RESEARCH ARTICLE | NOVEMBER 17 2016

Computing algebraic transfer entropy and coupling directions via transcripts

José M. Amigó ; Roberto Monetti; Beata Graff; Grzegorz Graff

Chaos 26, 113115 (2016)
https://doi.org/10.1063/1.4967803

Computing algebraic transfer entropy and coupling directions via transcripts

José M. Amigó,[1,a)] Roberto Monetti,[2,b)] Beata Graff,[3,c)] and Grzegorz Graff[4,d)]
[1]*Centro de Investigación Operativa, Universidad Miguel Hernández, 03202 Elche, Spain*
[2]*IngSoft GmbH, 90403 Nuernberg, Germany*
[3]*Department of Hypertension and Diabetology, Medical University of Gdansk, 80-952 Gdansk, Poland*
[4]*Faculty of Applied Physics and Mathematics, Gdansk University of Technology, 80-233 Gdansk, Poland*

(Received 25 May 2016; accepted 1 November 2016; published online 17 November 2016)

Most random processes studied in nonlinear time series analysis take values on sets endowed with a group structure, e.g., the real and rational numbers, and the integers. This fact allows to associate with each pair of group elements a third element, called their transcript, which is defined as the product of the second element in the pair times the first one. The transfer entropy of two such processes is called algebraic transfer entropy. It measures the information transferred between two coupled processes whose values belong to a group. In this paper, we show that, subject to one constraint, the algebraic transfer entropy matches the (in general, conditional) mutual information of certain transcripts with one variable less. This property has interesting practical applications, especially to the analysis of short time series. We also derive weak conditions for the 3-dimensional algebraic transfer entropy to yield the same coupling direction as the corresponding mutual information of transcripts. A related issue concerns the use of mutual information of transcripts to determine coupling directions in cases where the conditions just mentioned are not fulfilled. We checked the latter possibility in the lowest dimensional case with numerical simulations and cardiovascular data, and obtained positive results. *Published by AIP Publishing.*
[http://dx.doi.org/10.1063/1.4967803]

The elucidation of causality relations between interacting dynamical systems or random processes is an important though difficult undertaking in time series analysis. To this end a number of quantitative approaches have been proposed in the literature in which Wiener's criterion, based on predictability, was possibly the first one. In this paper we focus on transfer entropy, which can be considered for the implementation of Wiener's criterion which has become one of the most popular tools. Even in this limited setting, the determination of causality in practice remains subtle and challenging due to a number of factors such as (i) symbolization, (ii) observational noise, (iii) poor statistical estimations, or (iv) hidden drivers. The topic of this paper is related to (iii) since it deals with the possibility of computing transfer entropy and coupling directions with a lower dimensional quantity when the data being analyzed belong to an algebraic group. Such is the case of numerical observations and certain symbolic representations (e.g., ordinal patterns). This property can make a difference when studying information directionality with short time series. In any case, the use of directionality indicators with lower dimensionality enhances the statistical quality of the results.

I. INTRODUCTION

There are several ways of materializing the concept of causality between processes or systems wherein one of the proposals goes back to Wiener.[1] In short, a random process $(Y_n)_{n\in\mathbb{Z}}$ is said to (Wiener-)cause another process $(X_n)_{n\in\mathbb{Z}}$ if, given the past history of X_n, the additional knowledge of the past history of Y_n improves the prediction of the present value X_n. The implementation of this principle in linear time series analysis goes by the name of Granger causality.[2–4] In a deterministic setting, though, Granger causality may be problematic, as Granger himself recognized.[2] Causality is often referred to as coupling direction or temporal causality in nonlinear dynamics, where there are a number of alternative proposals for its determination. Just to mention a few, some use mutual prediction,[5,6] and others recurrences,[7] proximity of the embedding vectors in the reconstructed attractors,[8,9] or the so-called convergent cross mappings.[10] One has to move on to information theory to recover the original spirit of Wiener causality; this time it is formulated by means of asymmetric dependence measures. One of the most popular choices, since Schreiber[11] proposed transfer entropy to measure information transfer, has been conditional mutual information. This way one can quantify the reduction of uncertainty about X_n that supposes the additional knowledge of the past history of Y_n. If one repeats the calculation with the roles of the random processes $(X_n)_{n\in\mathbb{Z}}$ and $(Y_n)_{n\in\mathbb{Z}}$ interchanged and subtracts both results, the bottom line is a measure of the net amount of information between $(X_n)_{n\in\mathbb{Z}}$ and $(Y_n)_{n\in\mathbb{Z}}$. The sign of the net information transfer gives then the direction in which most information flows. If this direction goes, e.g., from $(Y_n)_{n\in\mathbb{Z}}$ to $(X_n)_{n\in\mathbb{Z}}$, then the former process drives the latter one more than in the opposite direction; that is, $(Y_n)_{n\in\mathbb{Z}}$ is the dominant or leading driver. For a general overview of causality detection based on information-theoretical methods, the interested reader is referred to Ref. 12.

[a)]jm.amigo@umh.es
[b)]r.monetti@gmail.com
[c)]bgraff@gumed.edu.pl
[d)]graff@mif.pg.gda.pl

In this paper, we follow the information-theoretical approach with a twist, since we are going to focus on the algebraic structure that the values of the processes considered or their symbolic representation might have. Formally, suppose that $\mathcal{G}$ is an algebraic group and $(X_n)_{n\in\mathbb{Z}}$ is a $\mathcal{G}$-valued random process. Any realization $(x_n)_{n\in\mathbb{Z}}$ is then called an algebraic time series. The examples of such algebraic time series are very common and include numerical observations, binning, instantaneous phases, and, of course, symbolic representations with elements of a group, e.g., permutations. Although $\mathcal{G}$ may have different cardinalities (finite, infinite countable, infinite uncountable) and topologies (discrete, continuous), in practice, the group elements with a positive probability (observable data) build a finite subset.

The transfer entropy of two random processes taking values on the same group will be called algebraic transfer entropy (ATE). Note that the qualifier "algebraic" adds nothing to the definition of transfer entropy; it only signalizes the fact that the data being analyzed are amenable to algebraic operations. In particular, the concept of *transcript* exploits this feature. Given two elements of the group $\mathcal{G}$, their transcript is the composition of the second element with the inverse of the first one, thus an element of $\mathcal{G}$ as well. However simple this concept may appear, it is fundamental for the core results of this paper. Transcripts were introduced in Ref. 13 for ordinal symbolic representations and further studied in Refs. 14–16. Transcripts in general algebraic time series representations were first considered in Ref. 17.

More importantly, in Ref. 17, it is shown that, under certain constraints, the algebraic (i.e., $\mathcal{G}$-valued) conditional multi-information can be calculated by a (in general conditional) mutual information of transcripts with one variable less. In this paper, we generalize this result along two lines. First of all, we consider a general algebraic transfer entropy (based on arbitrary, finite histories of $(X_n)_{n\in\mathbb{Z}}$ and $(Y_n)_{n\in\mathbb{Z}}$) instead. Therefore, the present results refer to ATEs of any dimension, while in Ref. 17, the connection to ATE could be established only in the lowest, three dimensional case. Second, the constraints that allow equating an ATE to a mutual information of transcripts of lower dimension (its "transcript dimensional reduction") are weaker than the constraints for the algebraic conditional multi-information. Specifically, two restrictions were needed in Ref. 17, to wit: one on the entropies of the processes and the other on the so-called coupling complexity coefficients. Here, only one restriction on the coupling complexity coefficients will be required. This is a fine result because, as we will review in Sec. V, that restriction can be often satisfied in practice by choosing the time delay adequately.

In sum, the results contained in this paper are not only interesting from a formal point of view but also from a practical one. Regarding this, we stress that the practical application of transfer entropy is challenging for a number of reasons that have been recently discussed in the literature. Among them, small data sets are, of course, directly related to the topic of this paper. Others include indirect influences,[18–20] redundancy,[21] low resolution,[19,22] anticipatory couplings,[20,22] choice of parameters,[23,24] and dominance of neighbors.[20] Possible remedies can be found in Ref. 25, as well as in the respective papers.

The rest of this paper is organized as follows. Sec. II contains the conceptual baggage needed to make this paper self-contained, most notably the concept of transcript (Sec. II B) and coupling complexity coefficients (Sec. II C). Once the notational and conceptual framework has been set, we derive in Sec. III the main theoretical result of the paper, Theorem 1, which states that an n-dimensional ATE equals an $(n-1)$-dimensional conditional mutual information of transcripts, subject to a constraint involving coupling complexity coefficients. Some results of Sec. III are then generalized in Sec. IV in that their hypotheses are weaker but still sufficient to guarantee that a given ATE and its transcript dimensional reduction yield the same coupling direction. The practical implementation of Theorem 1 is discussed in Sec. V. Finally, we illustrate the performance of the transcript 2-dimensional reduction as a directionality indicator in cases in which the constraints required in Secs. III and IV are not fulfilled. This is done with numerical simulations (Sec. VI A) and cardiovascular data (Sec. VI B).

II. THEORETICAL FRAMEWORK

The mathematical framework of our analysis is set by random variables and stationary random processes taking values on an algebraic group. The concept of transcript, presented in Sec. II B, is a simple way of exploiting that algebraic structure. Transcripts go into the definition of coupling complexity coefficients, Sec. II C, which are necessary for the main results of this paper.

