Delay independence of mutual-information rate of two symbolic sequences

Jean-Luc Blanc,[1] Laurent Pezard,[1] and Annick Lesne[2,3,*]

[1]*Laboratoire de Neurosciences Intégratives et Adaptatives UMR 6149 CNRS Aix-Marseille Université, 3 Place Victor Hugo, F-13331 Marseille Cedex 03, France*

[2]*Laboratoire de Physique Théorique de la Matière Condensée UMR 7600 CNRS Université Pierre et Marie Curie-Paris 6, 4 place Jussieu, F-75252 Paris Cedex 05, France*

[3]*Institut des Hautes Études Scientifiques Le Bois-Marie, 35 route de Chartres, F-91440 Bures-sur-Yvette, France*

(Received 11 November 2010; revised manuscript received 30 May 2011; published 23 September 2011)

Introduced by Shannon as a "rate of actual transmission," mutual information rate (MIR) is an extension of mutual information to a pair of dynamical processes. We show a delay-independence theorem, according to which MIR is not sensitive to a time shift between the two processes. Numerical studies of several benchmark situations confirm that this theoretical asymptotic property remains valid for realistic finite sequences. Estimations based on block entropies and a causal state machine algorithm perform better than an estimation based on a Lempel-Ziv compression algorithm provided that block length and maximum history length, respectively, can be chosen larger than the delay. MIR is thus a relevant index for measuring nonlinear correlations between two experimental or simulated sequences when the transmission delay (in input-output devices) or dephasing (in coupled systems) is variable or unknown.

DOI: 10.1103/PhysRevE.84.036214 PACS number(s): 05.45.Tp

I. INTRODUCTION

Shannon devised entropy per unit time (denoted h thereafter) of discrete-valued stochastic processes as one of the basic concepts of information theory [1]. This concept has been later developed in dynamical systems theory by Kolmogorov [2] and Sinai [3], and there known as metric entropy or Kolmogorov-Sinai entropy. Its relevance to quantify the overall temporal organization of an evolution has been recognized in numerous applications, and it is now a standard tool of nonlinear time series analysis [4]. Directly derived from Shannon entropy, mutual information measures nonlinear correlation between two discrete random variables. A natural extension crossing entropy per unit time with mutual information is mutual information per unit time or *mutual information rate* (MIR). This dynamic extension provides a global quantification of the overall joint probability distribution and captures mutual information of two *processes*.

In the section of the seminal paper [1] devoted to transmission in a noisy channel, Shannon already considered MIR of the input signal X and the output signal Y. He first defined the conditional entropy rate $h(X|Y) = h(X,Y) - h(Y)$ measuring the average ambiguity of the output signal, namely, the amount of additional information needed per unit time to correct the transmitted message Y and recover X. The symmetric conditional entropy rate, $h(Y|X)$, is the part due to noise in $h(Y)$. MIR, termed "rate of actual transmission" by Shannon, is defined as

$$i(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X). \quad (1)$$

Compared to entropy rate and mutual information, MIR did not arouse great interest for experimental time series analysis, since its faithful estimation from data was presumed to be impossible and its practical interpretation not straightforward. Many textbooks [5–7] offer only brief mentions of MIR. Investigation of signal transmission in a noisy channel (see, e.g., Ref. [8]), involving MIR between a Gaussian input $S(t)$ and the output $S(t) + N(t)$ where $N(t)$ is a Gaussian white noise, do not apply to the case where the two sequences have different state spaces. Information rates considered by Palus *et al.* [9] are, in fact, time-averaged delayed mutual information, whose descriptive power is that of coarse-grained rates [10]. They differ in their definition and properties from the mutual information rate considered in the present paper. Detailed investigations presented in Refs. [11–13] focus on asymptotic properties and mathematical aspects, for instance, ergodic decomposition, dependence of MIR on the discretization of an underlying continuous state space, and extension to continuous alphabets. They are not straightforwardly relevant for practical analysis of finite-length symbolic sequences. Zozor *et al.* [14] mentioned that extension of Lempel-Ziv (LZ) compression algorithm to multivariate data could be implemented for computing MIR (see Sec. III A). They did not investigate this direction further since the estimated quantity could be negative, whereas the theoretical quantity is always positive.[1] Several faithful estimation methods, including those used below and more computationally demanding ones [15], are now available and renew the practical relevance of this index.

The goal of the present paper is twofold. We first study the properties of MIR as a dynamic extension of mutual information suitable to analyze quantitatively the joint temporal organization and nonlinear correlation of two processes, in the context of input-output systems or coupled systems. We show that MIR is independent of the delay between the two processes. We then investigate whether this delay-independence theorem established in the asymptotic limit remains valid in experimental situations, i.e., for finite sequences. We consider synthetic data obtained in benchmark situations (discretized logistic maps and Markov chain). We show that estimations based on block entropies and a causal state machine algorithm perform better

*lesne@lptmc.jussieu.fr

[1]We will see below that negative estimated values are due to uncontrolled finite-size effects.

than an estimation by means of a LZ compression algorithm (free of any control setting) provided the block length and maximum history length, respectively, can be chosen larger than the delay. Our study delineates the conditions in which MIR provides an operational index to assess nonlinear correlations between two signals when an unknown or variable time shift (delay or dephasing) is present between them, whereas usual mutual information is strongly misleading in this case [16].

II. GENERAL PROPERTIES

A. Notations and definitions

Let us consider a discrete-time process $Z = (Z_t)_{t\geqslant 0}$ with values $z_t = (x_t, y_t)$ in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The alphabets $\mathcal{X}$ and $\mathcal{Y}$ can (and will often) be different. We assume statistical stationarity and ergodicity.[2] We denote X and Y the canonical projections (stochastic processes), X_t and Y_t their components at time t (random variables), and $\theta_\tau X$ the time-shifted process such that $[\theta_\tau X]_t = X_{t+\tau}$. The distribution of Z is not in general the product of the distributions of X and Y, and this is precisely the point to be quantitatively appreciated here. This mathematical setting covers the case of a bivariate recording of a source, the case of two different recordings (i.e., from two different sources), and the case of a communication channel or input-output device.

The overall distributions of Z, X, and Y are defined through the distribution of n-words $w_n = (z_t, z_{t+1}, \ldots, z_{t+n-1})$ independent of t by stationarity. Shannon entropies of these distributions are also called block entropies of order n and denoted, respectively, $H_n(Z)$, $H_n(X)$, and $H_n(Y)$. Explicitly[3] for Z [5],

$$H_n(Z) = -\sum_{w_n} p_Z(w_n) \log p_Z(w_n), \tag{2}$$

where w_n runs over the set of n-words written with alphabet $\mathcal{Z}$. Shannon entropy $H_1(X)$ is extended into the entropy per unit time $h(X)$ of the stochastic process according to

$$h(X) = \lim_{n\to\infty} \frac{H_n(X)}{n} \tag{3}$$

$$= \lim_{n\to\infty} [\, H_{n+1}(X) - H_n(X]. \tag{4}$$

Block mutual information of order n is written

$$I_n(X;Y) = H_n(X) + H_n(Y) - H_n(Z), \tag{5}$$

and the mutual information rate is defined as the limit $\{$recall that if $\lim_{n\to\infty}[I_{n+1}(X;Y) - I_n(X;Y)]$ exists, then $\lim_{n\to\infty} I_n(X;Y)/n$ also exists and they are equal$\}$:

