PHYSICAL REVIEW E **87**, 042917 (2013)

Spurious causalities with transfer entropy

Dmitry A. Smirnov*
Saratov Branch of V. A. Kotel'nikov Institute of Radio Engineering and Electronics of the Russian Academy of Sciences, 38 Zelyonaya St., Saratov 410019, Russia

(Received 21 February 2013; published 17 April 2013)

Transfer entropy (TE) seems currently to be the most widely used tool to characterize causal influences in ensembles of complex systems from observed time series. In particular, in an elemental case of two systems, nonzero TEs in both directions are usually interpreted as a sign of a bidirectional coupling. However, one often overlooks that both positive TEs may well be encountered for unidirectionally coupled systems so that a false detection of a causal influence on the basis of a nonzero TE is rather possible. This work highlights typical factors leading to such "spurious couplings": (i) unobserved state variables of the driving system, (ii) low temporal resolution, and (iii) observation errors. All are shown to be particular cases of a general problem: *imperfect observations of states of the driving system*. Importantly, *exact values* of TEs, rather than their *statistical estimates*, are computed here for selected benchmark systems. Conditions for a "spurious" TE to be large and even strongly exceed a "correct" TE are presented and discussed.

DOI: 10.1103/PhysRevE.87.042917 PACS number(s): 05.45.Tp, 05.45.Xt, 02.50.Sk, 02.50.Tt

I. INTRODUCTION

In studies of an ensemble of systems with nontrivial temporal evolution, it is fruitful to characterize their interactions which often determine basic features of the collective behavior [1–6]. For a deeper understanding, it is particularly important to reveal "directional couplings" or "causal influences" [7–9], i.e., to answer the question "who drives whom." To mention just a couple of fields with multiple examples of such a problem setting, one looks for couplings between brain areas or different rhythms in electroencephalograms (see, e.g., Refs. [10–16]) and between large-scale modes of climate variability, such as El Niño Southern Oscillation and North Atlantic Oscillation, or global climate processes (see, e.g., Refs. [17–21]). The detection of directional couplings is much more reliable when one can manipulate with the systems under study by performing special interventions into them [7,10,22,23]. If this is impossible, as often the case in biomedical or geophysical research, one must reveal couplings from passive observations of the systems behavior, e.g., from a time series of certain observed variables. Sometimes solving such a problem appears feasible based on the celebrated concept of the Granger causality [24]. Its generalized, modern, actively used, and highly trusted version is an information-theoretic measure called "transfer entropy" (TE) [25]. However, as is shown in this work, TE may lead to spuriously detected causalities under simple and widespread practical conditions which remain underestimated in many studies.

In the case of two systems, one says that a system X "Granger causes" a system Y if knowledge of the past of X improves predictions of Y as compared to self-predictions. Prediction improvement (PI) is usually defined as a decrease in the mean-squared prediction error, which is easily estimated from data and often appears sufficient for practical purposes. Such a PI is not invariant under a nonlinear invertible change of variables, and, hence, in practice it may be sensitive to nonlinear distortions of a measurement device. Instead of mean-squared errors, TE characterizes an "uncertainty reduction" in terms of Shannon entropies of the conditional probability distributions of the future of Y. It agrees with some early ideas of C. W. J. Granger as well [26]. TE is invariant to any invertible change of variables and seems currently the most universal and widely used characteristic of causal influences. Several techniques are developed for TE estimation from a time series [27,28], including optimization of time lags [29,30], a special approach to soften "the curse of dimensionality" problem [21], a symbolic version of TE [14], an expanded version of TE for ensembles [16], and analytic assessment of an estimator confidence band [31]. Moreover, conditional mutual information, a concept similar to TE, has been introduced and studied in a series of parallel works [28,32–36]. Similarly to PI, a statistically significant nonzero value of TE in the direction from X to Y is usually interpreted as a result of the influence $X \rightarrow Y$, and nonzero TEs in both directions as a result of the bidirectional coupling (BC).