A. Algebraic time series

Let $\mathbf{x} = (x_n)_{n\in\mathbb{T}}$ be a one sided ($\mathbb{T} = \mathbb{N}_0 := \{0\} \cup \mathbb{N}$) or two-sided ($\mathbb{T} = \mathbb{Z}$) time series taking values on a set $\mathcal{G}$. We say, that $\mathbf{x}$ is an *algebraic time series* if $\mathcal{G}$ is an algebraic group. Examples of groups that appear in applications are the following.

(G1): If x_n are measurements with finite precision, binned observations, or digitalized analog signals, then $\mathcal{G} = \mathbb{Z}$ or $\mathbb{Q}_k$, the rational numbers with $k \geq 1$ decimal digits. The composition law is the addition. Measurements with ideally infinite precision correspond to $\mathcal{G} = \mathbb{R}$.
(G2): Binary sequences are used in digital communications and cryptography. In this case, $\mathcal{G} = \{0, 1\}$ equipped with the XOR addition ($0 \oplus 0 = 1 \oplus 1 = 0$, $0 \oplus 1 = 1 \oplus 0 = 1$).
(G3): If x_n are instantaneous phases, then $\mathcal{G} = ([0, 1) \bmod 1)$. In practice, these phases will be also rational numbers with k decimal digits in the interval [0, 1), the composition law being addition modulo 1.
(G4): If $\mathbf{x}$ is symbolically represented by the elements of a group, then we obtain an *algebraic representation*. In particular, an *ordinal representation* is an algebraic representation by means of ordinal patterns[26,28] or, for that matter, permutations of length $L \geq 2$. Specifically, we trade off each $x_n \in \mathbb{R}$ for the rank vector $r = (\rho_0, \ldots, \rho_{L-1})$ of the block $x_n, x_{n+1} \ldots, x_{n+L-1}$, i.e., the permutation $(\rho_0, \ldots, \rho_{L-1})$ of $(0, \ldots, L-1)$ is the ordinal L-pattern r of x_n if

$$x_{n+\rho_0} < x_{n+\rho_1} < \cdots < x_{n+\rho_{L-1}}.$$

Thus, in this case, $\mathcal{G} = \mathcal{S}_L$, the symmetric group of degree $L \geq 2$. The composition law is the composition or "product" of permutations.[14] This group is non-commutative for $L \geq 3$. Other definitions, e.g., using the block $x_{n-L+1}, \ldots, x_{n-1}, x_n$, are also possible.

We assume throughout this paper that $\mathbf{x}$ is the output of a *stationary* random process $\mathbf{X} = (X_n)_{n\in\mathbb{T}}$ (possibly a noisy deterministic or "coarse grained" dynamics) or a symbolic representation thereof. Furthermore, we may assume that only a finite number of group elements have a positive probability because this is what happens in practice. Therefore, all the entropies considered below are finite (otherwise, we should suppose case by case that the corresponding numerical series is convergent). As an additional remark, finite groups may have also elements with zero probability. For instance, time series which are generated by piecewise strictly monotonic interval maps have forbidden ordinal patterns for sufficiently large pattern lengths.[27,28]

B. Transcripts

Let $\mathcal{G}$ be a group. Given $x, y \in \mathcal{G}$, there always exists a unique $t = t(x, y) \in \mathcal{G}$, called the *transcript from the element x to the element y*,[13] such that

$$t = yx^{-1}. \tag{1}$$

For generality, we use a multiplicative notation, i.e., the inverse $x \in \mathcal{G}$ is written as x^{-1}, and the composition or "product" of two elements of $\mathcal{G}$ is denoted by concatenating them in the correct order (since $\mathcal{G}$ might be non-commutative). When the elements being multiplied, x and y in (1), are important for the discussion, we write $t_{x,y}$. It follows from (1) that

$$t_{y,x} = (t_{x,y})^{-1}, \tag{2}$$

and

$$t_{y,z}t_{x,y} = zy^{-1}yx^{-1} = zx^{-1} = t_{x,z}.$$

In case that $\mathcal{G}$ has a finite cardinality $|\mathcal{G}|$, the map $\mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined by $(x, y) \mapsto t_{x,y}$ is $|\mathcal{G}|$-to- 1 since the $|\mathcal{G}|$ distinct pairs (x, tx) are sent to t for all $x \in \mathcal{G}$. Reciprocally, any pair $(x, y) \in \mathcal{G} \times \mathcal{G}$ whose transcript is t must have $y = tx$ by (1).

As two simple examples of transcripts, consider (i) $\mathcal{G} = \mathbb{R}$ (**G1** in Sec. II A) and (ii) $\mathcal{G} = \mathcal{S}_L$ (**G4** in Sec. II A). Then, $t_{x,y} = y - x$ in case (i). As for (ii), if $x = (\xi_0, \ldots, \xi_{L-1})$ and $y = (\eta_0, \ldots, \eta_{L-1})$, then

$$\begin{aligned} t_{x,y} &= (\eta_0, \ldots, \eta_{L-1})(\xi_0, \ldots, \xi_{L-1})^{-1} \\ &= (\eta_0, \ldots, \eta_{L-1})(\pi_0, \ldots, \pi_{L-1}) \\ &= (\pi_{\eta_0}, \ldots, \pi_{\eta_{L-1}}), \end{aligned}$$

where $(\pi_0, \ldots, \pi_{L-1})$ is the ordinal pattern of the string $\xi_0, \ldots, \xi_{L-1}$.[14]

C. Coupling complexity coefficients

The *coupling complexity coefficient* of the $\mathcal{G}$-valued random variables $X_1, \ldots, X_N$ ($N \geq 2$), denoted by $C(X_1, \ldots, X_N)$, is directly linked to their transcripts. They are defined as[14,15]

$$\begin{aligned} C(X_1, X_2, \ldots, X_N) = \min_{1\leq n\leq N} H(X_n) - H(X_1, X_2, \ldots, X_N) \\ + H(T_{x_1,x_2}, T_{x_2,x_3}, \ldots, T_{x_{N-1},x_N}), \end{aligned} \tag{3}$$

where $H(\ldots)$ is the (joint) entropy of the random variable(s) in the argument, and $T_{x_i,x_{i+1}}$ is a random variable that outputs the transcript $t_{x_i,x_{i+1}}$ whenever $X_i = x_i$ and $X_{i+1} = x_{i+1}$. It can be shown[15,17] that

$$C(X_1, X_2, \ldots, X_N) = \min_{1\leq n\leq N} I(X_n; T_{x_1,x_2}, T_{x_2,x_3}, \ldots, T_{x_{N-1},x_N}), \tag{4}$$

where $I(\cdot;\cdot)$ is the mutual information between the random variables in the argument separated by the semicolon. These coefficients were introduced in Ref. 14 to discriminate different synchronization regimes in coupled dynamics.

The coefficients $C(X_1, \ldots, X_N)$ have a number of interesting properties. Among the basic ones, they are invariant under permutation of their arguments. The next property will be needed in Sec. IV. We use the notation $C(X_1, \ldots, \bar{X}_k, \ldots, X_N)$ to indicate that the variable X_k has been omitted in the argument of $C(X_1, \ldots, X_k, \ldots, X_N)$. By convention we set

$$C(X) := 0 \tag{5}$$

to cover the one-dimensional case $C(X_1)$ or $C(X_2)$ of $C(X_1, \ldots, \bar{X}_k, \ldots, X_N)$ when $N = 2$.

Lemma 1. *(Monotonicity property of the coupling complexity coefficients.) Consider the $\mathcal{G}$- valued random variables $X_1, \ldots, X_N$, where $N \geq 2$.*

(i) *If there is only one variable $X_{n_{\min}}$ such that $\min_{1\leq n\leq N} H(X_n) = H(X_{n_{\min}})$, then $C(X_1, \ldots, X_N) \geq C(X_1, \ldots, \bar{X}_k, \ldots, X_N)$ for all $k \neq n_{\min}$.*

(ii) *If there are at least two variables with minimum entropy, then $C(X_1, \ldots, X_N) \geq C(X_1, \ldots, \bar{X}_k, \ldots, X_N)$ for all $k = 1, \ldots, N$.*

We conclude from Lemma 1 that $C(X_1, \ldots, X_N) < C(X_1, \ldots, \bar{X}_k, \ldots, X_N)$ is only possible when $H(X_k)$ is a unique minimum. For the proofs of the above and other properties of the coupling complexity coefficients, the interested reader is referred to Refs. 15 and 17.

To conclude this section, let us also mention the *equivalence property*. By this, we mean the fact that different sets of variables composed of group elements and their transcripts contain the same information just because the variables in one set univocally determine the variables in the other sets. Thus, given the triple $(x, y, t_{x,y})$, any pair of elements, i.e., (x, y), $(x, t_{x,y})$, or $(y, t_{x,y})$, univocally determines the remaining element. The same happens with $(x, y, t_{y,x})$ because of (2). This simple observation implies

$$H(\ldots, X, Y, \ldots) = H(\ldots, X, T_{x,y}, \ldots) = H(\ldots, X, T_{y,x}, \ldots) \tag{6}$$

$$= H(\ldots, T_{x,y}, Y, \ldots) = H(\ldots, T_{y,x}, Y, \ldots) \tag{7}$$

because any of the random variable pairs explicitly shown in (6) and (7) can be determined from any other variable pair. A similar property holds with higher numbers of symbols and their transcripts.