$$i(X;Y) = \lim_{n\to\infty} \frac{I_n(X;Y)}{n} \tag{6}$$

$$= \lim_{n\to\infty} [\, I_{n+1}(X;Y) - I_n(X;Y)\,] \tag{7}$$

$$= h(X) + h(Y) - h(Z). \tag{8}$$

B. Basic properties

It can be shown [13] that i always exists for stationary processes, and basic properties of MIR follow from the above-mentioned relation with conditional entropy rates $i(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$, namely:
(1) Symmetry $i(X;Y) = i(Y;X)$,
(2) Positivity $i(X;Y) \geqslant 0$,
(3) Upper bound $i(X;Y) \leqslant \inf[h(X),h(Y)]$,
(4) Special value $i(X;X) = h(X)$,
(5) Special value $i(X;Y) = I(X;Y)$ if $Z = (X,Y)$ is uncorrelated and identically distributed in time,
(6) $i(X;Y) = 0$ in the case of independent processes X and Y.

C. Interpretations of MIR

Exactly as the entropy rate $h(X)$ gives a better account of the whole temporal structure of the process X compared to linear statistical indices like the autocorrelation function or instantaneous indices like Shannon entropy, $i(X;Y)$ accounts for both the instantaneous correlation between X_t and Y_t and the temporal autocorrelation of the process $t \to (X_t, Y_t)$. It provides a more complete and global quantification of the interrelations between the two *processes* X and Y than mutual information and covariance (all the more since covariance is not straightforwardly defined for discrete-valued processes).

Further interpretation is based on the Shannon-McMillan-Breiman theorem [1,17,18], which states that for N large enough, a source X produces $e^{Nh(X)}$ typical equiprobable N-sequences $(x_1, \ldots, x_N)$. Using (8), MIR relates the number of product sequences $(x_1, y_1, \ldots, x_N, y_N)$ to the actual number of typical N-sequences $(z_1, \ldots, z_N)$ of the joint source $Z = (X,Y)$ according to

$$e^{Nh(Z)} = \frac{e^{N[h(X)+h(Y)]}}{e^{Ni(X;Y)}}. \tag{9}$$

The reduction of the set of typical joint realizations compared to the case of independent sources is thus quantified by the factor $e^{Ni(X;Y)}$. As soon as $i(X;Y) > 0$, the realization of one component of the pair (X,Y) partly determines the realization of the second one, all the more since $i(X;Y)$ is large. Small values of $i(X;Y)$ could correspond either to correlated signals of very small entropy rates or to independent signals.

In the context of communication theory [1], Shannon proved that the maximum over all sources X of $i(X;Y)$ defines the capacity of a given noisy transmission channel. According to (1) and the Shannon-McMillan-Breiman theorem, each typical output can be produced by about $e^{Nh(X|Y)}$ inputs ("reasonable causes of the output" in Shannon's words [1]) with $e^{Nh(X|Y)} = e^{Nh(X)}e^{-Ni(X;Y)}$. Similarly, each typical input message produces about $e^{Nh(Y|X)}$ outputs ("reasonable consequences of the input" [1]) with $e^{Nh(Y|X)} = e^{Nh(Y)}e^{-Ni(X;Y)}$. Thus, if $i(X;Y) = h(X)$, the output is associated to a single input message, and conversely if $i(X;Y) = h(Y)$, a single output is associated to each typical input message. Since $i(X;Y) \leqslant \inf[h(X),h(Y)]$, it is not possible to have a single input associated to an emitted message if $h(X) < h(Y)$. On the contrary, if $i(X;Y) = 0$, the input message X carries no knowledge about the output, which can be associated to any

[2] Ergodicity is a weak property for a stochastic process, which requires only that any state can be reached from any other one.

[3] We shall use throughout the paper the Neperian logarithm, denoted log, as generally done in dynamical systems theory. Using the binary logarithm $\log_2$ would yield entropies and mutual information in bits, as usual in information theory.

of the input messages, and the output message can be any of the typical outputs.

D. MIR delay independence

1. MIR delay independence theorem

Starting from a sequence $(x_t)_{t\geqslant 1}$, that is, a realization of X, we derive the corresponding realization of the shifted process $\theta_\tau X$ by simply shifting the time labels by an integral amount τ. Whereas in general, $I(X_t;Y_t)$ and $I(X_t;\theta_\tau Y)$ are different, an important feature of MIR is the following *delay independence theorem* (its proof is given in the Appendix):

For any processes X and Y, and any delay τ, $i(X;\theta_\tau Y) = i(X;Y)$.

MIR thus evaluates the correlation between the two sequences as a whole and notwithstanding a possible delay.

2. Shifted processes $Y = \theta_\tau X$

For any ergodic and stationary source X and any time shift τ, the MIR delay independence theorem implies that $i(X;\theta_\tau X) = i(X;X) = h(X)$. MIR properly captures that the *processes* X and $\theta_\tau X$ are essentially identical. In contrast, except in critical situations with long-range correlations, mutual information $I(X;\theta_\tau X)$ decreases with τ, typically as $e^{-\tau/\tau_{\mathrm{corr}}}$ [19,20], leading us to conclude that the *signals* are independent as soon as τ is larger than the correlation time τ_{corr}.

3. The special case of Markov processes

Note that $H_n(X,\theta_\tau X) = H_{n+\tau}(X)$ as soon as $n \geqslant \tau$ (for any process). For a Markov process of order 1, $H_n(X) = H_1(X) + (n-1)h(X)$; hence block mutual information is written $I_n(X;\theta_\tau X) = H_1(X) + (n-1-\tau)h(X)$ for $n \geqslant \tau$, which controls the convergence of $I_n(X;\theta_\tau X)/n$ toward $h(X)$. The convergence of the difference is even better since $I_{n+1}(X;\theta_\tau X) - I_n(X;\theta_\tau X) = h(X)$ for any $n \geqslant \tau$. Similar results hold for a Markov chain of any finite order q, using that $H_n(X) = H_q(X) + (n-q)h(X)$ for any $n > q$. For Markov processes, MIR delay independence can be contrasted analytically with the behavior of pointwise mutual information. Denoting M_{xy} the transition probabilities (from state x to y) and p the stationary distribution, mutual information of the time-shifted variables is written:

$$\begin{aligned} I(X_1;X_{1+\tau}) &= 2H_1 - H(X_1,X_{1+\tau}) \\ &= H_1 - \sum_x p_x \sum_y M_{xy}^\tau \log M_{xy}^\tau, \end{aligned} \tag{10}$$

which converges to 0 exponentially fast as $\tau \to \infty$, like $e^{-\tau/\tau_{\mathrm{corr}}}$ where τ_{corr} is the characteristic relaxation time of the Markov chain, given by the second largest eigenvalue $\lambda_1 = e^{-1/\tau_{\mathrm{corr}}}$ [21] (the largest one is $\lambda_0 = 1$). Mutual information fails to capture the full correlation between the time-shifted sequences as soon as a nonvanishing delay $\tau > 0$ is present. Note that *delayed* mutual information [9], that is, mutual information between the first component and a backward-shifted version of the second one, would be a relevant index if the goal were to determine the lag τ (for instance, in a second step, after having evidenced the presence of correlations between the two processes using MIR). MIR and delayed mutual information are thus complementary indices, each fulfilling a specific goal.