It was known for linear systems that spurious couplings can be inferred from nonzero PIs due to the following reasons: unobserved variables influencing both systems dynamics [26], low temporal resolution [37,38], and observational noise [39]. Yet such effects were not appreciated and systematically analyzed for TE, apart from a specific example of nonlinear maps of a special kind with low observation accuracy [40]. TE is often stated to reflect an "information flow" or "information transfer" in the respective direction by its very definition, so that many works concentrate mainly on its accurate estimation (e.g., Refs. [28,29]), and it seems to be a sufficiently widespread opinion that getting spurious causalities with TE is hardly possible as opposed to the mean-squared PI. This work demonstrates that TEs may well be nonzero in both directions even in the case of a unidirectional coupling (UC), so that "spurious couplings" are rather possible to be inferred from TE under the naive approach. Moreover, it appears possible for a "spurious" TE (in the direction of an absent influence) to be much greater than a "correct" TE (in the direction of an existing influence). The reasons are similar to those mentioned above for the linear problems and shown to be particular cases of a general circumstance, which is *imperfect observations of states of the driving system.*

*smirnovda@yandex.ru

Exact theoretical values of TEs rather than their *statistical estimates* are presented here, so that all the results do not depend on any method of TE estimation and are, therefore, highly reliable. This is achieved due to the special choice of benchmark systems for the analysis: linear autoregression processes and Markov chains. These paradigmatic systems may serve as basic mathematical models of a wide range of natural processes, including oscillatory ones, that ensure practical importance of the results and allow their vivid interpretation.

The paper is organized as follows. Section II presents a mathematical formulation of the problem under study and a definition of TEs. Section III describes the benchmark examples and formulas for the TE computation. Section IV contains numerical results demonstrating the effect of unobserved variables, low temporal resolution, and observational noise on TEs. Section V discusses the results and their practical consequences, including a possible test for BC. Conclusions are given in Sec. VI. Some cumbersome derivations concerning Markov chains are moved to appendices.

II. MATHEMATICAL FORMULATION OF THE PROBLEM

The problem under study is the following. (i) There are two systems X and Y. (ii) Simultaneous time series of the variables u and v represent the dynamics of X and Y, respectively. (iii) Any hidden system influencing both X and Y is absent. (iv) The task is to find out whether a coupling between X and Y is bidirectional or unidirectional, and TEs in both directions are computed for that. The question is whether the TE values are always sufficient to reveal the coupling character. If not always, when are they appropriate? Note that a capability to reveal directional couplings between two systems is a necessary step to address a more complicated question about couplings within larger sets of systems. Moreover, the distinction between UC and BC is often basically and practically important *per se*; see, e.g., Ref. [38] and references therein.

Let u and v be observed at discrete time instants: u_n, v_n, $n \in \mathbf{Z}$. Without loss of generality, scalar-valued quantities u and v are considered. The case of vector-valued observables can be treated in the same way and is briefly commented on in Sec. V. Speaking mathematically, u_n and v_n are strictly stationary random processes. Denote $\mathbf{u}_n^d = (u_{n-1}, u_{n-2}, \ldots, u_{n-d})^{\mathrm{T}}$ and $\mathbf{v}_n^d = (v_{n-1}, v_{n-2}, \ldots, v_{n-d})^{\mathrm{T}}$, where T stands for transposition. The dimension d may be called "depth of history" relatively to the current time instant n. Let $P(u_n|\mathbf{u}_n^d)$ be a probability distribution of u_n conditioned by a given history $\mathbf{u}_n^d$, and $P(u_n|\mathbf{u}_n^d, \mathbf{v}_n^d)$ a distribution conditioned by histories of both processes. The difference between the Shannon entropies of the two distributions quantifies to what extent the uncertainty in the value of u_n decreases if a history of another process $\mathbf{v}_n^d$ is taken into account. Denote that difference $T_{Y\to X}^d$. In case of discrete-valued u and v, it reads

$$T_{Y\to X}^d = -\sum_{u_n,\mathbf{u}_n^d} P(u_n,\mathbf{u}_n^d)\log_2 P(u_n|\mathbf{u}_n^d) + \sum_{u_n,\mathbf{u}_n^d,\mathbf{v}_n^d} P(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)\log_2 P(u_n|\mathbf{u}_n^d,\mathbf{v}_n^d). \quad (1)$$