III. COMPUTING ALGEBRAIC TRANSFER ENTROPY VIA TRANSCRIPTS

Let $\mathcal{G}$ be a group, and $\mathbf{X} = (X_n)_{n\in\mathbb{Z}}$ and $\mathbf{Y} = (Y_n)_{n\in\mathbb{Z}}$ two $\mathcal{G}$-valued, stationary random processes. Roughly speaking, the transfer entropy from $\mathbf{Y}$ to $\mathbf{X}$ is the reduction of uncertainty in future values of $\mathbf{X}$, given past values of $\mathbf{X}$, due to the additional knowledge of past values of $\mathbf{Y}$. We remind next its standard definition[11] by means of conditional entropies; thereby, we introduce the term "algebraic transfer entropy" to underline that the values of X_n and Y_n have an extra algebraic structure. For r, $s \geq 1$ set $X_n^{(r)} := X_n, \ldots, X_{n-r+1}$, and $Y_n^{(s)} := Y_n, \ldots, Y_{n-s+1}$.

Definition 1. *The* algebraic transfer entropy from the process $\mathbf{Y}$ to the process $\mathbf{X}$ *with coupling delay* $\Lambda \geq 1$ *is defined as*

$$\begin{aligned} AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) &= H\left(X_{n+\Lambda}|X_n^{(r)}\right) - H\left(X_{n+\Lambda}|X_n^{(r)}, Y_n^{(s)}\right) \\ &= \sum_{x_{n+\Lambda}, x_n^{(r)}, y_n^{(s)}} p\left(x_{n+\Lambda}, x_n^{(r)}, y_n^{(s)}\right) \\ &\quad \times \log \frac{p\left(x_{n+\Lambda}|x_n^{(r)}, y_n^{(s)}\right)}{p\left(x_{n+\Lambda}|x_n^{(r)}\right)}. \end{aligned} \tag{8}$$

Therefore, $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = 0$ if and only if $p(x_{n+\Lambda}|x_n^{(r)}, y_n^{(s)}) = p(x_{n+\Lambda}|x_n^{(r)})$ for all $x_{n+\Lambda}$, $x_n^{(r)} := x_n, \ldots, x_{n-r+1}$, and $y_n^{(s)} := y_n, \ldots, y_{n-s+1}$. If, otherwise, $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) > 0$ we say, that there is an information transfer from the process $\mathbf{Y}$ to the process $\mathbf{X}$ with coupling delay Λ. In Ref. 11 and often in applications, $\Lambda = 1$. Ideally, one should set $s = r = -\infty$ in (8) but, bearing in mind the applications, it is more convenient to consider rather finite histories.

Alternatively, $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ can be also written as conditional mutual information, to wit:

$$\begin{aligned} AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) &= I\left(X_{n+\Lambda}; Y_n^{(s)}|X_n^{(r)}\right) \\ &= \sum_{x_{n+\Lambda}, x_n^{(r)}, y_n^{(s)}} p\left(x_{n+\Lambda}, x_n^{(r)}, y_n^{(s)}\right) \\ &\quad \times \log \frac{p\left(x_{n+\Lambda}, y_n^{(s)}|x_n^{(r)}\right)}{p\left(x_{n+\Lambda}|x_n^{(r)}\right) p\left(y_n^{(s)}|x_n^{(r)}\right)}. \end{aligned} \tag{9}$$

Other definitions of transfer entropy (e.g., with $n-1$ instead of n, or with $s = r$) can be found in the literature too. For definiteness, logarithms are taken to be base 2 throughout the paper.

Unlike the (unconditioned) mutual information, $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ is not symmetric under the exchange of the processes $\mathbf{X}$ and $\mathbf{Y}$. Therefore, the *directionality indicator*

$$\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) - AT^{(r,s)}_{\mathbf{X}\to\mathbf{Y}}(\Lambda) = -\Delta AT^{(r,s)}_{\mathbf{X}\to\mathbf{Y}}(\Lambda) \tag{10}$$

measures the *net* transfer of information between the processes $\mathbf{X}$ and $\mathbf{Y}$. For example, if $\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) > 0$, then $\mathbf{Y}$ is the dominant driving process with a coupling delay Λ. This so-called coupling direction is one of the main objectives in the study of interacting systems.

The next lemma relates $\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ to transcripts of $X_{\Lambda+n}$, $X_n^{(r)}$, and $Y_n^{(s)}$ via coupling complexity coefficients. For notational convenience, we introduce the symbols

$$\begin{aligned} C(X_1, \ldots, X_N|Y_1, \ldots, Y_M) &:= C(X_1, \ldots, X_N, Y_1, \ldots, Y_M) \\ &\quad - C(Y_1, \ldots, Y_M). \end{aligned} \tag{11}$$

Lemma 2. If $r \geq 2$,

$$\begin{aligned} AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) &= I(T_{x_{n+\Lambda},x_n}; T_{x_{n-r+1},y_n}, T_{y_n,y_{n-1}}, \ldots, \\ &\quad T_{y_{n-s+2},y_{n-s+1}}|T_{x_n,x_{n-1}}, \ldots, T_{x_{n-r+2},x_{n-r+1}}) \\ &\quad + C(X_{n+\Lambda}|X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1}) \\ &\quad - C(X_{n+\Lambda}|X_n, \ldots, X_{n-r+1}). \end{aligned} \tag{12}$$

If $r = 1$,

$$\begin{aligned} AT^{(s,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) &= I(T_{x_{n+\Lambda},x_n}; T_{x_n,y_n}, T_{y_n,y_{n-1}}, \ldots, T_{y_{n-s+2},y_{n-s+1}}) \\ &\quad + C(X_{n+\Lambda}|X_n, Y_n, \ldots, Y_{n-s+1}) - C(X_{n+\Lambda}|X_n). \end{aligned} \tag{13}$$

Proof. Suppose $r \geq 2$. First of all, note that $H(X_{n+\Lambda}) = H(X_n) = \cdots = H(X_{n-r+1})$, and $H(Y_n) = \cdots = H(Y_{n-s+1})$. From (3) and (11), it follows:

$$\begin{aligned} &H(X_{n+\Lambda}|X_n, \ldots, X_{n-r+1}) \\ &\quad = H(X_{n+\Lambda}, X_n, \ldots, X_{n-r+1}) - H(X_n, \ldots, X_{n-r+1}) \\ &\quad = H(T_{x_{n+\Lambda},x_n}|T_{x_n,x_{n-1}}, \ldots, T_{x_{n-r+2},x_{n-r+1}}) \\ &\qquad - C(X_{n+\Lambda}|X_n, \ldots, X_{n-r+1}), \end{aligned}$$

since

$$\min\{H(X_{n+\Lambda}), H(X_n), \ldots, H(X_{n-r+1})\} = H(X_n)$$

and

$$\min\{H(X_n), \ldots, H(X_{n-r+1})\} = H(X_n).$$

Likewise,

$$\begin{aligned} &H(X_{n+\Lambda}|X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1}) \\ &\quad = H(X_{n+\Lambda}, X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1}) \\ &\qquad - H(X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1}) \\ &\quad = H(T_{x_{n+\Lambda},x_n}|T_{x_n,x_{n-1}}, \ldots, T_{x_{n-r+2},x_{n-r+1}}, T_{x_{n-r+1},y_n}, \\ &\qquad T_{y_n,y_{n-1}}, \ldots, T_{y_{n-s+2},y_{n-s+1}}) \\ &\qquad - C(X_{n+\Lambda}|X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1}), \end{aligned}$$

since

$$\begin{aligned} &\min\{H(X_{n+\Lambda}), H(X_n), \ldots, H(X_{n-r+1}), H(Y_n), \ldots, H(Y_{n-s+1})\} \\ &\quad = \min\{H(X_n), H(Y_n)\} \end{aligned}$$

and

$$\min\{H(X_n),\dots,H(X_{n-r+1}),H(Y_n),\dots,H(Y_{n-s+1})\} = \min\{H(X_n),H(Y_n)\}.$$

Plug now these two expressions into the definition (8).

The case $r=1$ follows analogously. □

The following result is a straightforward consequence of Lemma 2.

Theorem 1. If

$$C(X_{n+\Lambda}|X_n,\dots,X_{n-r+1},Y_n,\dots,Y_{n-s+1}) = C(X_{n+\Lambda}|X_n,\dots,X_{n-r+1}) \quad (14)$$

then

$$AT^{(s,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(T_{x_{n+\Lambda},x_n};T_{x_n,y_n},T_{y_n,y_{n-1}},\dots,T_{y_{n-s+2},y_{n-s+1}}) \quad (15)$$

and ($r\geq 2$)

$$AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(T_{x_{n+\Lambda},x_n};T_{x_{n-r+1},y_n},T_{y_n,y_{n-1}},\dots, T_{y_{n-s+2},y_{n-s+1}}|T_{x_n,x_{n-1}},\dots,T_{x_{n-r+2},x_{n-r+1}}). \quad (16)$$

Note in Theorem 1 that $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ has $s+r+1$ variables (of which $r\geq 1$ are conditioning variables, see (9)), while

$$I(T_{x_{n+\Lambda},x_n};T_{x_n,y_n},T_{y_n,y_{n-1}},\dots,T_{y_{n-s+2},y_{n-s+1}})$$

(rhs of (15)) and

$$I(T_{x_{n+\Lambda},x_n};T_{x_{n-r+1},y_n},T_{y_n,y_{n-1}},\dots,T_{y_{n-s+2},y_{n-s+1}}|T_{x_n,x_{n-1}},\dots, T_{x_{n-r+2},x_{n-r+1}}),$$

(rhs of (16) have $s+r$ variables (of which $r-1$ are conditioning in the second case, and none in the first). For convenience, we will refer to the rhs of (15) and (16) as the *transcript dimensional reduction* of $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$, even when the constraint (14) is not fulfilled.