4. Input-output devices

The use of MIR in the context of transmission channels, as originally proposed by Shannon [1], is specially relevant in view of our theorem. Indeed, in practical situations, the delay between the input and the output signals is usually unknown, and one has no simple way of achieving a temporal alignment of the two sequences. Our theorem states that delays do not affect the rate of actual transmission of the channel. If the output Y is simply a time-delayed copy of the input X (i.e., $Y = \theta_\tau X$), $h(X|Y)$ vanishes, in agreement with the fact that there is no ambiguity on the input message given the output message. Practical applications also cover the analysis of information transmitted by spike trains [22].

5. Coupled systems

MIR is also of interest in the context of self-organizing systems, where dynamics is induced by the interactions of elements and emergent organization is not discernible at the individual level [23,24]. Considering strongly coupled systems, the fact that the evolution of Y reflects that of X, $Y = \theta_\tau f(X)$, is revealed in MIR since then $i(X;Y) = i[X,f(X)]$. Characterization of interactions is completed by considering an asymmetric index such as transfer entropy [25,26] $I(X_{t+1};X_t,Y_t)$ to assess the direction of the coupling between X and Y. Note that this transfer entropy is nothing but conditional mutual information $I[X(t+1);Y(t)|X(t)]$ as shown in Ref. [27].

III. RELEVANCE OF MIR FOR DATA ANALYSIS

In order to appreciate the practical relevance of using MIR for investigating experimental input-output systems and coupled systems, we have to delineate the conditions in which the *asymptotic* independence with respect to a delay or time shift between the two sequences established in the previous section remains valid for *finite sequences*. We will investigate the behavior of MIR *estimated from finite numerical data*, considering two standard benchmark systems [28]: discretized logistic maps and two-symbol Markov chains.

A. MIR estimators

Since $i(X;Y) = h(X) + h(Y) - h(X,Y)$, there are as many methods for estimating MIR as for estimating entropy rates. Extending our previous investigations on entropy rate estimation from short sequences [29], we will use LZ estimators as a preliminary quantification free of any control setting, to be refined with block entropy estimators and a causal state machine-based estimator: the causal state splitting reconstruction (CSSR) algorithm [30–32]. All estimations require statistical stationarity of the sequences, which could compel one to restrict them to a suitable time window.

In the LZ compression algorithm [29,33], the sequence of length N is parsed recursively into $\mathcal{N}_w$ words, by considering as a new word the shortest one that has not yet been encountered. For instance, the sequence 10011011100101000101l . . . is parsed according to

$1 \bullet 0 \bullet 01 \bullet 10 \bullet 11 \bullet 100 \bullet 101 \bullet 00 \bullet 010 \bullet 11 \ldots$ One then computes

$$L = \frac{\mathcal{N}_w(1 + \log_k \mathcal{N}_w)}{N} \quad \text{with} \quad \lim_{N\to\infty} L = \frac{h}{\log k}. \qquad (11)$$

The limit holds for almost all sequences under the assumption of ergodicity of the process. LZ algorithms can be easily extended to compute the joint LZ complexity of two sequences $(x_i)_i$ and $(y_i)_i$, with respective alphabets $\mathcal{X}$ and $\mathcal{Y}$, by considering the sequence $(z_i)_i$ belonging to the finite alphabet $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ [14]. Denoting k_x, k_y, and $k_z = k_x k_y$ the respective sizes of the alphabets, MIR is estimated as

$$\widehat{\imath}_{LZ}(X;Y) = L_X \log k_x + L_Y \log k_y - L_Z \log(k_x k_y), \qquad (12)$$

where L_X (resp. L_Y, L_Z) is the LZ complexity of the symbolic sequence X (resp. Y, Z).

From Eqs. (6) and (7) one defines a *difference* block estimator $\widehat{\imath}_{n,\delta} = \widehat{I}_{n+1} - \widehat{I}_n$ and an *average* block estimator $\widehat{\imath}_{n,\mathrm{av}} = \widehat{I}_n/n$ of i, involving block entropy estimators $\widehat{H}_n$ through $\widehat{I}_n(X;Y) = \widehat{H}_n(X) + \widehat{H}_n(Y) - \widehat{H}_n(X,Y)$. Ideally, the MIR estimator is obtained asymptotically according to $\widehat{\imath} = \lim_{n\to\infty} \widehat{\imath}_{n,\delta} = \lim_{n\to\infty} \widehat{\imath}_{n,\mathrm{av}}$.

Alternatively, entropy rates could be computed using the CSSR algorithm [31] which reconstructs hidden Markov models from a symbolic sequence [34]. The basic idea of this algorithm is to infer causal states from the data and obtain the most compact representation of the probabilistic model generating the symbolic sequence [30]. The CSSR algorithm starts by assuming that the process is an independent and identically distributed symbolic sequence with only one causal state and adds states recursively until the current set of states is statistically sufficient. The algorithm contains three phases: (1) initialization, (2) finding a next-step with sufficient statistic for optimal one-step-ahead prediction according to the maximum sequence history length l, and (3) making the set of states recursively calculable, by splitting the states until they have deterministic transitions. Then the relative frequency of strings of maximum length l for each state gives the probability distribution of states, directly usable for entropy computation.

Several other estimations, more accurate though more computationally demanding, can be implemented. Let us cite those based on the recurrence of some words of motifs [35,36], those based on a variable length Markov chain models [37] (similar in spirit to CSSR), and a Bayesian algorithm based on context-tree reconstruction and supplemented with a Monte Carlo sampling to compute MIR [15]. This latter method, extending to MIR the entropy rate estimation method proposed in Ref. [28], provides a confidence interval and is well suited when the input signal is continuous. Since our aim is to show the applicability and benefit of the MIR delay-independence theorem in the analysis of experimental data, we here deliberately trade accuracy for robustness and simplicity. Development of more accurate and sophisticated estimation algorithms would only strengthen our point.

B. Finite-size errors

The convergence of LZ estimators has been studied for binary sequences in Ref. [29] and shows good performance

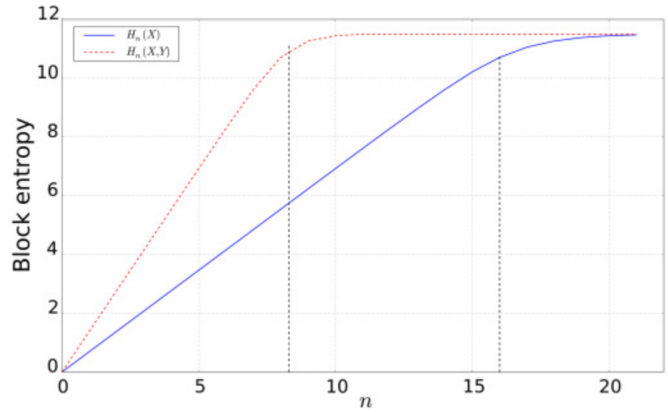


FIG. 1. (Color online) Finite-size behavior of block entropy estimators $H_n(X)$ and $H_n(X,Y)$ when increasing the word length n in the reference case of independent and fully random binary sequences of length $N = 10^5$. The corresponding entropy values are $h_X = h_Y = \log 2$ and $h_{XY} = \log 4$. The vertical dashed lines indicate the theoretical location $n^* = \log N/h$ of the crossover between good and bad statistics above which H_n saturates to the value $\log N$ [29,39].

even for sequences of a few hundreds of symbols. Here a specific difficulty originates in the different convergence rates of LZ estimators for $h(X)$, or $h(Y)$, and for $h(X,Y)$, due to the fact that (X,Y) is written with the product alphabet, of size $k_z = k_x k_y$. This may lead to negative values for the estimated MIR, in contradiction with the positivity of its exact value [38]. Note that convergence of LZ estimator to entropy rate is true only for almost all sequences. As we consider experimental data, a result for almost all sequences, that is, for all the probable sequences, with full measure, is, however, sufficient.