The base of the logarithms determines just the units of measurement. Binary logarithms mean that all the entropies are measured in bits. TE in the direction $Y \to X$ is defined as $T_{Y\to X} = \lim_{d\to\infty} T_{Y\to X}^d$, if the limit exists. "Practical" convergence over d is reached for many systems at quite moderate values of d, which is the case in all examples below. Therefore, TEs are computed according to Eq. (1) with d large enough for the convergence at a given small error to occur. Thus, the "exact" values of TE imply here "exact up to a given pre-defined error." However, this is quite a small error in comparison with any analysis based on statistical estimates of TE from a simulated time series due to the "curse of dimensionality" [21,28]. For brevity, the notation $T_{Y\to X}$ without the superscript d is used below for such "almost exact" values of TE.

If u and v are continuous-valued and characterized by (piecewise-) continuous probability density functions, then in the definition of TE (1) the probability distributions P should be replaced by the probability densities p and the sums by the integrals:

$$T_{Y\to X}^d = -\int p(u_n,\mathbf{u}_n^d)\log_2 p(u_n|\mathbf{u}_n^d)\,du_n\,d\mathbf{u}_n^d + \int p(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)\log_2 p(u_n|\mathbf{u}_n^d,\mathbf{v}_n^d)\,du_n\,d\mathbf{u}_n^d\,d\mathbf{v}_n^d. \quad (2)$$

Here TE is the difference of the differential Shannon entropies, while the usual entropies are infinitely large for continuous-valued variables. However, the TE interpretation does not differ from the discrete case. Indeed, consider coarse-grained distributions of the continuous-valued quantities with a certain bin size δ and introduce a "coarse-grained TE" according to Eq. (1). It would converge to (2) at $\delta \to 0$. Thus, it makes a clear sense that the "differential" TE (2) is also measured in bits and characterizes the decrease in uncertainty in the same way as in the discrete case. Below the definition (2) is used for Gaussian processes (Sec. III A) and the definition (1) for Markov chains (Sec. III B).

III. TRANSFER ENTROPY FORMULAS FOR BENCHMARK SYSTEMS

Denote with $\mathbf{z}$ a d_Z-dimensional state vector of the combined system Z consisting of the systems X and Y. "State vector" relates to the concept of state space models and means that the value $\mathbf{z}_n$ at a current time instant n completely determines the probability distribution of $\mathbf{z}_{n+1}$. If one denotes $\mathbf{z}_n^d = (\mathbf{z}_{n-1}^{\mathrm{T}}, \mathbf{z}_{n-2}^{\mathrm{T}}, \ldots, \mathbf{z}_{n-d}^{\mathrm{T}})^{\mathrm{T}}$, then the latter property reads as equality of the conditional distributions: $P(\mathbf{z}_{n+1}|\mathbf{z}_n) = P(\mathbf{z}_{n+1}|\mathbf{z}_n, \mathbf{z}_n^d)$ for any $d > 0$; i.e., $\mathbf{z}_n$ "shields" the future from the past so that $\mathbf{z}_n$ is a first-order vector-valued Markovian process. The vector $\mathbf{z}$ consists of the two components: $\mathbf{x}$ of dimension d_X and $\mathbf{y}$ of dimension d_Y, which characterize the systems X and Y, respectively; i.e., $\mathbf{z}_n = (\mathbf{x}_n^{\mathrm{T}}, \mathbf{y}_n^{\mathrm{T}})^{\mathrm{T}}$ and $d_Z = d_X + d_Y$.

