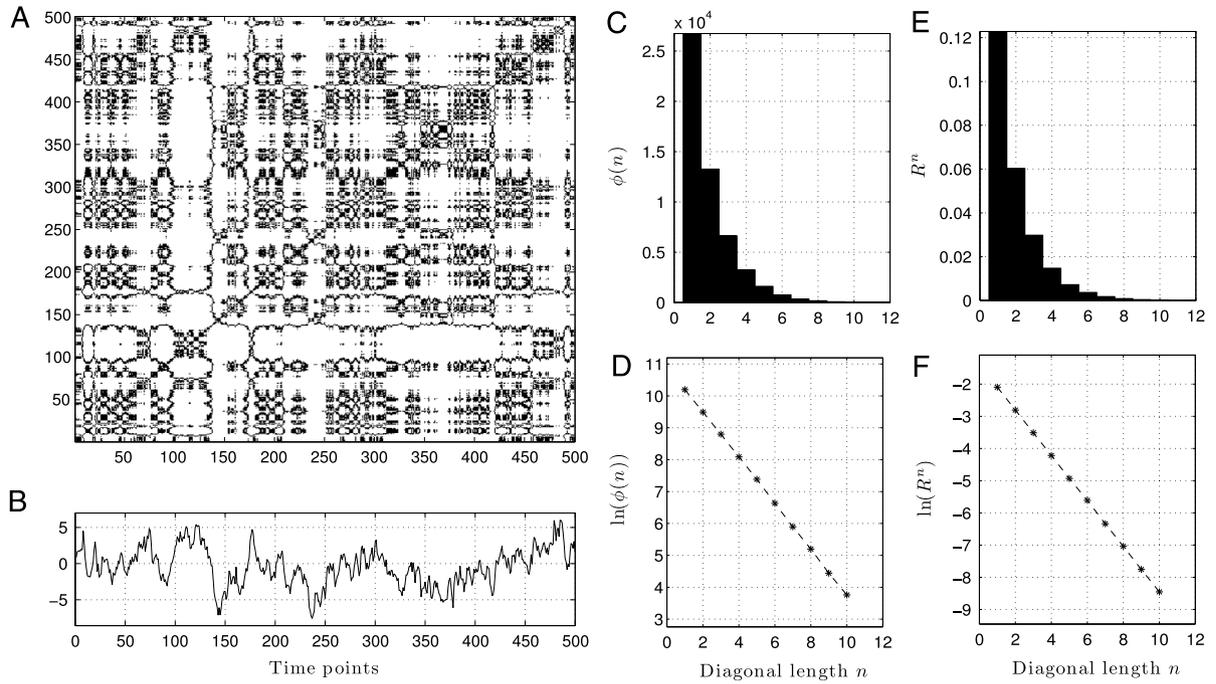


Fig. 9. Same as Fig. 8 for AR(1) process with $\varphi = 0.9$. In this case, the empirical and theoretical values of the correlation entropy are respectively $\kappa_e(\varepsilon) = 0.7163$ and $\kappa(\varepsilon) = 0.7057$.