Using block-entropy estimation requires us to find a suitable balance in the choice of the block size n, since statistical errors increase as n increases, while convergence toward the asymptotic value requires large enough n values. Figure 1 illustrates the behavior of $n \to \widehat{H}_n$ in the reference case of random sequences (with $N = 10^5$). Both $\widehat{H}_n(X,Y)$ and $\widehat{H}_n(X)$ saturate to the same value $\log N$ due to finite sampling. Using a straightforward argument, we have shown in Ref. [29] that the crossover between linear regime and saturation in the regime of bad statistics occurs roughly for a value $n^* = \log N/h$, recovering the result rigorously derived in Ref. [39]. A conservative bound is $\log N/\log k$, where k is the size of the alphabet. Accordingly, $\widehat{H}_n(X,Y)$ saturates faster than $\widehat{H}_n(X)$ due to different alphabet sizes (respectively, $k_x k_y$ and k_x). In a similar way,[4] biases and statistical fluctuations hamper the estimation of $H_n(X,Y)$ more strongly than the estimation of $H_n(X)$ and $H_n(Y)$. In all cases the limiting step of the computation of $i(X;Y)$ will be the estimation of $h_{X,Y}$. As in the case of estimation of entropy rate of single sequences [29],

[4]Recall [29] that the variance of the estimator H_n scales like $h.\log k.n^3 e^{nh}/N$ at the leading order; hence we expect that the variance of the estimation of $H_n(X,Y)$ is in general far larger than the variance of $H_n(X) + H_n(Y)$.

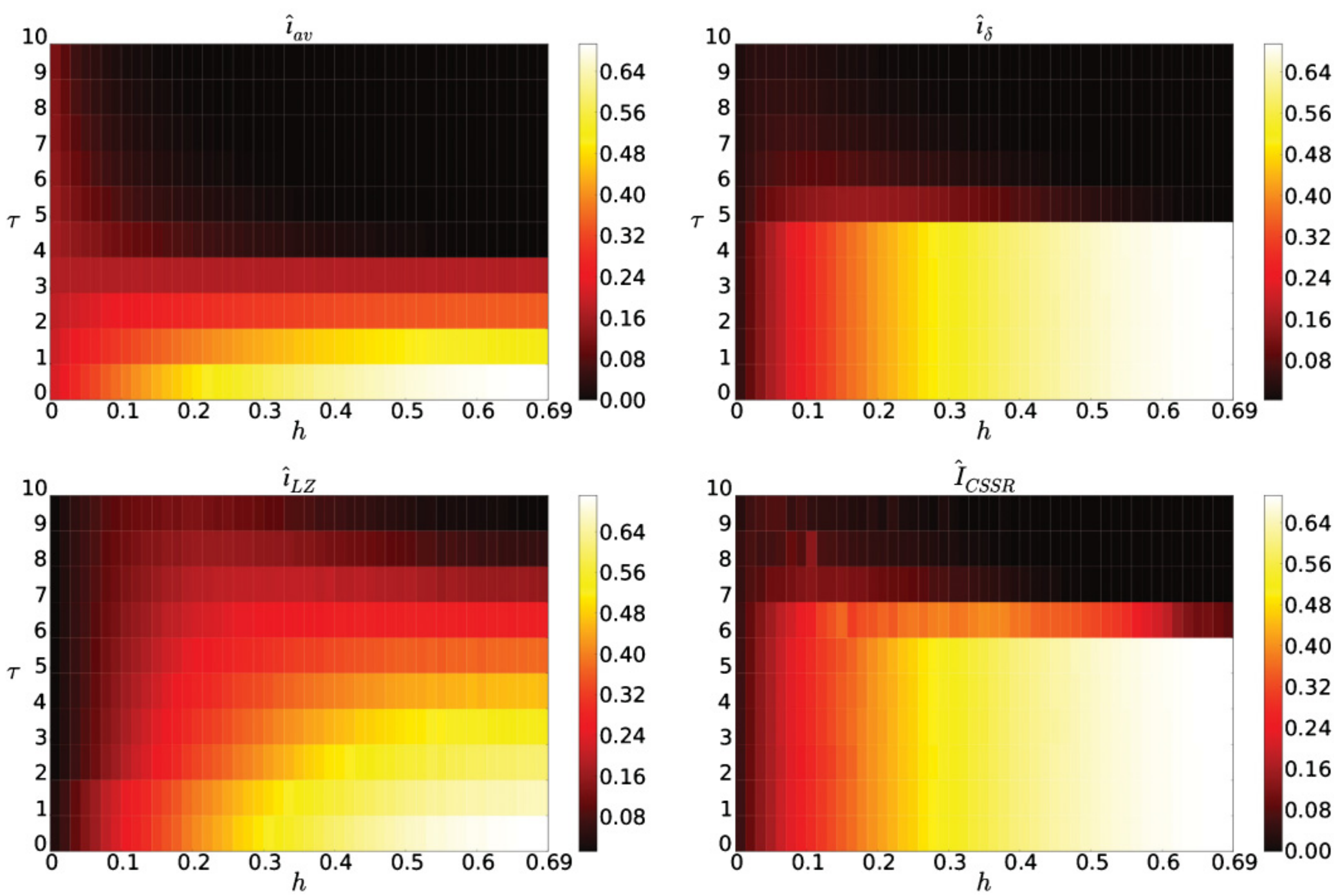
we suggest a two-step estimation: in a first step, an estimate of $h_{X,Y}$ is obtained at low computational cost using the LZ method. This preliminary knowledge guides the choice of the optimal value of n, namely, the larger value for which statistical errors are controlled, which is satisfied provided $n \log k_z e^{nh_{X,Y}} \leqslant N h_{X,Y}$ [29].

Excepting special instances or coincidental cancellations, MIR block estimators are biased ($\langle \widehat{\iota}_{n,\delta} \rangle \neq I_{n+1} - I_n$ and $\langle \widehat{\iota}_{n,\text{av}} \rangle \neq I_n/n$) due to the negative bias affecting block entropy estimators [29]. Correction of the bias proposed in Refs. [40] and [41] is derived for entropy assuming an uncorrelated sample, and hence do not apply in general in our context. In our simulations where $n = 4$, $N = 10^5$, and $k = 2$, the Miller and Madow correction [42] is of order 10^{-3} and can be neglected. Asymptotic bounds rigorously derived in Ref. [43] are not straightforwardly relevant for finite-size sequences.