Being isolated from each other, the systems X and Y would be completely described by their own state vectors $\mathbf{x}$ and $\mathbf{y}$ so that the processes $\mathbf{x}_n$ and $\mathbf{y}_n$ would satisfy the generalized

Markov property

$$P(\mathbf{x}_{n+1}|\mathbf{x}_n) = P(\mathbf{x}_{n+1}|\mathbf{x}_n,\mathbf{y}_n,\mathbf{z}_n^d), \tag{3}$$

$$P(\mathbf{y}_{n+1}|\mathbf{y}_n) = P(\mathbf{y}_{n+1}|\mathbf{y}_n,\mathbf{x}_n,\mathbf{z}_n^d), \tag{4}$$

for any $d > 0$; i.e., given $\mathbf{x}_n$ ($\mathbf{y}_n$), the future $\mathbf{x}$ ($\mathbf{y}$) does not depend on a deeper history of X (Y) and any current or past values of $\mathbf{y}$ ($\mathbf{x}$). An influence $X \to Y$ would manifest itself in a dependence of $\mathbf{y}_{n+1}$ on some values of $\mathbf{x}$, given $\mathbf{y}_n$, i.e., in violation of Eq. (4). An influence $Y \to X$ violates Eq. (3). Below examples with a UC $X \to Y$ are considered that corresponds to the system Z for which Eq. (3) holds true and Eq. (4) is violated.

Let u and v be single-valued functions of the state vectors $u = h_u(\mathbf{x})$ and $v = h_v(\mathbf{y})$. Without loss of generality, it is further always assumed that $u = x_1$ and $v = y_1$; i.e., the first component of each state vector is observed.

A. Transfer entropy for Gaussian processes

The first exemplary class of systems is given by stochastic linear difference equation

$$\mathbf{z}_n = \mathbf{C} \cdot \mathbf{z}_{n-1} + \boldsymbol{\xi}_n, \tag{5}$$

where $\mathbf{C}$ is a constant matrix of dimension $d_X + d_Y$, $\boldsymbol{\xi}_n$ is a stationary Gaussian white noise, i.e., a sequence of independent random vectors identically distributed according to a Gaussian law with zero mean and a covariance matrix $\boldsymbol{\Sigma}$. The submatrix C_{ij}, $1 \leqslant i,j \leqslant d_X$, reflects individual dynamical properties of the system X; the submatrix C_{ij}, $d_X + 1 \leqslant i,j \leqslant d_X + d_Y$, specifies the dynamics of the system Y. The submatrix C_{ij}, $1 \leqslant i \leqslant d_X, d_X + 1 \leqslant j \leqslant d_X + d_Y$, determines the influence of Y on X and equals zero. The submatrix C_{ij}, $d_X + 1 \leqslant i \leqslant d_X + d_Y$, $1 \leqslant j \leqslant d_X$, determines the influence $X \to Y$. Similarly, the respective "on-diagonal" submatrices of the symmetric matrix $\boldsymbol{\Sigma}$ determine individual properties of noises within the systems X and Y, while the off-diagonal submatrix determines noise covariances across the systems.

To compute $T_{Y\to X}$, one must find the densities $p(u_n,\mathbf{u}_n^d)$, $p(\mathbf{u}_n^d)$, $p(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)$, and $p(\mathbf{u}_n^d,\mathbf{v}_n^d)$ which completely determine the right-hand side of Eq. (2) since $p(u_n|\mathbf{u}_n^d) = p(u_n,\mathbf{u}_n^d)/p(\mathbf{u}_n^d)$ and $p(u_n|\mathbf{u}_n^d,\mathbf{v}_n^d) = p(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)/p(\mathbf{u}_n^d,\mathbf{v}_n^d)$. These probability densities are Gaussian with zero means, and their covariance matrices can be denoted $\mathbf{R}(u_n,\mathbf{u}_n^d)$, $\mathbf{R}(\mathbf{u}_n^d)$, $\mathbf{R}(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)$, and $\mathbf{R}(\mathbf{u}_n^d,\mathbf{v}_n^d)$. These are square matrices of dimensions $d+1$, d, $2d+1$, and $2d$, respectively. According to Ref. [41], the conditional entropies in Eq. (2) relate to the determinants of these matrices as