Remark 1.

(i) A similar result for the algebraic multi-information function[29]

$$I(X_1,\dots,X_N) = \sum_{n=1}^{N} H(X_n) - H(X_1,\dots,X_N)$$

can be found in Theorem 1 of Ref. 17. Note that $I(X_1, X_2)$ is mutual information; hence, it leads to a transfer entropy after proper conditioning. The comparison of both theorems shows a common feature and a crucial difference. The common feature is a constraint involving the coupling complexity coefficients. The constraint for the multi-information function is however different from (14), except for $r=s=1$. The crucial difference is that Theorem 1 of Ref. 17 requires as well a further constraint involving the entropies of the processes considered. No such a constraint is needed in the case of an ATE.

(ii) If

$$C(X_{n+\Lambda},X_n,X_{n-1},\dots,X_{n-r+1},Y_n,Y_{n-1},\dots,Y_{n-s+1}) = 0 \quad (17)$$

and

$$C(X_{n+\Lambda},X_n,X_{n-1},\dots,X_{n-r+1}) = 0 \quad (18)$$

then,

$$C(X_{n+\Lambda}|X_n,X_{n-1},\dots,X_{n-r+1},Y_n,Y_{n-1},\dots,Y_{n-s+1}) = C(X_{n+\Lambda}|X_n,X_{n-1},\dots,X_{n-r+1}) = 0$$

by the monotonicity property of the coupling complexity coefficients, Lemma 1. This being the case, the condition (14) holds then trivially.

(iii) Owing to the *equivalence property* mentioned in Sec. II, Eqs. (15) and (16) can be written in different, equivalent forms.

The computation of the transfer entropy supposes the estimation of the probabilities in (8) or (9), usually by their relative frequencies. Due to the difficulty of estimating reliably those probabilities in high dimensions, researchers mostly use the lowest dimensional case, namely:

$$AT^{(1,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(X_{n+\Lambda};Y_n|X_n). \quad (19)$$

For notational simplicity, $AT^{(1,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) =: AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ hereafter. Setting $s=r=1$ in Theorem 1, and noting that $C(X_{n+\Lambda}|X_n,Y_n) = C(X_{n+\Lambda},X_n,Y_n) - C(X_n,Y_n)$, and $C(X_{n+\Lambda}|X_n) = C(X_{n+\Lambda},X_n)$ (since $C(X_n)=0$ by (5)), we obtain the following result.

Corollary 1. If

$$C(X_{n+\Lambda},X_n,Y_n) = C(X_{n+\Lambda},X_n) + C(X_n,Y_n) \quad (20)$$

then

$$AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(T_{x_{n+\Lambda},x_n};T_{x_n,y_n}). \quad (21)$$

Corollary 1 spells out that if the condition on the coupling complexity coefficients (20) holds, then the algebraic transfer entropy $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ can be calculated as a mutual information of certain transcripts. Let us set

$$TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) := I(T_{x_{n+\Lambda},x_n};T_{x_n,y_n}) \quad (22)$$

for further reference. Note that the mutual information $I(T_{x_{n+\Lambda},x_n};T_{x_n,y_n})$ is not symmetric under the swap of $\mathbf{X}$ and $\mathbf{Y}$, then $TI_{\mathbf{X}\to\mathbf{Y}}(\Lambda) := I(T_{y_{n+\Lambda},y_n};T_{y_n,x_n}) = I(T_{y_{n+\Lambda},y_n};T_{x_n,y_n})$ (in virtue of the equivalence property, Sec. II B); hence, $TI_{\mathbf{X}\to\mathbf{Y}}(\Lambda) \neq TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ in general.

By setting $r=s=1$ in Remark 1(ii), we obtain the following special version of Corollary 1.

Corollary 2. If

$$C(X_{n+\Lambda},X_n,Y_n) = C(X_{n+\Lambda},X_n) = 0, \quad (23)$$

then $I(X_{n+\Lambda};Y_n|X_n) = I(T_{x_{n+\Lambda},x_n};T_{x_n,y_n})$, i.e.,

$$AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda).$$

From Corollaries 1 and 2, we conclude that, under conditions (20) or (23) on the coupling complexity coefficients, the algebraic transfer entropy $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ can be calculated

by means of mutual information of transcripts. In some cases, the condition (20) can be satisfied by selecting the time delay appropriately. More frequently, the conditions (23) can be met approximately by choosing the time delay sufficiently large. These practical issues will be considered in Sec. V.

IV. DIRECTIONALITY INDICATORS

In view of the definition (10) of $\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ and Lemma 2, set

$$\begin{aligned}\Delta TI^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(&T_{x_{n+\Lambda},x_n}; T_{x_{n-r+1},y_n}, T_{y_n,y_{n-1}}, \ldots,\\ &T_{y_{n-s+2},y_{n-s+1}} | T_{x_n,x_{n-1}}, \ldots, T_{x_{n-r+2},x_{n-r+1}})\\ -I(&T_{y_{n+\Lambda},y_n}; T_{y_{n-s+1},x_n}, T_{x_n,x_{n-1}}, \ldots,\\ &T_{x_{n-r+2},x_{n-r+1}} | T_{y_n,y_{n-1}}, \ldots, T_{y_{n-s+2},y_{n-s+1}})\end{aligned}$$

if $r, s \geq 2$, and

$$\begin{aligned}\Delta TI^{(s,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = &I(T_{x_{n+\Lambda},x_n}; T_{x_n,y_n}, T_{y_n,y_{n-1}}, \ldots, T_{y_{n-s+2},y_{n-s+1}})\\ &-I(T_{y_{n+\Lambda},y_n}; T_{y_{n-s+1},x_n} | T_{y_n,y_{n-1}}, \ldots, T_{y_{n-s+2},y_{n-s+1}}),\end{aligned} \tag{24}$$

$$\begin{aligned}\Delta TI^{(1,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = &I(T_{x_{n+\Lambda},x_n}; T_{x_{n-r+1},y_n} | T_{x_n,x_{n-1}}, \ldots, T_{x_{n-r+2},x_{n-r+1}})\\ &-I(T_{y_{n+\Lambda},y_n}; T_{y_n,x_n}, T_{x_n,x_{n-1}}, \ldots, T_{x_{n-r+2},x_{n-r+1}}).\end{aligned} \tag{25}$$

Lemma 3. The following identity holds:

$$\begin{aligned}\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = &\Delta TI^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)\\ &+ C(Y_n, \ldots, Y_{n-s+1} | X_{n+\Lambda}, X_n, \ldots, X_{n-r+1})\\ &- C(Y_n, \ldots, Y_{n-s+1})\\ &- C(X_n, \ldots, X_{n-r+1} | Y_{n+\Lambda}, Y_n, \ldots, Y_{n-s+1})\\ &+ C(X_n, \ldots, X_{n-r+1}).\end{aligned} \tag{26}$$

Proof. Consider the case $r, s \geq 2$. From Lemma 2, Eq. (12), it follows

$$\begin{aligned}\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = &\Delta TI^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)\\ &+ C(X_{n+\Lambda} | X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1})\\ &- C(X_{n+\Lambda} | X_n, \ldots, X_{n-r+1})\\ &- C(Y_{n+\Lambda} | X_n, \ldots, X_{n-r+1}, Y_n, \ldots, Y_{n-s+1})\\ &+ C(Y_{n+\Lambda} | Y_n, \ldots, Y_{n-s+1})\end{aligned}$$

Replace now the definition of the different coupling complexity coefficients, see Eq. (11), to derive the simplified expression (26).

The other cases are checked similarly. □

Lemma 3 yields readily sufficient conditions for the computation of $\Delta AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ with one variable less. Rather than delving into the general case, we are going to focus, as in Sec. III, on the lowest dimensional case because of the applications. Thus, set $s = r = 1$ in (26) and remember that $C(X_n) = C(Y_n) := 0$ to obtain

$$\begin{aligned}\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = &\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) + C(Y_n | X_{n+\Lambda}, X_n)\\ &- C(X_n | Y_{n+\Lambda}, Y_n),\end{aligned} \tag{27}$$

where

$$\begin{aligned}\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) &:= \Delta AT^{(1,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)\\ &= I(X_{n+\Lambda}; Y_n | X_n) - I(Y_{n+\Lambda}; X_n | Y_n),\end{aligned} \tag{28}$$

(see (10) and (19)), and

$$\begin{aligned}\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) &:= \Delta TI^{(1,1)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)\\ &= I(T_{x_{n+\Lambda},x_n}; T_{x_n,y_n}) - I(T_{y_{n+\Lambda},y_n}; T_{x_n,y_n}),\end{aligned} \tag{29}$$

(see (24) or (25)).

Eq. (27) gives sufficient conditions for $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ to coincide in magnitude and sign with the directionality indicator $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$. More often than not, however, the data analyst is only interested in the coupling direction. This being the case, we derive in the following also sufficient conditions for $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ to have the same sign as $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$.

Theorem 2. (a) If

$$C(Y_n | X_{n+\Lambda}, X_n) \geq C(X_n | Y_{n+\Lambda}, Y_n), \tag{30}$$

then

$$\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) > 0 \Rightarrow \Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) > 0.$$

(b) If, otherwise,

$$C(Y_n | X_{n+\Lambda}, X_n) \leq C(X_n | Y_{n+\Lambda}, Y_n), \tag{31}$$

then

$$\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) < 0 \Rightarrow \Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) < 0.$$

(c) Finally, if

$$C(Y_n | X_{n+\Lambda}, X_n) = C(X_n | Y_{n+\Lambda}, Y_n), \tag{32}$$

then $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = \Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$.