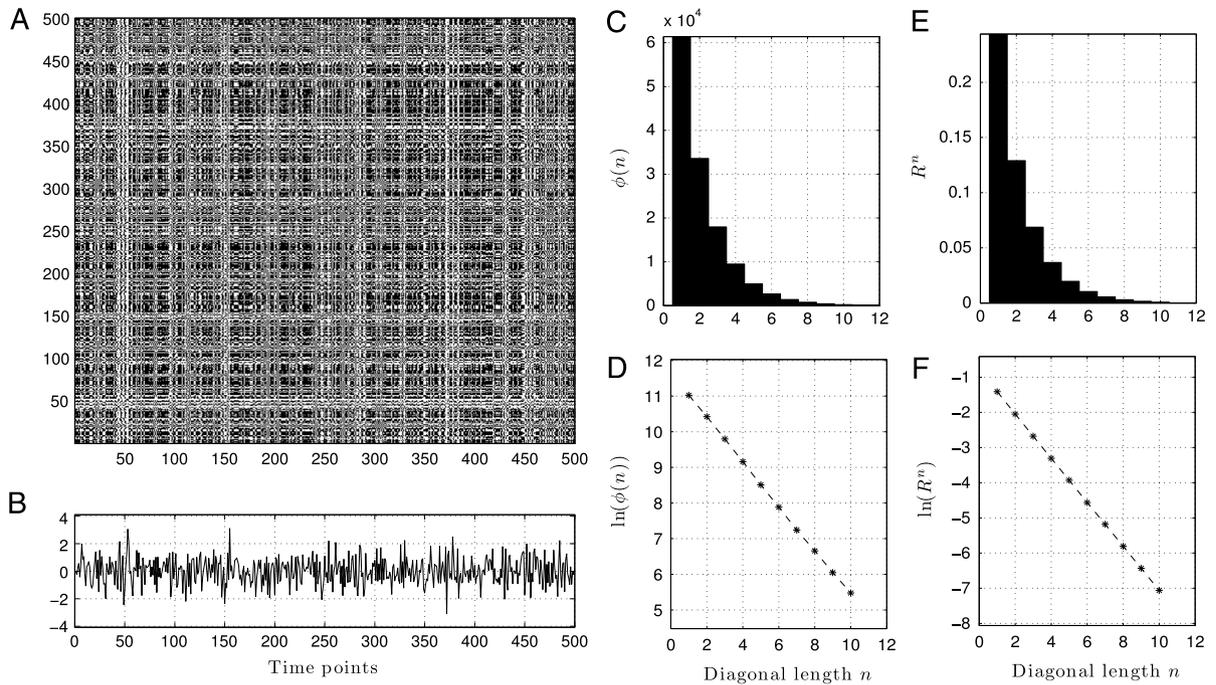


Fig. 10. Same as Fig. 8 for fGn process with $H = 0.3$. In this case, the empirical and theoretical values of the correlation entropy are respectively $\kappa_e(\varepsilon) = 0.6294$ and $\kappa(\varepsilon) = 0.6270$.

facilitates the computations involved in the analysis but can also be justified by the discrete-time nature of the stochastic processes under consideration. Indeed, the statistics of diagonal lines for different embedding dimensions d can be related since a diagonal line of length n in dimension d appears as a diagonal line of length $(n + d - 1)$ when the embedding dimension is equal to 1. In the context of chaotic dynamics and dynamical invariants estimation, the relevance of RPs constructed without embedding has been demonstrated [43,36,6]. In addition, it should be noted that in the case of one-dimensional embedding RPs, the maximum and Euclidean norms used to define recurrences are equal. In this

work, we have used the maximum norm to define the domain corresponding to a diagonal line: $\mathbb{M}(\varepsilon) = \{\mathbf{z} : \mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\|_\infty \leq \varepsilon\}$, involved in the integration of the multivariate Gaussian pdf for computing $P_{ij}^n(\varepsilon)$ (see Section 3.2). We emphasize the fact that the maximum norm is here used as a shortcut of a mathematical notation (and not as a convenient choice, like in the definition of the RP, see Eq. (2)).

The proposed approach can be easily extended to other measures based on diagonal lines such as the *average diagonal line length* L (see for instance [7], where L is related to the mean prediction time of the process). The definition of this measure