$$T_{Y\to X}^d = \frac{1}{2}\log_2 \frac{|\mathbf{R}(u_n,\mathbf{u}_n^d)|}{|\mathbf{R}(\mathbf{u}_n^d)|} - \frac{1}{2}\log_2 \frac{|\mathbf{R}(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)|}{|\mathbf{R}(\mathbf{u}_n^d,\mathbf{v}_n^d)|}. \tag{6}$$

All these matrices can be found via the selection of appropriate elements from the full covariance matrix of the vector $\mathbf{z}_{n+1}^{d+1}$. This matrix is formed by elements of the following matrices: $\mathbf{R}_{\mathbf{z},0} = E[\mathbf{z}_n\mathbf{z}_n^{\mathrm{T}}]$ (covariance matrix of the vector $\mathbf{z}_n$), $\mathbf{R}_{\mathbf{z},1} = E[\mathbf{z}_n\mathbf{z}_{n-1}^{\mathrm{T}}]$ (covariance between the vectors $\mathbf{z}_n$ and $\mathbf{z}_{n-1}$), ..., $\mathbf{R}_{\mathbf{z},d} = E[\mathbf{z}_n\mathbf{z}_{n-d}^{\mathrm{T}}]$ (covariance between the vectors $\mathbf{z}_n$ and $\mathbf{z}_{n-d}$), where E stands for the expectation value. To find $\mathbf{R}_{\mathbf{z},0}$, transpose both sides of (5) and get $\mathbf{z}_n^{\mathrm{T}} = \mathbf{z}_{n-1}^{\mathrm{T}} \cdot \mathbf{C}^{\mathrm{T}} + \xi_n^{\mathrm{T}}$, then multiply both sides of (5) by their transposed versions, and take the expectation value. Thereby, one gets $\mathbf{R}_{\mathbf{z},0} = \mathbf{C} \cdot \mathbf{R}_{\mathbf{z},0} \cdot \mathbf{C}^{\mathrm{T}} + \boldsymbol{\Sigma}$. This is a linear set of algebraic equations with respect to the elements of $\mathbf{R}_{\mathbf{z},0}$ which can be solved by any standard technique. By multiplying both sides of (5) by $\mathbf{z}_{n-k}^{\mathrm{T}}$ ($k > 0$) from the right and taking the expectation value, one further gets a recursive formula $\mathbf{R}_{\mathbf{z},k} = \mathbf{C} \cdot \mathbf{R}_{\mathbf{z},k-1}$, which allows us to compute all the necessary $\mathbf{R}_{\mathbf{z},k}$, $1 \leqslant k \leqslant d$. Now, the matrix $\mathbf{R}(\mathbf{u}_n^d)$ is formed by the elements of $\mathbf{R}_{\mathbf{z},k}$ with the index (1,1), $0 \leqslant k \leqslant d-1$ (recall that $u = x_1 = z_1$ by convention). Similarly, all the other matrices entering Eq. (6) are formed by the appropriate elements of $\mathbf{R}_{\mathbf{z},k}$ and, thus, $T_{Y\to X}$ can be computed via Eq. (6).