Fitting a causal state machine model to experimental data depends on the maximum history length l, constrained by the finite length of the data sequence. In the following investigations, CSSR estimation will be done using $l = 5$ according to the heuristic procedure associated with the algorithm. In case of context trees or causal state machines estimation method, bootstrap methods (i.e., resampling by simulating the reconstructed process) can be developed to determine precise confidence interval [37]. A simpler clue to estimation errors accounting for the dependence among the observations is obtained by replacing the actual sequence length N by an effective length N_{eff} proportional to the sequence length N and the entropy rate h (roughly $N_{\text{eff}} = hN/\log k$ for an alphabet of size k) [29].

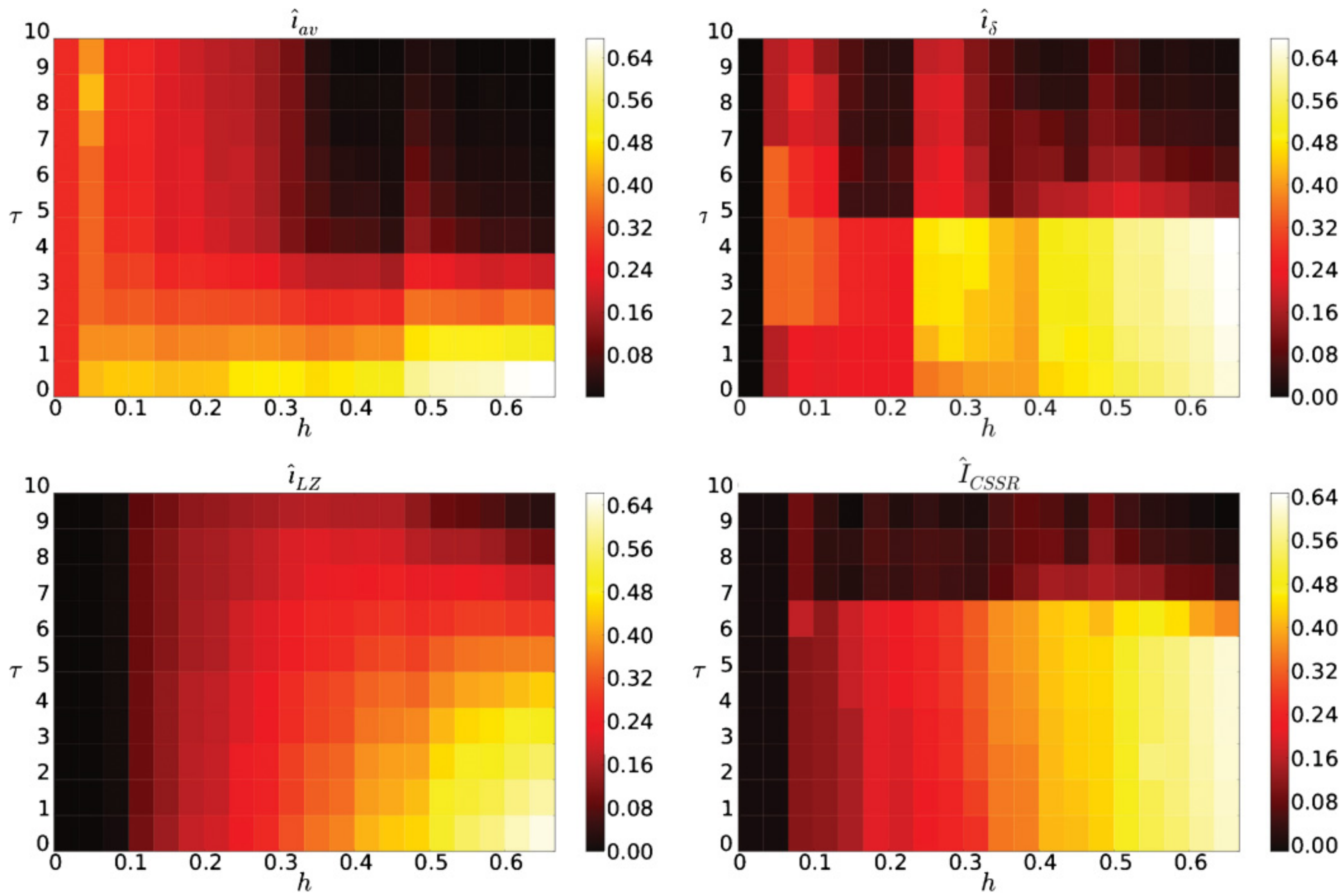
C. Influence of an input-output delay

We investigated discrepancy between the asymptotic delay independence of MIR and the behavior of finite-size estimators $\widehat{\iota}(X;Y)$ in the simple case where Y is a time-shifted version of X, i.e., $Y = \theta_\tau X$. We considered two situations where X presents a specific temporal organization, namely, a Markov



.

FIG. 2. (Color online) MIR of time-shifted binary Markov chains of length $N = 10^5$ as a function of the entropy rate h (horizontal axis) and the time delay τ (vertical axis). The first sequence is generated with a transition matrix control parameter $a \in [0,1/2]$, and the second one is obtained by shifting the first one by a lag τ varying from 0 to 10. The entropy rate, known analytically as a function $h(a)$, is taken as a measure of the temporal correlations. It is scanned unevenly when a is varied at constant step $\delta a = 0.01$, hence represented with a varying step δh on the linear scale. The color (grayscale) code used for the values of i is defined in the color (grayscale) bar on the right. MIR estimation is implemented using $\widehat{\iota}_{\text{av}} = I_n/n$ with $n = 4$ (top left), $\widehat{\iota}_\delta = I_{n+1} - I_n$ (top right), LZ algorithm (bottom left), and CSSR algorithm with $l = 5$ (bottom right).



.

FIG. 3. (Color online) MIR of time-shifted binary sequences of length $N = 10^5$ in the case of an underlying deterministic dynamics, as a function of the entropy rate h (horizontal axis) and the time delay τ (vertical axis). The first sequence is generated from a logistic map discretized according to the partition $[0,1/2[\cup[1/2,1]$, and the second one by shifting the first one with a lag τ varying from 0 to 10. The entropy rate is taken as a measure of the temporal correlations of the source; its value is computed as the Lyapunov exponent of the map from a very long typical trajectory (hence with a very good accuracy). It is scanned unevenly when a is varied at constant step $\delta a = 0.01$, hence represented with a varying step δh on the linear scale. The color (grayscale) code used for the values of i is defined in the color (grayscale) bar on the right. MIR estimation is implemented using $\widehat{\imath}_{\rm av} = I_n/n$ (top left), $\widehat{\imath}_\delta = I_{n+1} - I_n$ with $n = 4$ (top right), LZ algorithm (bottom left), and CSSR algorithm with $l = 5$ (bottom right).

chain and a discretized logistic map trajectory. In both cases $h(X)$ is taken as a measure of the temporal correlations in X and MIR is studied as a function of h and τ.