The following theorem formulates similar results with the help of entropies:

Theorem 3. (a) If $H(X_n) \leq H(Y_n)$ and

$$C(Y_{n+\Lambda}, X_n, Y_n) \leq C(Y_{n+\Lambda}, Y_n), \tag{33}$$

then

$$\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) > 0 \Rightarrow \Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) > 0.$$

(b) If, otherwise, $H(Y_n) \leq H(X_n)$ and

$$C(X_{n+\Lambda}, Y_n, X_n) \leq C(X_{n+\Lambda}, X_n), \tag{34}$$

then

$$\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) < 0 \Rightarrow \Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) < 0.$$

(c) If $H(X_n) = H(Y_n)$ and

$$C(X_{n+\Lambda}, Y_n, X_n) = C(Y_{n+\Lambda}, X_n, Y_n) = 0, \tag{35}$$

then,

$$\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = \Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda).$$

Proof. (a) If $H(X_n) \leq H(Y_n)$, then

$$C(X_{n+\Lambda}, Y_n, X_n) \geq C(X_{n+\Lambda}, X_n)$$

by the monotonicity property of the coupling complexity coefficients, Lemma 1 (note again that $H(X_{n+\Lambda}) = H(X_n)$). Assumption (33) entails then that the condition (30) of Theorem 2(a) is fulfilled.

(b) Follows by interchanging X_n and Y_n in (a)

(c) By the monotonicity property of the coupling complexity coefficients (see Lemma 1), the assumption $H(X_n) = H(Y_n)$ implies both

$$C(X_{n+\Lambda}, Y_n, X_n) \geq C(X_{n+\Lambda}, X_n) \quad \text{and}$$
$$C(Y_{n+\Lambda}, X_n, Y_n) \geq C(Y_{n+\Lambda}, Y_n).$$

The second assumption $C(X_{n+\Lambda},Y_n,X_n)=C(Y_{n+\Lambda},X_n,Y_n) =0$ implies then

$$C(X_{n+\Lambda}, Y_n, X_n) = C(X_{n+\Lambda}, X_n) = 0 \quad \text{and}$$
$$C(Y_{n+\Lambda}, X_n, Y_n) = C(Y_{n+\Lambda}, Y_n) = 0,$$

respectively, because the coupling complexity coefficients are nonnegative. Apply now Theorem 2(c). □

According to Lemma 1, the condition

$$C(Y_{n+\Lambda}, X_n, Y_n) < C(Y_{n+\Lambda}, Y_n)$$

in (33) is only possible if $H(X_n) < H(Y_n)$. Likewise, the condition

$$C(X_{n+\Lambda}, Y_n, X_n) < C(X_{n+\Lambda}, X_n)$$

in (34) is only possible if $H(Y_n) < H(X_n)$. So, the conditions on the entropies and coupling complexity coefficients in Theorems 3(a) and (b) are consistent.

V. PRACTICAL CONSIDERATIONS

We consider next some practical issues concerning the theoretical results in Sections III and IV. For this reason, we concentrate on the lowest, 3-dimensional transfer entropy, $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(X_{n+\Lambda}; Y_n|X_n)$, and its transcript 2-dimensional reduction $TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = I(T_{x_{n+\Lambda},x_n}; T_{x_n,y_n})$.

To begin with, the possibility of computing or estimating the ATE with a lower dimensional quantity is a welcome property because it can avoid undersampling and, in any case, it improves the quality of the statistical estimates which go into the computation of the transfer entropy. For example, if one uses two decimal digits for the observations, the computation of $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ needs to estimate the $O(10^6)$ 3-variate probabilities $p(x_{n+\Lambda}, x_n, y_n)$ (see (8)), whereas $TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ requires the estimation of the $O(10^4)$ 2-variate probabilities $p(t_{x_{n+\Lambda},x_n}, t_{x_n,y_n})$.

The practical advantage of using the transcript dimensional reduction of an ATE becomes more evident if the data belong to a small group. This is the case of symbolic representations, in particular, with ordinal patterns of small or moderate lengths. Thus, if $\mathcal{G} = \mathcal{S}_4$, the computation of $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ requires the estimation of, in general, $(4!)^3 = 13824$ 3-variate probabilities, while for $TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ only $(4!)^2 = 576$ 2-variate probability estimations will do. Actually, these numbers are, in general, upper bounds because a deterministic dynamics, whether one-dimensional or higher dimensional, has always forbidden ordinal patterns for sufficiently long pattern lengths.[30] Numerical simulations show that the same thing happens with scalar observations (e.g., projections from an attractor[30]), even if non-uniformly sampled.[31] Ordinal patterns are becoming increasingly popular for discretizing $\mathbb{R}$-valued signals.

According to Corollaries 1 and 2, the equality $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ holds under the general condition (20) and, of course, under any particularization thereof such as (23). The practical implementation of these conditions in the case of ordinal representations was investigated in Refs. 16 and 17. The results can be summarized as follows.

- The coupling complexity coefficients (along with the entropies involved in their definition) depend, in general, on the time delay T used. It turns out that sometimes T can be chosen so that the conditions (20) or (23) are satisfied. But both possibilities seem to depend on the nature of the data being analyzed.
- First, it is generally observed that the 2- and 3-variate coupling complexity coefficients decrease monotonically with increasing T. For random and (noiseless) chaotic signals, their limits lie below the 10% of the maximal amplitude of the signal. Thus, in this case, one can take T sufficiently large so that the complexity coefficients in (23) may be considered "small" for practical purposes. For real-world data, though, the 3-variate limit remains above the 10% threshold. See Refs. 17 (Fig. 1) and 16 (Fig. 1).
- The same difference happens also with respect to the general condition (20). Indeed, one can often find small T's for which (20) holds in case of random and chaotic signals, while this is generally not possible for real-world data. See Ref. 17 (Fig. 1).

VI. DIRECTIONALITY WITH TRANSCRIPT MUTUAL INFORMATION

As mentioned above, the conditions for the equality $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda) = \Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ are difficult to meet when dealing with real-world data. This empirical observation raises the question of how good $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ (actually, its sign) performs as a directionality indicator regardless of those conditions. This question was first tackled in the study[32] of the spontaneous information flow within the visual corticothalamocortical circuit (comprising the six layers of the visual cortex and the lateral geniculate nucleus of the thalamus). $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ has been used as well, together with other two directionality indicators based on time-delay coordinates, to analyze climatological and EEG data.[33] In the first case, $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ performed as $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ both with the actual data and ordinal patterns of length 4. In the second case, the performance of $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ was comparable to one of the other two indicators. In the following two sections we pursue this intriguing issue with both synthetic (hence, noiseless)

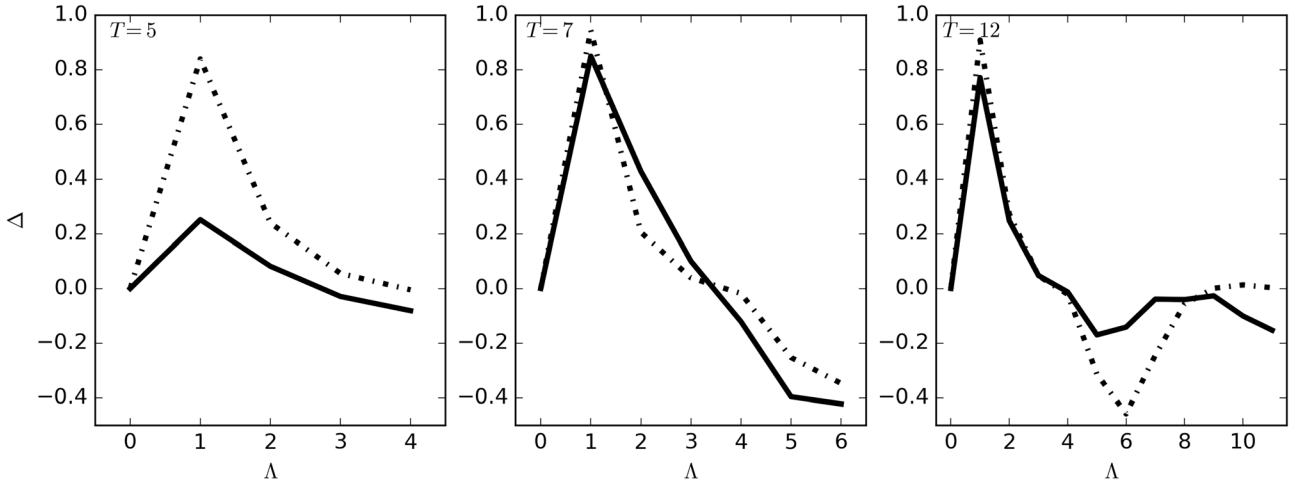

FIG. 1. Plots of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (solid line) and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (dashed line) in bits vs Λ for $N=2^{15}$ and $T=5$ (left), $T=7$ (center), $T=12$ (right), with $0\le\Lambda\le T-1$ ($\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(0)=\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(0)=0$ plotted for a better visualization).

data and real world data using ordinal patterns of lengths 4 and 3, respectively. All numerical results are given in bits.