B. Transfer entropy for Markov chains

Suppose that each of the variables $x_1, x_2, \ldots, x_{d_X}, y_1, y_2, \ldots, y_{d_Y}$ can take on B different values. Then $\mathbf{z}$ can take on B^{d_Z} values. Denote them ζ_k, $k = 1, \ldots, B^{d_Z}$. The dynamics of a Markov chain are determined by its transition probabilities matrix $\mathbf{A}$ of a dimension B^{d_Z}, whose elements A_{ik} are probabilities of transitions from a state ζ_k to a state ζ_i:

$$A_{ik} = \mathbf{P}\{\mathbf{z}_n = \zeta_i|\mathbf{z}_{n-1} = \zeta_k\}. \tag{7}$$

If a probability distribution of $\mathbf{z}$ at time instant $n-1$ is specified by a vector $\mathbf{P}_{n-1}$ with elements $P_{n-1,k} = \mathbf{P}\{\mathbf{z}_{n-1} = \zeta_k\}$, then the distribution at the next instant is given by $\mathbf{P}_n = \mathbf{A} \cdot \mathbf{P}_{n-1}$. A stationary distribution $P(\mathbf{z})$ is then represented by a vector $\mathbf{P}_{\mathrm{st}}$ which satisfies $\mathbf{P}_{\mathrm{st}} = \mathbf{A} \cdot \mathbf{P}_{\mathrm{st}}$ and the normalization condition (the sum of all its elements equals unity). These two conditions allow us to find $\mathbf{P}_{\mathrm{st}}$ uniquely from a linear set of algebraic equations. Starting from that one-dimensional distribution $P(\mathbf{z})$, one finds a stationary $(d+1)$-dimensional distribution $P(\mathbf{z}_{n+1}^{d+1})$ by recursively multiplying a lower-dimensional distribution by the respective transition probabilities (7). Having the stationary distribution $P(\mathbf{z}_{n+1}^{d+1})$, one finds all the distributions $P(u_n,\mathbf{u}_n^d)$, $P(\mathbf{u}_n^d)$, $P(u_n,\mathbf{u}_n^d,\mathbf{v}_n^d)$, and $P(\mathbf{u}_n^d,\mathbf{v}_n^d)$ determining the right-hand side of Eq. (1) via the summation over the respective variables. In particular, to find $P(\mathbf{u}_n^d)$, one sums (marginalizes) $P(\mathbf{z}_{n+1}^{d+1})$ over all the variables which are not components of the vector $\mathbf{u}_n^d$. Everything is analogous for the other three distributions. Next, the value of $T_{Y\to X}$ can be computed directly from Eq. (1).

IV. TYPICAL EXAMPLES OF SPURIOUS CAUSALITIES

Gaussian processes (5) and Markov chains (7) with different concrete structures and parameter values illustrate below typical practical situations leading to spurious causalities: unobserved state variable (Sec. IV A), low temporal resolution (Sec. IV B), and observational noise (Sec. IV C). Namely, for the case of UC $X \to Y$ considered, it is shown that the values of $T_{Y\to X}$ can be nonzero and, moreover, even greater than the "correct" nonzero $T_{X\to Y}$. Gaussian processes represent either "relaxation systems" or oscillators: An initial perturbation in the noise-free case tends to zero either in a nonoscillatory or an oscillatory way. Markov chains are constructed in some analogy with the Gaussian processes to mimic their positive or negative autocorrelations. However, being simple versions of nonlinear systems, Markov chains are particularly useful to show the general character of the results.

For the Gaussian processes at each set of the parameter values, TEs (2) are computed for d ranging from 1 to 20. Convergence with a relative error $\varepsilon = |T^{d+1}_{X\to Y} - T^{d}_{X\to Y}|/T^{d}_{X\to Y} < 10^{-7}$ is achieved in all the examples at most for $d \leqslant 9$ and often for $d = 4$ or $d = 5$. For uniformity, all the results are presented for $d = 10$ to ensure the convergence. For the Markov chains, $B = 2$ is used, and convergence of (1) over d with $\varepsilon < 10^{-2}$ is achieved for $d = 6$ or earlier. For uniformity, all the results are presented for $d = 7$.

TE values are studied versus different parameters of the systems as follows. First, certain starting values (a starting set) of the parameters are specified. Then one of the parameters is varied in order to maintain stability of the system under study, while all the others are kept equal to their starting values. For each set of values of the parameters, TEs in both directions are computed along with their ratio $r = T_{Y\to X}/T_{X\to Y}$, which is a relative measure of the spurious coupling effect. It is necessary to have $r = 0$ or at least $r << 1$ to infer the UC $X \to Y$ from TEs correctly. However, this is often not the case, as shown below.