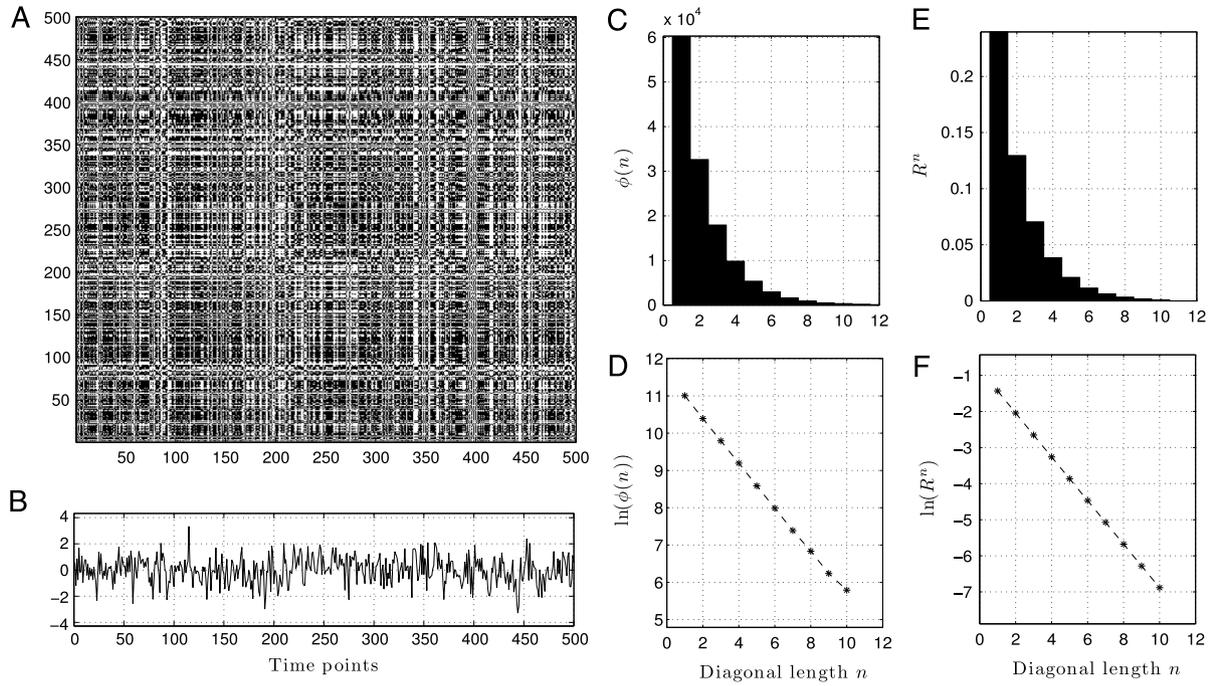


Fig. 11. Same as Fig. 8 for fGn process with $H = 0.7$. In this case, the empirical and theoretical values of the correlation entropy are respectively $\kappa_c(\varepsilon) = 0.5984$ and $\kappa(\varepsilon) = 0.6057$.

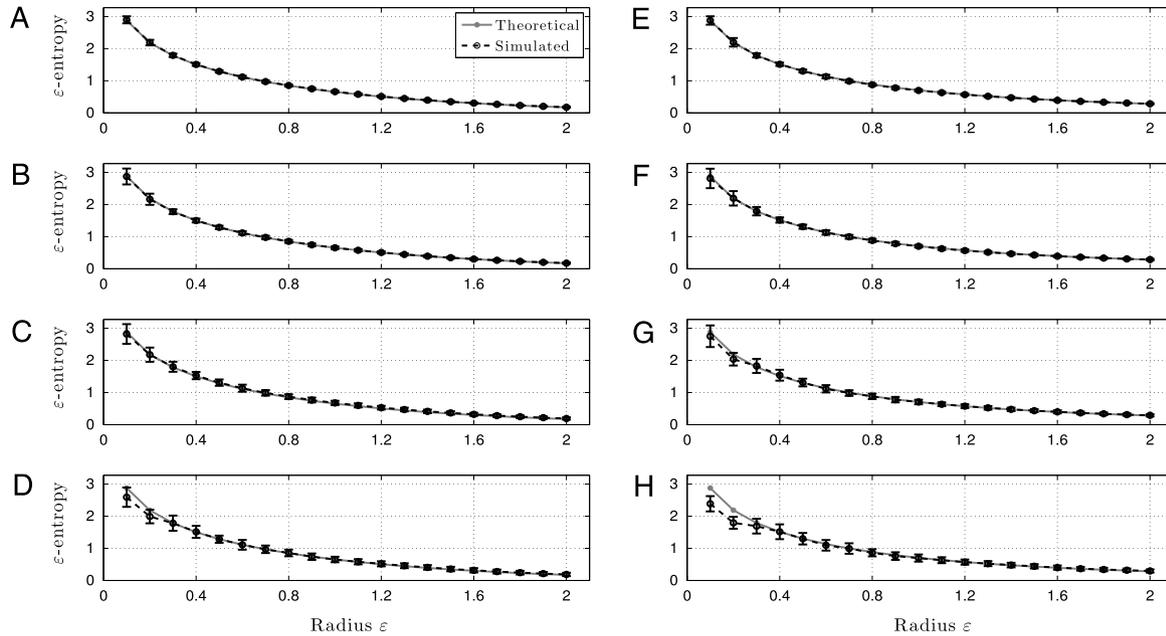


Fig. 12. The theoretical ε -entropy $\kappa(\varepsilon)$ compared to the mean ε -entropy $\kappa_c(\varepsilon)$ obtained from 50 simulated series of the AR(1) processes. The error bars indicate standard deviations. (A) AR(1) with $\varphi = 0.1$, and $N = 500$, (B) $N = 250$, (C) $N = 125$, (D) $N = 75$, (E) AR(1) with $\varphi = 0.9$, and $N = 500$, (F) $N = 250$, (G) $N = 125$, (H) $N = 75$.