A. Numerical simulations

To compare the performances of $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ and $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ as directionality indicators, we consider two bidirectionally delayed-coupled logistic maps $f:[0,1]\to[0,1]$, $f(t)=4t(1-t)$ defined by the equations[34]

$$\begin{aligned} x(t) &= f(g_{yx}(t) \bmod 1),\\ y(t) &= f(g_{xy}(t) \bmod 1), \end{aligned} \tag{36}$$

where

$$\begin{aligned} g_{yx}(t) &= 0.5y(t-1)+0.5x(t-1),\\ g_{xy}(t) &= 0.5x(t-6)+0.5y(t-1). \end{aligned}$$

For the numerical simulation, N points $x_n := x(n)$ and $y_n := y(n)$ of each time series were generated and a symbolic representation of the time series with ordinal patterns of length $L=4$ was used. To be more specific, let $\hat{x}_n=(\xi_0,\xi_1,\xi_2,\xi_3)\in\mathcal{S}_4$ be the ordinal 4-pattern of the time delay vector $\mathbf{v}(x_n)=(x_n,x_{n+T},x_{n+2T},x_{n+3T})$ (i.e., $x_{n+\xi_0 T}<x_{n+\xi_1 T}<x_{n+\xi_2 T}<x_{n+\xi_4 T}$, with $\xi_i\in\{0,1,2,3\}$), where $1\le n\le N-3T$ and $T\ge 1$, and similarly with (y_n). Then, the data analyzed were the $\mathcal{S}_4$-valued time series $(\hat{x}_n)$, and $(\hat{y}_n)$.

In view of the number of probabilities needed to estimate $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (see Sec. V), we consider next time series of lengths $N=2^{15}$ and 2^{11}. In the first case, the number of data is sufficient for an unbiased estimate of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (this can be checked by taking N larger). On the contrary, $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ is undersampled if $N=2^{11}$, so we expect a better performance of $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ in this case. To avoid direct information leaks due to overlaps of the ordinal 4-patterns, we chose $1\le\Lambda\le T-1$. This means that $T\ge 2$.

Fig. 1 shows $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (solid line) and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (dashed line) vs Λ for $N=2^{15}$ and time delays $T=5,7,12$; the values have been linearly interpolated for a better visualization. Here, $\hat{\mathbf{X}}$ (resp. $\hat{\mathbf{Y}}$) is the $\mathcal{S}_4$-valued random process which outputs $(\hat{x}_n)$ (resp. $(\hat{y}_n)$). Needless to say, since the plots of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ are different, the conditions on coupling complexity coefficients stated in Theorems 2(c) or 3(c) cannot hold in this case. Nevertheless, according to Fig. 1, $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ performs as good as, if not better than, $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ as far as the direction of the net information is concerned. In particular, both indicators signalize a change of direction around $\Lambda=4$, and the coupling delays of the system (36) at $\Lambda=1$ (for $T=5,7,12$) and $\Lambda=6$ (for $T=12$). However, $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ has sharp peaks both at $\Lambda=1$ and 6, while $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ displays an inverted hump around $\Lambda=6$, taking actually a misleading, more negative value at $\Lambda=5$. Similar performances are also observed with other choices of the coupling constants and delays.

As N decreases, the performance of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ deteriorates but much slower than one would expect based on the hypothetical number of 3-variate probabilities to be estimated. The reason for this is the existence of forbidden patterns, which lowers that count in noiseless, nonlinear time series analysis. For a short series of length $N=2^{11}$ (Fig. 2), $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ still marks (with a smaller amplitude) the coupling delay $\Lambda=1$ but misses $\Lambda=6$. Indeed, the curve $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ is much flatter now for $\Lambda\ge 3$ than in Fig. 1, which makes difficult the location of, e.g., direction changes even in this noiseless simulation. On the contrary, the curves $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ on the four panels of Fig. 2 look the same as in Fig. 1, which confirms that the probability estimates for $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ are not undersampled for $N=2^{11}$.

To complete the picture, we have calculated $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ for $\Lambda=1,6$ (the two coupling delays of the system (36)), $T=12$, and $N=2^k$, $11\le k\le 17$. Fig. 3 (lower panel) shows that $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(6)$ performs poorly over the whole range, being only marginally negative for $k\ge 12$ and negligible for $k=11$; furthermore, $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(5)\le\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(6)$ for $13\le k\le 17$ (not shown). As for $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)$ (upper panel), its performance improves with N. At variance with these results, $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ has always a positive sharp peak at $\Lambda=1$, and a negative sharp peak at $\Lambda=6$, with $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)>\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)>0$ and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(6)<$

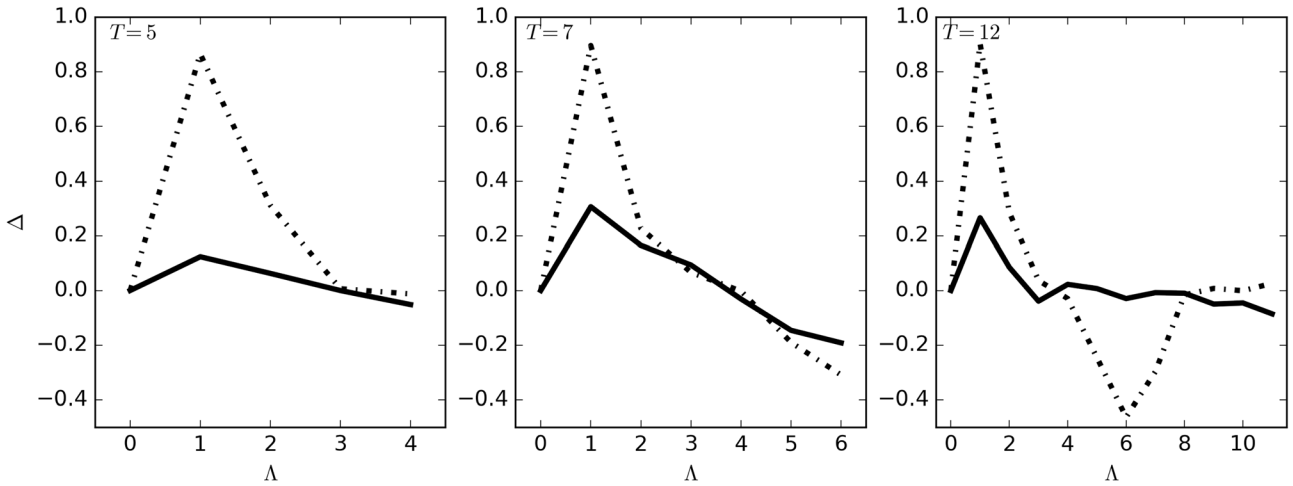

FIG. 2. Plots of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (solid line) and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (dashed line) in bits vs Λ for $N=2^{11}$ and $T=5$ (left), $T=7$ (center), $T=12$ (right), with $0\le\Lambda\le T-1$ ($\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(0)=\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(0)=0$ plotted for a better visualization).

$\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)\le 0$ over the whole range of k. Furthermore, both $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)$ and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(6)$ exhibit a stable, almost constant behavior with N. In conclusion, $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ proves once again to be a reliable directionality indicator.

B. Cardiovascular data

The ordinal methodology in general biomedical applications can be found in Ref. 35. For specific applications to cardiology, the interested reader is referred to Refs. 36–38.

Oxygen is widely used in patients as well as in aviation or during diving activities. However, the previous studies have shown that in some cases the impact of oxygen might not be beneficial but harmful. The mechanism of cardiovascular changes during oxygen breathing (OXB) is not fully understood. Therefore, a clinical study was designed to determine whether the relation between heart rate and blood pressure differs between medical air and 100% oxygen breathing. Since the time series recorded in this study were rather short, they were an ideal test bed for comparing the performances of $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$, Eq. (28), and its transcript dimensional reduction $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$, Eq. (29), with real world data. No medical implications from the analysis below are claimed.

Twelve healthy young volunteers (5 men; age within the interval 33.8 ± 7.4 years) participated in the study. All patients underwent short-term electrocardiographic recording (ECG) using PowerLab system with Lab Chart software (ADInstruments, Australia). The sampling rate was 1000 Hz. Non-invasive beat-to-beat blood pressure (BP) was recorded by a FINOMETER device (Finapres Medical Systems). ECG fragments of 512 inter-beat (RR) intervals were used for

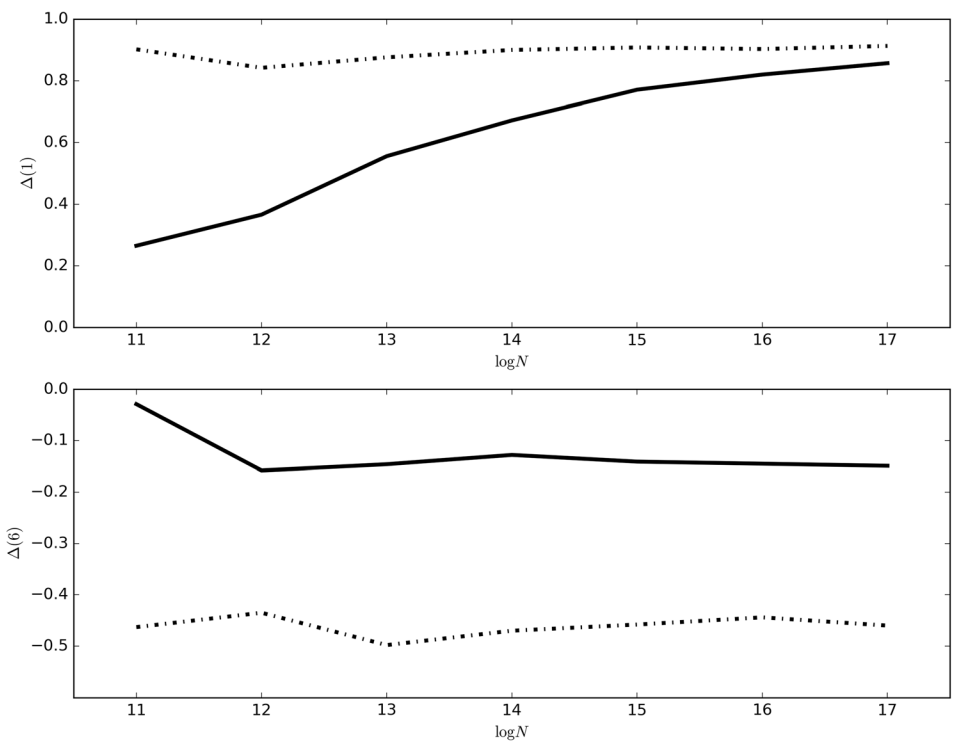

FIG. 3. Top panel: Plots of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)$ (solid line) and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(1)$ (dashed line) in bits for $N=2^k$, $11\le k\le 17$. Bottom panel: Plots of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(6)$ (solid line) and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(6)$ (dashed line) in bits for $N=2^k$, $11\le k\le 17$. In both panels, $T=12$.

further analyses. All measurements took place twice: first during *medical air breathing* (MAB) and then during *100% oxygen breathing* (OXB).