A. Unobserved state variable

This subsection deals with a widespread practical situation where certain state variables of the driving system X (some components of the vector $\mathbf{x}$) are hidden; i.e., the state of the system X is not completely observed.

1. Transfer entropies for Gaussian processes with unobserved state variables

Consider a two-dimensional process X given by the equations

$$x_{1,n} = c_{11}x_{1,n-1} + c_{12}x_{2,n-1} + \xi_{1,n},$$
$$x_{2,n} = c_{21}x_{1,n-1} + c_{22}x_{2,n-1} + \xi_{2,n}, \quad (8)$$

where Gaussian white noises ξ_1,ξ_2 have variances Σ_{11},Σ_{22} and a covariance Σ_{12}. The elements c_{ij} determine whether the system (8) is an oscillator or a relaxation system. Namely, the roots of the characteristic equation for this linear system are $\gamma_{1,2} = (a \pm \sqrt{a^2 - 4b})/2$, where $a = c_{11} + c_{22}$ and $b = c_{11}c_{22} - c_{12}c_{21}$. If $|\gamma_{1,2}| < 1$, then the system (8) at zero noise level has a stable fixed point at the origin. If $a^2 - 4b < 0$, then a transient process represents exponentially decaying oscillations; otherwise the system converges to the fixed point in a nonoscillatory way. There exist many real-world examples described by similar equations [42] where two variables determine a state, but only one of them is observed ($u = x_1$). As the response system Y, consider a one-dimensional relaxation process

$$y_{1,n} = c_{32}x_{2,n-1} + c_{33}y_{1,n-1} + \xi_{3,n}, \quad (9)$$

where Gaussian white noise ξ_3 has a variance Σ_{33} and does not depend on the noise in the system X, i.e., $\Sigma_{13} = \Sigma_{23} = 0$. At $\Sigma_{33} = 0$ and $c_{32} = 0$, the system Y has a fixed stable point if $|c_{33}| < 1$ and the value of $|c_{33}|$ determines the speed of relaxation. The UC $X \to Y$ is realized here via the hidden variable x_2. An observable is $v = y_1$.

TEs are studied versus c_{ij} and Σ_{ij} at a starting set of the parameters $c_{11} = c_{22} = 1/2$, $c_{12} = c_{21} = 1/4$, $c_{32} = 1$, $c_{33} = 0.1$, $\Sigma_{11} = \Sigma_{22} = 1$, and $\Sigma_{33} = 0.1$, where the system X is a relaxation system consisting of two identical relaxation processes with moderate (as compared to c_{11}) coupling coefficients c_{12} and c_{21}.