1. *Shifted stochastic sequences (Markov chains)*

We first consider the case where X is a first-order Markov chain with transition matrix:

$$M(a) = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix}. \tag{13}$$

$H_1(a) = \log 2$ and $h(a) = -a\log a - (1-a)\log(1-a)$ for this process. Results are presented in Fig. 2 for $n = 4$ and $l = 5$. MIR delay independence is observed with $\widehat{\imath}_{n,\delta}$ and $\widehat{\imath}_{l,\rm CSSR}$ provided $\tau \leqslant n$ and $\tau \leqslant l$, respectively. For $\tau > n$ or $\tau > l$, the estimation is dramatically impaired, and $\widehat{\imath}_{n,\delta}(a,\tau)$, $\widehat{\imath}_{l,\rm CSSR}(a,\tau)$ markedly differ from $i(a,\tau) = h(a)$. The upper limit $n^* \sim (\log N)/h$ on n imposed to get a proper statistical quality for block-entropy estimation (Sec. III B) also bounds above the time lags for which correlation between time-shifted sequences is well captured by $\widehat{\imath}_{n,\delta}$. At low h, a larger block size n could possibly be considered, which enlarges the range of time lags for which delay independence is satisfied in practice.

The same requirement $\tau \leqslant n$ is also observed with $\widehat{\imath}_{n,\rm av}$, but this estimator performs less satisfactorily than $\widehat{\imath}_\delta$. For $\tau > n$, $\widehat{\imath}_{n,\rm av}(a,\tau)$ could even be negative, whereas the exact value $i(a,\tau) = h(a)$ is always positive, leading us to definitely drop this estimator.

LZ estimation displays the merit of a control-setting-free estimator: It is less accurate than $\widehat{\imath}_{n,\delta}$ and $\widehat{\imath}_{l,\rm CSSR}$ in their range of validity, but it properly evidences the existence of a correlation between the original and shifted sequences in a wider range of delay values.

Overall, these results show that estimated MIR captures the correlation between the original and shifted sequences, more or less accurately according to the estimator, whereas this full correlation is lost when considering classic mutual information.

2. Shifted discretized trajectories of a logistic map

We then considered a sequence X obtained from the discretization,[5] according to the generating partition $[0,1/2[\cup[1/2,1]$, of the deterministic dynamics generated by the one-parameter family of logistic maps in $[0,1]$:

$$x_{n+1} = f_a(x_n) \equiv a x_n(1 - x_n) \tag{14}$$

with control parameter $a \in [3.5,4]$. Results are presented in Fig. 3 and are qualitatively comparable to those obtained in the Markovian case. Block and CSSR estimators properly evidence the correlation in the domain where the block size or the maximum history length is longer than the delay. LZ estimator reveals a correlation between the original and the shifted sequence in a wider range for the delay.

D. Coupled systems

We now consider the case of a process Z whose components X and Y evolve in a coupled fashion.

1. Symmetrically coupled logistic maps

A two-dimensional discrete-time system is obtained by a symmetric coupling of logistic maps as follows:

$$\begin{aligned} x_{n+1} &= \gamma f_a(x_n) + (1-\gamma) f_a(y_n), \\ y_{n+1} &= (1-\gamma) f_a(x_n) + \gamma f_a(y_n), \end{aligned} \tag{15}$$

where γ is the coupling coefficient and varies between 0 and 1. The contribution h_X (resp. h_Y) to MIR is the entropy rate of the projection X (resp. Y) differing from the entropy rate of the source X (resp. Y) in the absence of coupling.

According to Ref. [45], the overall dynamics of the coupled maps can be broadly divided into seven zones as in Fig. 4. The transition from one zone to another is marked by a bifurcation with control parameter γ. There is a quasisymmetry about $\gamma = 1/2$, so we can restrict to the four zones observed for $\gamma \in [0,1/2]$: Aone I is a region of complex dynamics, mostly chaotic but with several periodic orbits; zone II is characterized by a periodic behavior; zone III is completely chaotic; zone IV is also chaotic, but the maps X and Y are then perfectly synchronized.

In Fig. 4 we observe that MIR qualitatively follows the behavior of the difference $X - Y$ (displayed on the lower panel of the bifurcation diagram). Satisfactorily, a high value of MIR is observed when the two components X and Y are synchronized (i.e., the difference $X - Y$ is null). Here both the LZ estimator and CSSR estimator give the most faithful results.

2. Coupled time-uncorrelated sequences with delay

We finally consider the case where X is a time-uncorrelated equiprobable binary sequence and Y a randomly flipped version of X[6] (with probability ϵ), moreover shifted by a delay τ. This setting is suitable to study estimated MIR as a function of both the statistical correlation, as measured by the correlation coefficient $r = 1 - 2\epsilon$, and the delay τ between the sources X and Y. Results are presented in Fig. 5. The statistical dependence between X and Y is not captured by mutual information as soon as $\tau > 0$, nor by block estimators for $\tau > n$. For $\tau \leqslant n$, the difference block estimator $\widehat{\imath}_{n,\delta}$ behaves as theoretically expected (delay independence) and faithfully captures the correlation between the sequences. The same criterion $\tau \leqslant n$ is also required with $\widehat{\imath}_{n,\mathrm{av}}$, but this estimator performs less satisfactorily than $\widehat{\imath}_\delta$. The CSSR estimator gives good results within the domain $\tau \leqslant l_{\max}$. The LZ estimator $\widehat{\imath}_{lz}$ smoothly decreases, corresponding to a smooth loss of the ability to detect coupling as the delay increases. Overall, the LZ estimator reflects the presence of the coupling for a wider range of r and τ values, with no need of prior knowledge.

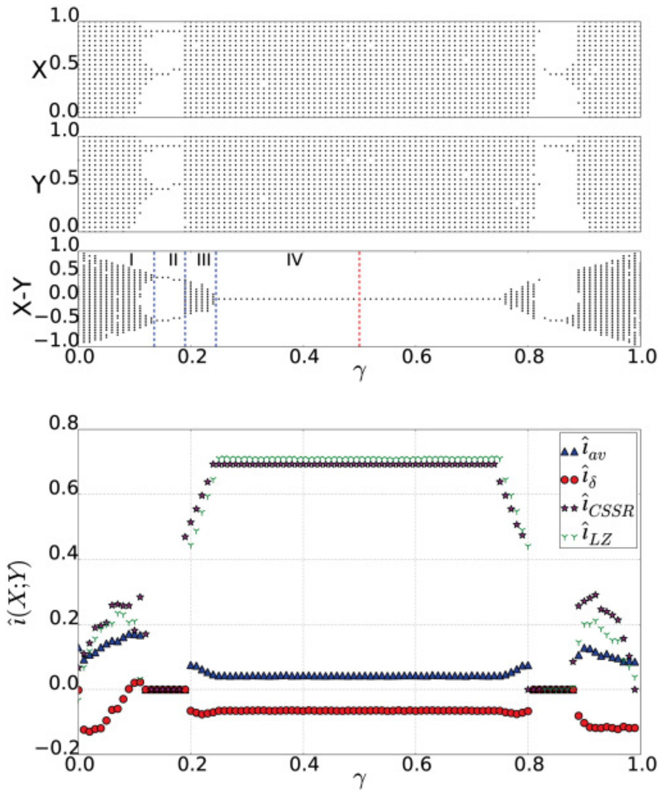


FIG. 4. (Color online) Top panel: Bifurcation diagram of two coupled logistic maps (for a control parameter value $a = 4$ with initial conditions $x_0 = 0.1$ and $y_0 = 0.75$, corresponding to fully chaotic maps when uncoupled) as a function the coupling coefficient γ. The vertical dashed lines delineate the bifurcation points and the regions (I, II, III, IV) with qualitatively different dynamic features. Bottom panel: MIR estimated from the discretized trajectories of these coupled maps as a function of the coupling coefficient γ.