Consider first the RR data in either case MAB or OXB. The raw data consisted then of 12 time series of the form

$$(\widetilde{y}_{t_n})_{n=1}^{512} = (\widetilde{RR}_{t_n})_{n=1}^{512}, \tag{37}$$

where t_n are the times in seconds when the heart beats, and $\widetilde{RR}_{t_n} = 1000(t_{n+1} - t_n)$, i.e., the consecutive inter-beat time intervals in milliseconds. For notational simplicity, we do not label each of the 12 time series nor the case (MAB, OXB). Since the times t_n were not evenly spaced, new, uniformly spaced time series

$$(y_n)_{n=1}^{512} = (RR_n)_{n=1}^{512} \tag{38}$$

were constructed by cubic spline interpolation, a rather common technique that is adequate for our illustrative purposes. Specifically, for each time series, the RR_n values were calculated by spline interpolating the measured $\widetilde{RR}_{t_n}$ values at times $t = 0.93n$ s, where 0.93 is the average of $t_{n+1} - t_n$, $1 \le n \le 512$ over the sample and the two cases MAB and OXB.

The same procedure was applied to the blood pressure data

$$(\widetilde{x}_{t_n})_{n=1}^{512} = (\widetilde{BP}_{t_n})_{n=1}^{512}, \tag{39}$$

where $\widetilde{BP}_{t_n}$ is blood pressure measured in mmHg. Again, all 12 time series per case were adjusted to a uniformly spaced time series

$$(x_n)_{n=1}^{512} = (BP_n)_{n=1}^{512}, \tag{40}$$

via spline interpolation of the blood pressures $\widetilde{BP}_{t_n}$ at times $t = 0.93n$ s, $1 \le n \le 512$. The data (37) and (39) are typical examples of short biomedical time series.

Next, we represented the $12 + 12$ time series $(x_n)_{n=1}^{512}$ and $(y_n)_{n=1}^{512}$ by the ordinal patterns of sliding time delay vectors of length $L = 3$ and time delay $T \ge 1$. That is, if $\mathbf{v}(x_n) = (x_n, x_{n+T}, x_{n+2T})$ and $\mathbf{v}(y_n) = (y_n, y_{n+T}, y_{n+2T})$ are such vectors, $1 \le n \le 512 - 2T$, and $\hat{x}_n = \widehat{BP}_n \in \mathcal{S}_3$, $\hat{y}_n = \widehat{RR}_n \in \mathcal{S}_3$ are the ordinal patterns of $\mathbf{v}(x_n)$ and $\mathbf{v}(y_n)$, respectively, then the symbolic time series to be analyzed were

$$(\hat{x}_n)_{n=1}^{512-2T} = (\widehat{BP}_n)_{n=1}^{512-2T}, \quad (\hat{y}_n)_{n=1}^{512-2T} = (\widehat{RR}_n)_{n=1}^{512-2T}. \tag{41}$$

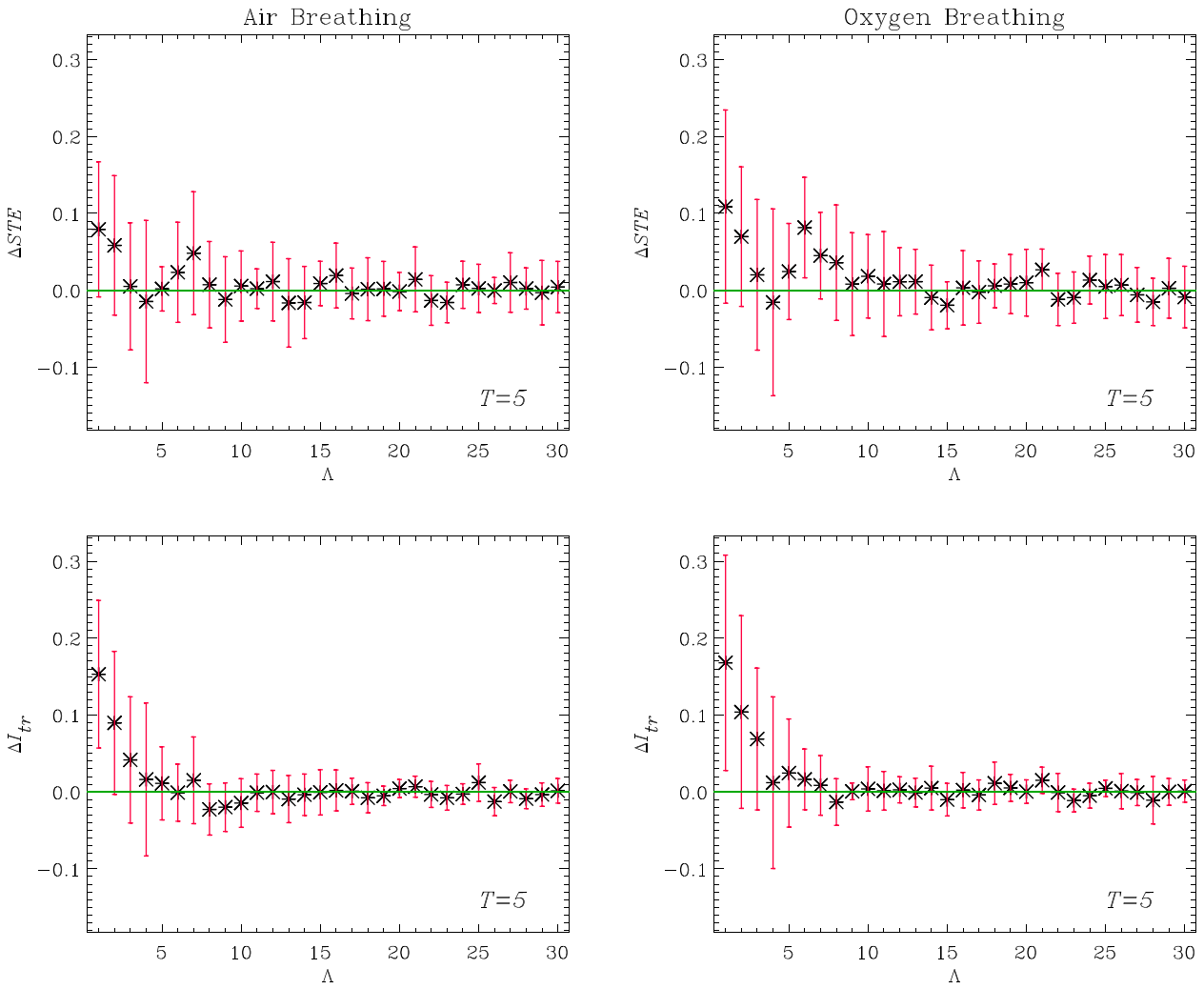

FIG. 4. $\hat{\mathbf{Y}} = (\hat{Y}_n)$ are the RR data and $\hat{\mathbf{X}} = (\hat{X}_n)$ the blood pressure data represented by ordinal 3-patterns, $1 \le n \le 450$. The top row shows $\Delta AT_{\hat{\mathbf{Y}} \to \hat{\mathbf{X}}}(\Lambda)$ in bits vs Λ in the MAB case (left column) and OXB case (right column) for $1 \le \Lambda \le 30$ ($T = 5$). The bottom row shows the corresponding plots for $\Delta TI_{\hat{\mathbf{Y}} \to \hat{\mathbf{X}}}(\Lambda)$.

If $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda) = \Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ would hold, then one could exploit the lower dimensionality of $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ at this point. For this, one needs to check first whether some of the conditions stated in Theorem 2 (a)–(c) or in Theorem 3 (a)–(c) are fulfilled. Since, for instance, $H(\hat{X}_n) = 2.56\pm0.03$ bits and $H(\hat{Y}_n) = 2.58\pm0.01$ bits for the MAB data, one might check the condition (35) of Theorem 3(c), but neither $C(\hat{X}_{n+\Lambda}, \hat{X}_n, \hat{Y}_n)$ nor $C(\hat{Y}_{n+1}, \hat{X}_n, \hat{Y}_n)$ may be considered small with respect to the entropies for $1 \le T \le 30$ (not shown). Therefore, $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda) \neq \Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ for all the time delays considered. Actually, this is the situation we are interested in here.