involves the number of diagonal lines of length exactly k , which is also involved in the definition of the percent determinism $DET(\varepsilon, n)$, and as such investigated in our manuscript. In particular, the theoretical value of L could be calculated using the probability $Q^k(\varepsilon)$ of these diagonal lines. Similarly, the theoretical value of the longest diagonal L_{\max} could also be extracted from $Q^k(\varepsilon)$. Another RQA measure based on diagonal lines that can be theoretically computed from $Q^k(\varepsilon)$ is the Shannon entropy of the distribution of diagonal line lengths, denoted ENT (see [7]). Less frequent RQA measures are based on vertical lines, for instance *laminarity*, which

is equivalent to DET for the vertical structures of the RP, and *trapping time* given by the average length of vertical lines [7]. These measures are generally used for the quantification of the occurrence of laminar phases or identify chaos-chaos transitions and intermittency phenomena. In the present work, we have not considered these measures which are dedicated to analyze deterministic dynamics. Due to the very specificity of recurrence networks, the associated measures are not related to diagonal-lines statistics (which reflects the closeness of states over the course of time). Accordingly, the extension of our results is not

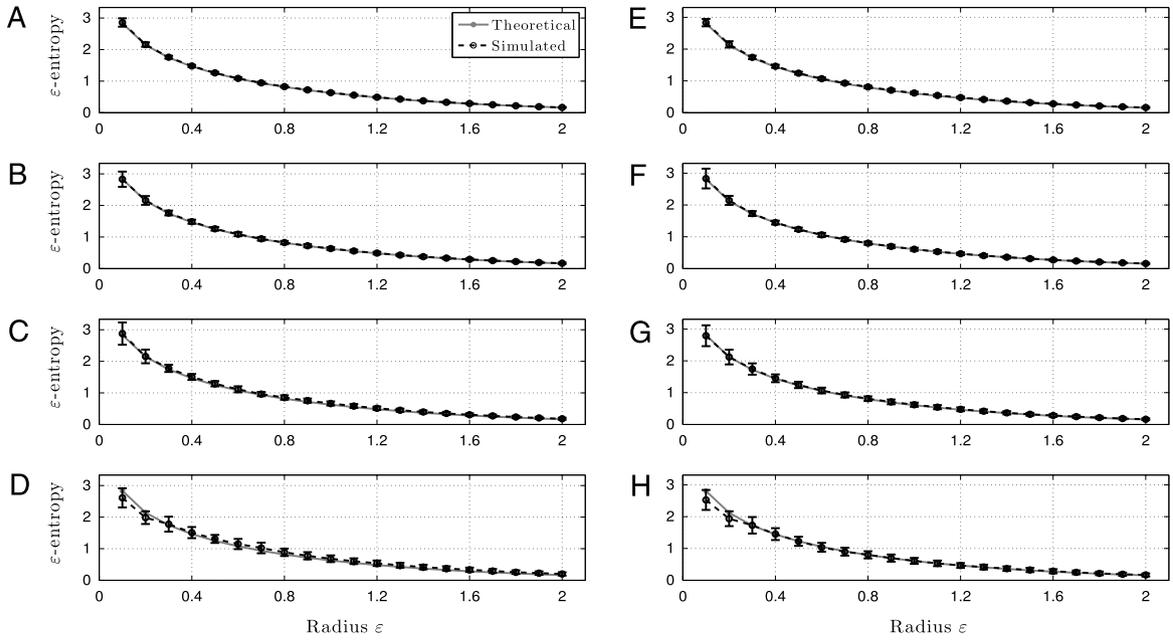


Fig. 13. Same as Fig. 12 for fGn processes. (A) fGn with $H = 0.3$, and $N = 500$, (B) $N = 250$, (C) $N = 125$, (D) $N = 75$, (E) fGn with $H = 0.7$, and $N = 500$, (F) $N = 250$, (G) $N = 125$, (H) $N = 75$.

straightforward. Only the so-called edge density, which is the formulation in network language of the recurrence rate, can be covered by our study (see [33–35]).