[5]We do not discuss here the discretization of a continuous trajectory. Most often, generating partitions are unknown or even do not exist, and the discretization is rather chosen according to its significance with respect to the investigated data set and questions, in particular to buffer individual variability [20,44]. The dependence of MIR on the partition chosen for discretizing the underlying continuous state space is discussed with full mathematical rigor in Ref. [13].

[6]Each symbol of the realization $(x_t)_t$ of X is inverted ($0 \to 1$ and $1 \to 0$) independently of the others with a probability ϵ.

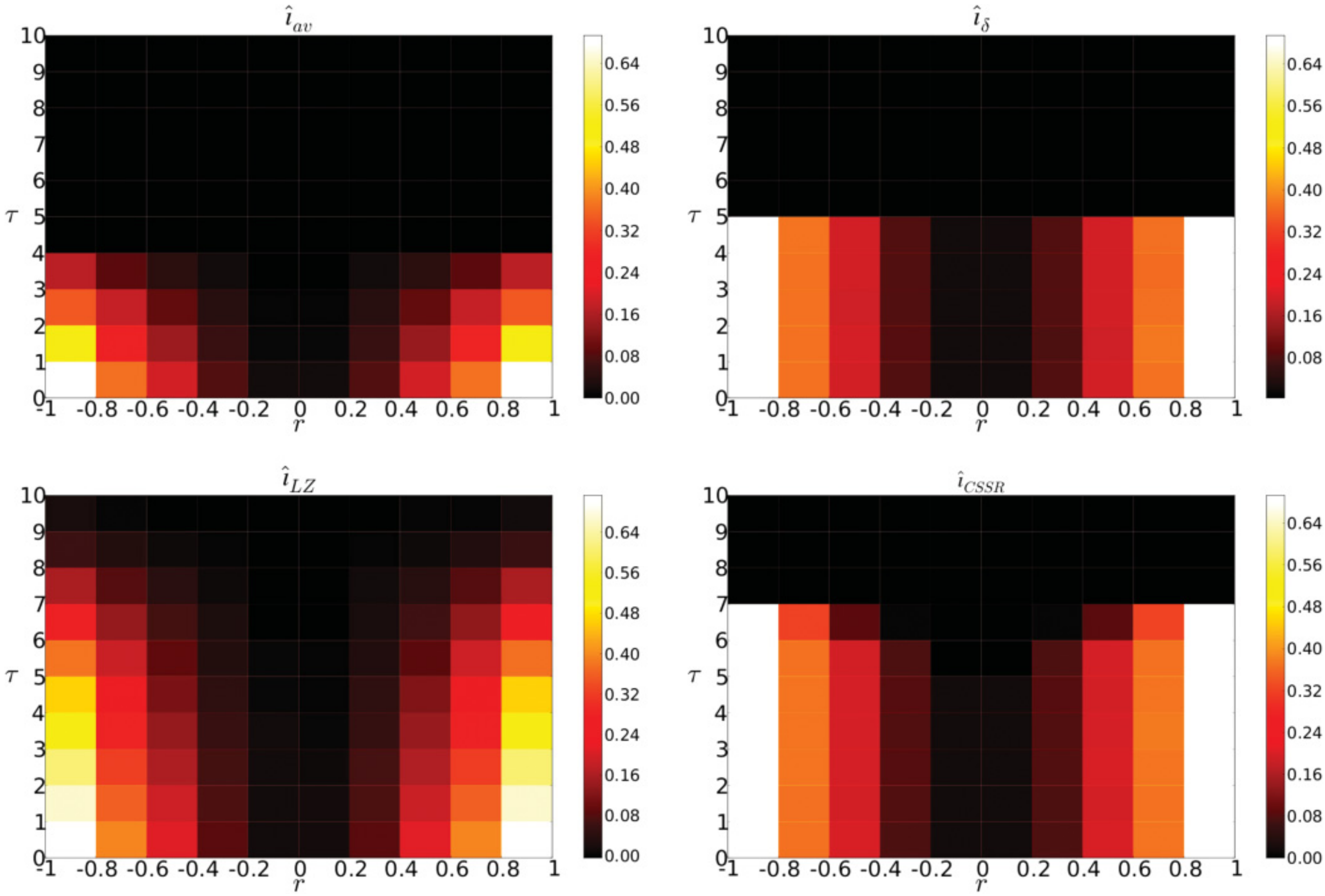


FIG. 5. (Color online) Mutual information rate estimation of two time-uncorrelated binary sequences X and Y (length $N = 10^5$) as a function of their correlation coefficient r (horizontal axis) and the time delay τ between them (vertical axis). The estimation is performed considering either block estimators: $\widehat{\iota}_{\text{av}} = I_n/n$ (top left) and $\widehat{\iota}_{\delta} = I_{n+1} - I_n$ (top right) for a word length $n = 4$, or LZ estimator (bottom left), CSSR algorithm with $l = 5$ (bottom right). The color (grayscale) code used for the values of i is defined in the color (grayscale) bar on the right.

IV. CONCLUSION

We have investigated analytically and numerically properties of MIR pertinent to its use in time-series analysis. As for other entropy-based indexes, no prior assessment of the deterministic nature of the evolution is needed. MIR thus offers a unified framework for analyzing both deterministic and stochastic evolutions. It is able to capture the degree of temporal organization of a coupled system by measuring how the coupling creates redundancy in the information production of the two sources. Since the theoretical behavior of MIR is a strict delay independence, this index is especially well suited to quantify mutual information of two processes, as a whole, when an unknown time shift or dephasing is present between them. In such a situation, delayed mutual information is not advisable because it misleadingly leads one to conclude that the two sources or signals are independent when it is computed with an ill-chosen delay. In practical applications, delay independence is satisfied by block- and CSSR-estimated MIR provided the delay is smaller than the block size and memory length, respectively. It is satisfied more roughly, but also more smoothly as a function of the delay by LZ-estimated MIR. We thus recommend the joint use of several estimation methods, especially in the exploratory analysis of the data, to benefit from their complementary advantages. These results demonstrate that mutual information rate provides a suitable index to evidence correlations between two signals despite a possible time shift or dephasing between them, in a wide range of experimental research domains.

ACKNOWLEDGMENTS

We warmly acknowledge the anonymous referees for constructive criticisms and suggestions, in particular indications for alleviating the proof of our independence theorem and several references. This study was partly supported by the NeuroInformatique interdisciplinary program 2008 of the Centre National de la Recherche Scientifique through the CODYSEP project.

APPENDIX: PROOF OF THE DELAY INDEPENDENCE THEOREM (SEC. II D)

For any processes X and Y with discrete values on alphabets $\mathcal{X}$ and $\mathcal{Y}$ (not necessarily identical) and any delay τ, we want to show that $i(X;\theta_\tau Y) = i(X;Y)$, where $\theta_\tau Y$ is the time-shifted version of Y, namely, $\theta_\tau Y(t) = Y(t+\tau)$. We introduce the standard notation X_i^j for abbreviating $X_i, X_{I+1}, \ldots, X_{j-1}, X_j$.