This being the case, we evaluated $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ and $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ with time delay $T = 5$ and $1 \le \Lambda \le 30$. The count of probability estimations is $(3!)^3 = 216$ for $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ (clearly undersampled in view of the $512 - 2T = 502$ points available), and only $(3!)^2 = 36$ for $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$. The results are represented in Fig. 4, where the error bars correspond to standard deviations over the 12 subjects. Fig. 4, top row, shows $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ vs Λ for the MAB data (left column) and OXB data (right column). Fig. 4, bottom row, shows the corresponding plots for $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$. In both cases, $\Delta TI_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ detects for $\Lambda = 1$ that the process $\hat{\mathbf{Y}}$ is leading the process $\hat{\mathbf{X}}$, which is in concordance with the previous findings. On the contrary, the error bars of $\Delta AT_{\hat{\mathbf{Y}}\to\hat{\mathbf{X}}}(\Lambda)$ do not allow any conclusion for $1 \le \Lambda \le 30$ (not even for $\Lambda = 5$, 10, in which case $\hat{y}_n$ overlaps with $\hat{x}_{n+\Lambda}$, and $\hat{x}_n$ with $\hat{y}_{n+\Lambda}$). Let us finally mention that $\Lambda = 1$ is the usual choice in applications.

VII. CONCLUSION

This paper continues the study of algebraic transfer entropy initiated in Ref. 16 for ordinal representations and generalized in Ref. 17 to algebraic symbolic representations and conditional multi-information. This time we showed (Theorem 1) that, subject to the restriction (14), the algebraic transfer entropy $AT^{(s,r)}_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ can be computed via a mutual information of transcripts (if $r = 1$), or a conditional mutual information of transcripts (if $r > 1$), with one variable less. More practical formulations were derived in Corollaries 1 and 2, as well as in Theorems 2 and 3. As a by-product, we generalized a previous result concerning the algebraic transfer entropy of dimension 3 (see Remark 1(i)).

The role of the restriction (14) in applications was discussed in Sec. V in the particular case of the 3-dimensional ATE, $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$, and the corresponding directionality indicator $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$. Our experience[32,33] shows that the restrictions for $AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ and $\Delta AT_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ can be satisfied in general (at least approximately) for random and chaotic data by choosing appropriate time delays. Therefore, in such cases, those quantities may be replaced by the transcript mutual information $TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ and the corresponding directionality indicator $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$, respectively. Intriguingly, our experience supports the suggestion that $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ is also a good directionality indicator even when the corresponding restriction is not fulfilled, such as it usually happens when analyzing real world data. This question was tackled in Secs. VI A (numerical simulations) and VI B (cardiovascular data) with satisfactory results. The performance of $\Delta TI_{\mathbf{Y}\to\mathbf{X}}(\Lambda)$ as a directionality indicator, regardless of any condition on the coupling complexity coefficients, is a subject that deserves further study.

ACKNOWLEDGMENTS

We thank our reviewers for their constructive criticism. J.M.A. was supported by the Spanish Ministry of Economy and Competitiveness, Grant No. MTM2012-31698. B.G. was supported by the Foundation for Polish Science within the SKILLS project co-financed with European Union funds, from the European Social Fund (No. 172/UD/SKILLS/2012). G.G. was supported by the National Science Centre, Poland, Grant No. UMO-2014/15/B/ST1/01710.

[1]N. Wiener, in *Modern Mathematics for Engineers*, edited by E. F. Beckenbach (McGraw-Hill, New York, 1956).
[2]C. W. J. Granger, "Investigating causal relations by econometric and cross-spectral methods," Econometrica **37**, 424–438 (1969).
[3]C. W. J. Granger, "Testing for causality: A personal viewpoint," J. Econ. Dyn. Control **2**, 329–352 (1980).
[4]C. W. J. Granger, "Some recent developments in a concept of causality," J. Econometrics **39**, 199–211 (1988).
[5]S. J. Schiff, P. So, T. Chang, R. E. Burke, and T. Sauer, "Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble," Phys. Rev. E **54**, 6708–6724 (1996).
[6]Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing multiple nonlinear time series with extended Granger causality," Phys. Lett. A **324**, 26–35 (2004).
[7]M. C. Romano, M. Thiel, J. Kurths, and C. Grebogi, "Estimation of the direction of the coupling by conditional probabilities of recurrence," Phys. Rev. E **76**, 036211 (2007).
[8]N. F. Rulkov, M. M. Sushchik, L. S. Tsimring, and H. D. I. Abarbanel, "Generalized synchronization of chaos in directionally coupled chaotic systems," Phys. Rev. E **51**, 980–994 (1995).
[9]Y. Hirata and K. Aihara, "Identifying hidden common causes from bivariate time series: A method using recurrence plots," Phys. Rev. E **81**, 016203 (2010).
[10]G. Sugihara, R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," Science **338**, 496–500 (2012).
[11]T. Schreiber, "Measuring information transfer," Phys. Rev. Lett. **85**, 461–464 (2000).
[12]K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," Phys. Rep. **441**, 1–46 (2007).
[13]R. Monetti, W. Bunk, T. Aschenbrenner, and F. Jamitzky, "Characterizing synchronization in time series using information measures extracted from symbolic representations," Phys. Rev. E **79**, 046207 (2009).
[14]J. M. Amigó, R. Monetti, T. Aschenbrenner, and W. Bunk, "Transcripts: An algebraic approach to coupled time series," Chaos **22**, 013105 (2012).
[15]R. Monetti, J. M. Amigó, T. Aschenbrenner, and W. Bunk, "Permutation complexity of interacting dynamical systems," Eur. Phys. J. Spec. Top. **222**, 421–436 (2013).
[16]R. Monetti, W. Bunk, T. Aschenbrenner, S. Springer, and J. M. Amigó, "Information directionality in coupled time series using transcripts," Phys. Rev. E **88**, 022911 (2013).
[17]J. M. Amigó, T. Aschenbrenner, W. Bunk, and R. Monetti, "Dimensional reduction of conditional algebraic multi-information via transcripts," Inf. Sci. **278**, 298–310 (2014).
[18]S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," Phys. Rev. Lett. **99**, 204101 (2007).
[19]D. Smirnov, "Spurious causalities with transfer entropy," Phys. Rev. E **87**, 042917 (2013).
[20]J. Sun and E. M. Bollt, "Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings," Physica D **267**, 49–57 (2014).
[21]I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," Phys. Rev. E **82**, 016207 (2010).

[22]D. W. Hahs and S. D. Pethel, "Distinguishing anticipation from causality: Anticipatory bias in the estimation of information flow," Phys. Rev. Lett. **107**, 128701 (2011).
[23]J. Runge, J. Heitzig, N. Marwan, and J. Kurths, "Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy," Phys. Rev. E **86**, 061121 (2012).
[24]C. Cafaro, W. M. Lord, J. Sun, and E. M. Bollt, "Causation entropy from symbolic representations of dynamical systems," Chaos **25**, 043106 (2015).
[25]M. Paluš and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," Phys. Rev. E **75**, 056211 (2007).
[26]C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," Phys. Rev. Lett. **88**, 174102 (2002).
[27]J. M. Amigó, L. Kocarev, and J. Szczepanski, "Order patterns and chaos," Phys. Lett. A **355**, 27–31 (2006).
[28]J. M. Amigó, *Permutation Complexity in Dynamical Systems—Ordinal Patterns, Permutation Entropy, and All That* (Springer, Verlag, Berlin, 2010).
[29]M. Studený and J. Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in Graphical Models*, edited by M. I. Jordan (MIT Press, Cambridge, MA, 1999), pp. 261–296.
[30]J. M. Amigó and M. B. Kennel, "Forbidden ordinal patterns in higher dimensional dynamics," Physica D **237**, 2893–2899 (2008).
[31]C. W. Kulp, J. M. Chobot, B. J. Niskala, and C. J. Needhammer, "Using forbidden ordinal patterns to detect determism in irregularly sampled time series," Chaos **26**, 023107 (2016).
[32]J. M. Amigó, R. Monetti, N. Tort-Colet, and M. V. Sanchez-Vives, "Infragranular layers lead information flow during slow oscillations according to information directionality indicators," J. Comput. Neurosci. **39**, 53–62 (2015).
[33]Y. Hirata, J. M. Amigó, Y. Matsuzaka, R. Yokota, H. Mushiake, and K. Aihara, "Detecting causality by combined use of multiple methods: Climate and brain examples," PLoS ONE **11**, e0158572 (2016).
[34]B. Pompe and J. Runge, "Momentary information transfer as a coupling measure of time series," Phys. Rev. E **83**, 051122 (2011).
[35]J. M. Amigó, K. Keller, and V. Unakafova, "Ordinal symbolic analysis and its application to biomedical recordings," Philos. Trans. R. Soc. A **373**, 20140091 (2015).
[36]U. Parlitz, S. Berg, S. Luther, A. Schirdewan, J. Kurths, and N. Wessel, "Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics," Comput. Biol. Med. **42**, 319–327 (2012).
[37]G. Graff, B. Graff, A. Kaczkowska, D. Makowiec, J. M. Amigó, J. Piskorski, K. Narkiewicz, and P. Guzik, "Ordinal pattern statistics for the assessment of heart rate variability," Eur. Phys. J. Spec. Top. **222**, 525–534 (2013).
[38]B. Graff, G. Graff, D. Makowiec, A. Kaczkowska, D. Wejer, S. Budrejko, D. Kozłowski, and K. Narkiewicz, "Entropy measures in the assessment of heart rate variability in patients with cardiodepressive vasovagal syncope," Entropy **17**, 1007–1022 (2015).