In a recent work [18], Zou et al. emphasized the importance of the stationarity of the studied stochastic process to correctly achieve a RN analysis. For our analytical study of diagonal-based measures of RPs such as ε -entropy (in the sense of correlation entropy), the stationarity assumption is necessary to ensure that the statistics are identical over all sub-RPs that could be constructed from the original RP and that the extracted measures will be equivalently computed over different segments of the dynamics. The authors also reported that defining an embedding dimension for a stationary stochastic process such as fGn is theoretically impossible arguing that this process is infinite-dimensional. However, they proposed that an embedding delay can be formally obtained from time series generated by fGn through the estimation of autocorrelation function. They reported that a low-dimensional embedding is not recommended because it will potentially lead to a loss of information. Our results suggest that the considered RQA measures are accurately estimated using one-dimensional embedding, even with $H = 0.7$, which corresponds to a positively correlated (or persistent) process. In the case of ε -entropy, the result of the analytical computation described here is coherent with the expected scaling as a function of ε ($\kappa(\varepsilon) \sim -\ln(\varepsilon)$, see [41,42,44]). As mentioned previously, one can relate the diagonal-based measures computed for different embedding dimensions. We also observe very similar results for fGn with $H = 0.3$ and $H = 0.7$. This is in accordance with the theoretical results regarding the scaling of ε -entropy, which cannot distinguish fGns with different H exponent [44].

From a more general and conceptual point of view, it should be noted that time delay embedding was originally defined and developed rigorously in the context of deterministic dynamics [22], allowing the reconstruction of phase space trajectories of multivariate systems from univariate time series. In a probabilistic framework, for an univariate discrete-time random process, recurrence plots are used to characterize the statistical behavior of predictability measures from the observation of a realization of the process. The findings reported here are in favor of the relevance of the construction of one-dimensional RPs

for discrete-time stochastic processes, given that stationarity is assumed. Note that the proposed approach can be extended to non-Gaussian stationary processes if the corresponding multivariate integrals can be computed.

The numerical experiments performed on two classical Gaussian processes show that the theoretical values of the considered measures provide useful information to assess the quality of the RP-based empirical estimations. The main conclusions that can be made from the statistics obtained through these numerical experiments are that the REC estimations are very good except in the case of processes for which short series ($N = 250$) significantly deviate from stationarity such as the AR(1) process for $\varphi = 0.9$ (see panel (F) of Fig. 4). For DET measure, the estimation is globally accurate for long series ($N = 1000$ and $N = 500$) and it is less efficient for shorter data set ($N = 250$) and more crucially for values of ε lower than 0.4 or 0.6. In this case, the overestimation is more pronounced when n is increased. For the AR(1) process with $\varphi = 0.9$, the overestimation is noted for both $N = 500$ and $N = 250$, for $n = 2, 3$ and 4, especially when ε is small (see panels (E) and (F) of Fig. 6). Once again, this result can be explained by the deviation from the stationarity assumption. We underline that for problematical values $\varepsilon = 0.2$ and $\varepsilon = 0.4$, the estimated and theoretical values of REC approximately range from 0.05 to 0.20. Such values, or even lower ones, are commonly used in the literature for various applications. This observation is not in favor of the approach based on targeting a low REC value as a criterion for selecting the threshold radius ε optimizing DET estimation. It should be noted that for values of the threshold ε larger than the range explored here (0.2 to 1.2), the estimation of $DET_{th}(\varepsilon, n)$ from Eq. (32) can be improved by using larger maximal k values to include more terms in the involved sums. This is a natural consequence of the slower vanishing of probability $Q^k(\varepsilon)$ with k , when ε is larger. Concerning the RP-based ε -entropy estimation, the results suggest a more robust behavior regarding the length of data. Using 50 realizations, the statistics showed that the estimations were globally efficient for lengths $N = 125$ and even $N = 75$ for a threshold radius ε larger or equal than 0.3. The poor performance of the RP-based estimations of the different measures observed for the smaller values of ε can be explained by the lack of neighbors in the state space, which naturally reduces the quality of the corresponding statistics. In addition, for the estimation of the ε -entropy, it should be

noted that the linear region of the histogram $\phi(\varepsilon, n)$ is reduced for the smaller values of ε . The deviations between the theoretical and empirical values observed for the shortest time series and the smallest threshold radius values are also related to the choice of the Theiler window [22,7]. For our simulations, we have chosen not to use such a window because its determination is not straightforward for correlated stochastic processes. However, we observed a better agreement between the theoretical and empirical values by excluding only the main diagonal, especially for the *DET* measure for $n = 3$ and $n = 4$. This is coherent since, for short series and low ε values, the contribution of the main diagonal is relatively more important. The theoretical expressions derived in Section 3 precisely provide a benchmark to validate such estimation choices. In particular, the selection of an appropriate Theiler window based on the probabilities $P_{i,j}$ and $P_{i,j}^n$ (see Figs. 1, 2, 3) could be an interesting research perspective. The quantitative comparison of the estimations of these RQA measures to their theoretical values can be refined by performing appropriate statistical tests.