We recall that the statistical stationarity of a process Y ensures the time-shift invariance of its entropy rate, i.e., $h(Y) = h(\theta_\tau Y)$. It is a straightforward consequence of the asymptotic nature of h and its dependence on the whole process. Similarly, stationarity of the pair (X,Y) implies that $i(X,\theta_\tau Y) = i(\theta_{-\tau} X, Y)$, allowing us to restrict to the case $\tau > 0$ without loss of generality. By definition of the mutual information rate, $i(X;\theta_\tau Y) = h(X) + h(\theta_\tau Y) - h(X;\theta_\tau Y)$, while $i(X;Y) = h(X) + h(Y) - h(X;Y)$. By stationarity, $h(\theta_\tau Y) = h(Y)$, so it remains to show that the joint entropy rates are equal:[7]

$$
\begin{aligned}
h(X;\theta_\tau Y) &= \lim_{n\to\infty} \frac{H\left[X_1^n,(\theta_\tau Y)_1^n\right]}{n} = \lim_{n\to\infty} \frac{H\left[X_1^n, Y_{1+\tau}^{n+\tau}\right]}{n} \\
&= \lim_{n\to\infty} \frac{H\left[(X,Y)_{1+\tau}^n\right] + H\left[X_1^\tau, Y_{n+1}^{n+\tau} \,\middle|\, (X,Y)_{1+\tau}^n\right] Y_{1+\tau}^{n+\tau}\right]}{n}
\end{aligned}
$$

using the chain rule for entropy. Since

$$
\begin{aligned}
0 &\leqslant H\left[X_1^\tau, Y_{n+1}^{n+\tau} \,\middle|\, (X,Y)_{1+\tau}^n\right] Y_{1+\tau}^{n+\tau}\right] \\
&\leqslant H\left[X_1^\tau, Y_{n+1}^{n+\tau}\right] \leqslant \tau \log(k_x k_y),
\end{aligned}
$$

the second contribution to the limit vanishes, that is, $\lim_{n\to\infty}(1/n) H[X_1^\tau, Y_{n+1}^{n+\tau} \mid (X,Y)_{1+\tau}^n] Y_{1+\tau}^{n+\tau}] = 0$. Thus

$$
\begin{aligned}
h(X;\theta_\tau Y) &= \lim_{n\to\infty} \frac{H\left[(X,Y)_{1+\tau}^n\right]}{n} \\
&= \lim_{n\to\infty} \frac{H\left[(X,Y)_{1+\tau}^n\right]}{n-\tau} = h(X,Y),
\end{aligned}
$$

which completes the proof.

[7] We credit an anonymous referee for this elegant formulation simplifying our original proof.

[1] C. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[2] A. N. Kolmogorov, Dokl. Acad. Nauk SSSR **124**, 754 (1959).
[3] Ya. G. Sinai, Dokl. Acad. Nauk SSSR **124**, 768 (1959).
[4] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, England, 1997).
[5] T. M. Cover and J. A Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[6] S. Ihara, *Information Theory for Continuous Systems* (World Scientific, Singapore, 1993).
[7] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, translated and edited by A. Feinstein (Holden-Day, San Francisco, 1964).
[8] J. L. Good and D. B. Osteyee, *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*, Chap. 12 (Springer, Berlin, 1974).
[9] M. Paluš, V. Komárek, Z. Hrnčir, and K. Štěrbová, Phys. Rev. E **63**, 046211 (2001).
[10] M. Paluš, Physica D **93**, 64 (1996).
[11] R. L. Dobrushin, Usp. Mat. Nauk **14**, 3 (1959); Amer. Math. Soc. Trans. **33**, 323 (1959).
[12] R. Gray and J. Kieffer, IEEE Trans. Inf. Theory **26**, 412 (1980).
[13] R. M. Gray, *Entropy and Information Theory* (Springer, New York, 1990).
[14] S. Zozor, P. Ravier, and O. Buttelli, Phys. A **345**, 285 (2005)
[15] J. Shlens, M. B. Kennel, H. D. I. Abarbanel, and E. J. Chichilnisky, Neural Comput. **19**, 1683 (2007).
[16] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, Phys. Rev. E **65**, 041903 (2002).
[17] L. Breiman, Ann. Math. Statist. **28**, 809 (1957); **31**, 809 (1957).
[18] B. McMillan, Ann. Math. Stastist. **24**, 196 (1953).
[19] W. Li, J. Stat. Phys. **60**, 823 (1990).
[20] M. P. Paulus, M. A. Geyer, L. H. Gold, and A. J. Mandell, Proc. Natl. Acad. Sci. USA **87**, 723 (1990).
[21] B. Gaveau, A. Lesne, and L. S. Schulman, Phys. Lett. A **258**, 222 (1999).
[22] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA, 1997).
[23] C. R. Shalizi, K. L. Shalizi, and R. Haslinger, Phys. Rev. Lett. **93**, 118701 (2004).
[24] J. Delgado and R. V. Solé, Phys. Rev. E **55**, 2338 (1997).
[25] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
[26] W. Bialek, M. DeWeese, F. Rieke, and D. Warland, Phys. A **200**, 581 (1993).
[27] M. Paluš and M. Vejmelka, Phys. Rev. E **75**, 056211 (2007).
[28] M. B. Kennel, J. Shlens, H. D. I. Abarbanel, and E. J. Chichilnisky, Neural Comput. **17**, 1531 (2005).
[29] A. Lesne, J. L. Blanc, and L. Pezard, Phys. Rev. E **79**, 046208 (2009).
[30] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).
[31] C. R. Shalizi and K. L. Klinkner, *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)* (AUAI Press, Arlington, VA, 2004).
[32] R. Haslinger, K. L. Klinkner, and C. R. Shalizi, Neural Comput. **22**, 121 (2010).
[33] A. Lempel and J. Ziv, IEEE Trans. Inf. Theory **22**, 75 (1976).
[34] The source code and documentation of CSSR are available at [http://bactra.org/CSSR].
[35] I. Kontoyiannis, P. H. Algoet, Yu. M. Suhov, and A. J. Wyner, IEEE Trans. Inf. Theory **44**, 1319 (1998).
[36] P. Faure and A. Lesne, Int. J. Bif. Chaos **20**, 1731 (2010).
[37] P. Bühlmann and A. J. Wyner, Ann. Stat. **27**, 480 (1999).
[38] J. L. Blanc, N. Schmidt, L. Bonnier, L. Pezard, and A. Lesne, in *Proceedings of the Second French Conference on Computational Neuroscience (Neurocomp'08)*, edited by L. U. Perrinet and E. Daucé (Marseille, 2008), pp. 40–43.
[39] K. Marton and P. C. Shields, Ann. Prob. **22**, 960 (1994); **24**, 541 (1996).
[40] A. Treves and S. Panzeri, Neural Comput. **7**, 399 (1995).
[41] J. D. Victor, Neural Comput. **12**, 2797 (2000).
[42] L. Paninski, Neural Comput. **15**, 1191 (2003).
[43] J. R. Chazottes and D. Gabrielli, Nonlinearity **18**, 2545 (2005).
[44] K. Doba, L. Pezard, A. Lesne, V. Christophe, and J. L. Nandrino, Psych. Rep. **101**, 237 (2007).
[45] F. E. Udwadia and N. Raju, Phys. D **111**, 16 (1998).