For our numerical experiments, we used 50 realizations of each considered process. This choice was made to ensure a meaningful estimation of the mean and standard deviation of the RP-based measures when compared to their theoretical counterparts. However, by reducing the number of realizations, we observed a very good robustness of the resulting statistics up to less than 10 realizations (6–9, depending on the measure), even for the shortest series. The assumption of ergodicity is thus relevant here, such that the number of realizations compensates for the finite length of the sample paths.

As an example of real-world data presumed to be satisfactorily modeled by a Gaussian process, hence in the scope of our study, we may cite electroencephalographic (EEG) signals, which have been already explored through RQA (see for instance [45,46]). A preprocessing step is to model the experimental time series using AR(p) autoregressive processes [47,48], generally with a high value of p. Implementing our general results for the case of these processes (as done in Section 4 for the special case of AR(1)) would allow one to predict the statistical behavior of RQA measures, with respect to the length of the analyzed signals and the number of available realizations of these signals.

In the present work, we proposed a methodological approach to analytically study the RPs (with an embedding dimension equal to 1) of a widespread class of random processes, namely discrete-time stationary Gaussian processes. This approach allows the theoretical computation of the recurrence rate and diagonal-based measures such as percent determinism and ε -entropy given that the covariance of the considered process is known.

These findings provide useful information and guidelines for the processing of real-world Gaussian data in terms of the minimal length of the series, number of realizations, threshold radius ε used to construct the RP and the minimal length of diagonals for the estimation of percent determinism. Further research should confirm these statistical results for other classes of stationary random processes.

Acknowledgment

S. Ramdani would like to thank V. Kleptsyn for his useful advices regarding probability calculations.

Appendix

For $n \geq 2$, there is no closed-form expression for (23) and a numerical method must be applied to compute the multivariate integrals involved in this equation.

This is a classical problem encountered in various statistical applications and many contributions can be found in the literature.

The specific bivariate and trivariate normal problems have been studied by several authors [49–53]. Other methods consider the more general case of the multivariate normal integral [54,30]. In our application, we have used the stochastic method proposed by Genz [30]. The principle of this algorithm is to express the original integral as an integral over the unit hyper-cube thanks to a sequence of transformations based on a Cholesky decomposition of the covariance matrix. Denoting $\mathbf{C}\mathbf{C}^T$ the Cholesky decomposition of the covariance matrix $\mathbf{\Omega}^{i,j}$ and $(c_{k,l})$ the elements of matrix \mathbf{C} , we get [30]

$$P_{i,j}^n(\varepsilon) = (u_1 - v_1) \int_0^1 (u_2 - v_2) \dots \int_0^1 (u_n - v_n) \int_0^1 d\mathbf{w} \quad (45)$$

where we have

$$u_1 = \Phi\left(\frac{b_1}{c_{1,1}}\right) \quad (46)$$

$$v_1 = \Phi\left(\frac{a_1}{c_{1,1}}\right) \quad (47)$$

and for $k = 2, \dots, n$

$$u_k = \Phi\left(\frac{1}{c_{k,k}} \left[b_k - \sum_{l=1}^{k-1} c_{k,l} \Phi^{-1}(v_l + w_l(u_l - v_l)) \right]\right) \quad (48)$$

$$v_k = \Phi\left(\frac{1}{c_{k,k}} \left[a_k - \sum_{l=1}^{k-1} c_{k,l} \Phi^{-1}(v_l + w_l(u_l - v_l)) \right]\right) \quad (49)$$

with $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$ and $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp(-\frac{1}{2}\theta^2) d\theta$ is the standard univariate normal distribution function. The real numbers (a_k) and (b_k) correspond to the limits of the integration domain. In our application, we set $a_k = -\varepsilon$ and $b_k = \varepsilon$.

From the formulation given by Eq. (45), a classical Monte-Carlo algorithm is used to compute the integral (see [30] for more details). To achieve these computations, we used the freely available source code provided by A. Genz.

References

- [1] J.P. Eckmann, S.O. Kamphorst, D. Ruelle, Recurrence plots of dynamical systems, *Europhys. Lett.* 4 (1987) 973–977.
- [2] J.P. Zbilut, C.L. Webber, Embeddings and delays as derived from quantification of recurrence plots, *Phys. Lett. A* 171 (1992) 199–203.
- [3] C.L. Webber, J.P. Zbilut, Dynamical assessment of physiological systems and states using recurrence plot strategies, *J. Appl. Physiol.* 76 (1994) 965–973.
- [4] P. Faure, H. Korn, A new method to estimate the Kolmogorov entropy from recurrence plots: its application to neuronal signals, *Physica D* 122 (1998) 265–279.
- [5] K. Urbanowicz, J.A. Holyst, Noise-level estimation of time series using coarse-grained entropy, *Phys. Rev. E* 67 (2003) 46218.
- [6] M. Thiel, M.C. Romano, P.L. Read, J. Kurths, Estimation of dynamical invariants without embedding by recurrence plots, *Chaos* 14 (2004) 234–243.
- [7] N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, *Phys. Rep.* 438 (2007) 237–329.
- [8] G. Robinson, M. Thiel, Estimation of dynamical invariants without embedding by recurrence plots, *Chaos* 19 (2009) 023104.
- [9] C.L. Webber, N. Marwan, *Recurrence Quantification Analysis*, Springer International Publishing, 2015.
- [10] N. Marwan, J. Kurths, Comment on “Stochastic analysis of recurrence plots with applications to the detection of deterministic signals” by Rohde et al. [*Physica D* 237 (2008) 619–629], *Physica D* 438 (2009) 1711–1715.
- [11] N. Marwan, How to avoid potential pitfalls in recurrence plot based data analysis, *Internat. J. Bifur. Chaos* 21 (2011) 1003–1017.
- [12] G.K. Rohde, J.M. Nichols, B.M. Dissinger, F. Bucholtz, Stochastic analysis of recurrence plots with applications to the detection of deterministic signals, *Physica D* 237 (2008) 619–629.
- [13] J.P. Zbilut, N. Marwan, The Wiener-Khinchin theorem and recurrence quantification, *Phys. Lett. A* 372 (2008) 6622–6626.
- [14] P. Faure, A. Lesne, Recurrence plots for symbolic sequences, *Internat. J. Bifur. Chaos* 20 (2010) 1731–1749.

- [15] A. Sipers, P. Borm, R. Peeters, On the unique reconstruction of a signal from its unthresholded recurrence plot, *Phys. Lett. A* 375 (2011) 2309–2321.
- [16] M. Grendár, J. Majerová, V. Špitalský, Strong laws for recurrence quantification analysis, *Internat. J. Bifur. Chaos* 23 (2013) 1350147.
- [17] D. Schultz, S. Spiegel, N. Marwan, S. Albayrak, Approximation of diagonal line based measures in recurrence quantification analysis, *Phys. Lett. A* 379 (2015) 997–1011.
- [18] Y. Zou, R.V. Donner, J. Kurths, Analyzing long-term correlated stochastic processes by means of recurrence networks: Potentials and pitfalls, *Phys. Rev. E* 91 (2015) 022926.
- [19] J.-L. Liu, Z.-G. Yu, V. Anh, Topological properties and fractal analysis of recurrence network constructed from fractional Brownian motions, *Phys. Rev. E* 89 (2014) 032814.
- [20] N.H. Packard, J.P. Crutchfield, J.D. Farmer, R.S. Shaw, Geometry from a time series, *Phys. Rev. Lett.* 45 (1980) 712–716.
- [21] F. Takens, Detecting strange attractors in turbulence, in: *Dynamical Systems and Turbulence*, in: *Lecture Notes in Mathematics*, vol. 898, Springer, Berlin, 1981, pp. 366–381.
- [22] H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, 2004.
- [23] N. Marwan, C.L. Webber, *Mathematical and computational foundations of recurrence quantifications*, in: *Recurrence Quantification Analysis*, Springer International Publishing, 2015, pp. 3–43.
- [24] P. Faure, A. Lesne, Estimating Kolmogorov entropy from recurrence plots, in: *Recurrence Quantification Analysis*, Springer International Publishing, 2015, pp. 45–63.
- [25] D.T. Kaplan, L. Glass, Coarse-grained embeddings of time series: random walks, Gaussian random processes, and deterministic chaos, *Physica D* 64 (1993) 431–454.
- [26] M.C. Casdagli, Recurrence plots revisited, *Physica D* 108 (1997) 12–44.
- [27] A. Papoulis, S.U. Pillai, *Probability, Random Variables and Stochastic Processes*, fourth ed., McGraw-Hill Higher Education, 2002.
- [28] J.S. Bendat, A.G. Piersol, *Random Data: Analysis and Measurement Procedures*, second ed., Wiley, New-York, 1986.
- [29] C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
- [30] A. Genz, Numerical computation of multivariate normal probabilities, *J. Comput. Graph. Statist.* 1 (1992) 141–149.
- [31] J. Sheil, I. O’Muircheartaigh, Algorithm AS 106: The distribution of non-negative quadratic forms in normal variables, *Appl. Stat.* 26 (1977) 92–98.
- [32] M. Thiel, M.C. Romano, J. Kurths, R. Meucci, E. Allaria, F.T. Arecchi, Influence of observational noise on the recurrence quantification analysis, *Physica D* 171 (2002) 138–152.
- [33] R.V. Donner, Y. Zou, J.F. Donges, N. Marwan, J. Kurths, Recurrence networks—a novel paradigm for nonlinear time series analysis, *New J. Phys.* 12 (2010) 033025.
- [34] J.F. Donges, J. Heitzig, R.V. Donner, J. Kurths, Analytical framework for recurrence network analysis of time series, *Phys. Rev. E* 85 (2012) 046105.
- [35] Y. Zou, J. Heitzig, R.V. Donner, J.F. Donges, J.D. Farmer, R. Meucci, S. Euzzor, N. Marwan, J. Kurths, Power-laws in recurrence networks from dynamical systems, *Europhys. Lett.* 98 (2012) 48001.
- [36] M. Thiel, *Recurrences: exploiting naturally occurring analogues* (Ph.D. thesis), University of Potsdam, Germany, 2004.
- [37] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, J. Kurths, Recurrence plot based measures of complexity and its application to heart rate variability data, *Phys. Rev. E* 66 (2002) 026702.
- [38] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, *Time Series Analysis: Forecasting and Control*, fourth ed., Wiley, 2008.
- [39] B.B. Mandelbrot, J.W. van Ness, Fractional Brownian motions, fractional noises and applications, *SIAM Rev.* 10 (1968) 422–437.
- [40] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, 1994.
- [41] P. Gaspard, X.-J. Wang, Noise, chaos, and (ε, τ) -entropy per unit time, *Phys. Rep.* 235 (1993) 291–343.
- [42] G. Nicolis, P. Gaspard, Toward a probabilistic approach to complex systems, *Chaos Solitons Fractals* 4 (1994) 41–57.
- [43] J. Iwanski, E. Bradley, Recurrence plot analysis: to embed or not to embed? *Chaos* 8 (1998) 861–871.
- [44] J. Gao, J. Hu, F. Liu, Y. Cao, Multiscale entropy analysis of biological signals: a fundamental bi-scaling law, *Front. Comput. Neurosci.* 9 (2015) 64.
- [45] U.R. Acharya, S.V. Sree, S. Chattopadhyay, W. Yu, P.C.A. Ang, Application of recurrence quantification analysis for the automated identification of epileptic EEG signals, *Int. J. Neural Syst.* 21 (2011) 199–211.
- [46] E.J. Ngamga, S. Bialonski, N. Marwan, J. Kurths, C. Geier, K. Lehnertz, Evaluation of selected recurrence measures in discriminating pre-ictal and inter-ictal periods from epileptic EEG data, *Phys. Lett. A* 380 (2016) 1419–1425.
- [47] J. Pardey, S. Roberts, L. Tarassenko, A review of parametric modeling techniques for EEG analysis, *Med. Eng. Phys.* 18 (1996) 2–11.
- [48] E. Pereda, R.Q. Quiroga, J. Bhattacharya, Nonlinear multivariate analysis of neurophysiological signal, *Prog. Neurobiol.* 77 (2005) 1–37.
- [49] D.B. Owen, Tables for computing bivariate normal probability, *Ann. Math. Stat.* 27 (1956) 1075–1090.
- [50] J.V. Terza, U. Welland, A Comparison of bivariate normal algorithms, *J. Stat. Comput. Simul.* 19 (1988) 115–127.
- [51] Z. Drezner, G.O. Wesolowsky, On the computation of the bivariate normal integral, *J. Stat. Comput. Simul.* 35 (1990) 101–107.
- [52] Z. Drezner, Computation of the trivariate normal integral, *Math. Comp.* 62 (1994) 289–294.
- [53] A. Genz, Numerical computation of rectangular bivariate and trivariate normal and t probabilities, *Stat. Comput.* 14 (2004) 251–260.
- [54] Z. Drezner, Computation of the multivariate normal integral, *ACM Trans. Math. Software* 18 (1992) 450–460.