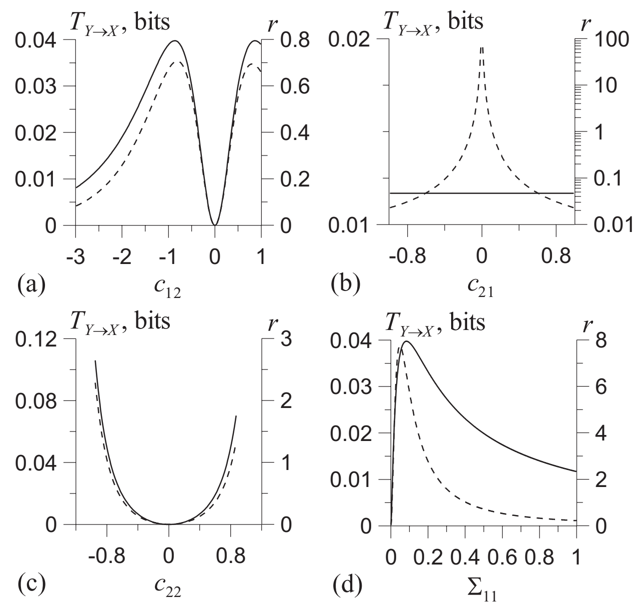

FIG. 1. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the Gaussian process (8) and (9) versus parameters: (a) c_{12} determines the influence of the hidden variable x_2 on the observable $u = x_1$, spurious causality is not considerable if c_{12} is small enough; (b) c_{21} determines the influence of $u = x_1$ on x_2 and, thereby, on Y so that r is small if c_{21} is large enough; (c) c_{22} controls individual properties of the hidden process x_2 whose amplitude rises with $|c_{22}|$ making stronger the spurious coupling effect; (d) Σ_{11} controls individual properties (amplitude) of the observed process $u = x_1$.

First, Fig. 1 shows positive "spurious" $T_{Y\to X}$ in wide ranges of the parameter values, which is the most important point. The principal existence of the spurious coupling effect is explained as follows. If one observed a full state vector of the driving system X [$\mathbf{u} = (x_1,x_2)^{\mathrm{T}}$ or $\mathbf{u} = \mathbf{h}_u(\mathbf{x})$ with $\mathbf{h}_u$ a one-to-one function], then a distribution of $\mathbf{u}_n$ would be completely determined by the previous value $\mathbf{u}_{n-1}$ and, given $\mathbf{u}_{n-1}$, would not depend on the previous dynamics of Y, i.e., one would get $T_{Y\to X} = 0$ naturally reflecting the absence of the influence $Y \to X$. Since one of the components of the state vector of X is unobserved, then the uncertainty in u_n, given the entire history u_{n-k}, $k > 0$, is greater than the uncertainty in u_n, given $\mathbf{x}_{n-1} = (x_{1,n-1},x_{2,n-1})^{\mathrm{T}}$. This is because the unobserved value $x_{2,n-1}$ cannot be restored absolutely accurately from all the observed u_{n-k}, $k > 0$. At that, an interdependence between the variables v and x_2 exists due to the influence $X \to Y$. Therefore, the variable v may carry additional information about the hidden variable x_2, and, thereby, knowledge of the previous values of v may decrease uncertainty in u_n as compared to knowledge of only the previous values of u. Thus, it is an incomplete observation of the driving system state vector which leads to nonzero $T_{Y\to X}$ and r.

Second, numerical values of the "spurious" TE reach 0.1 bits [Fig. 1(c)] that should be considered a large value for the following reason. TE multiplied by 2 ln 2 equals approximately the normalized mean-squared prediction improvement [41], since TE is proportional to the logarithm of the ratio of the conditional variances (6). Thus, the TE value of 0.1 bits corresponds to the normalized prediction improvement of u (when v is taken into account) by more than 10%, which is quite considerable. Third, the relative spurious causality measure r can take on arbitrarily large values [Fig. 1(b)] and often exceeds unity. The latter means that the "spurious" $T_{Y\to X}$ exceeds the "correct" $T_{X\to Y}$. Fourth, the spurious coupling effect does not depend on whether the system X is an oscillator or a relaxation system: indeed, X is an oscillator at $c_{12} < 0$ in Fig. 1(a) [at $c_{21} < 0$ in Fig. 1(b)] and a relaxation system at $c_{12} > 0$ (at $c_{21} > 0$), while the values of TE are symmetric with respect to zero.

Fifth, how TE and r depend on different parameters can be explained by properties of the systems under study. For example, consider Fig. 1(b) where r is especially large at small c_{21}. Indeed, at $c_{21} = 0$ there is no influence of the observed variable $u = x_1$ on the hidden variable x_2, and, hence, there is no influence of u on v. Thus, one has $T_{X\to Y} = 0$, while the value of $T_{Y\to X}$ is nonzero for the same reasons as described above (correlation between v and x_2). Hence, it holds $r = \infty$. This situation can be also interpreted as the influence of the unobserved third process (x_2) on the observed processes u and v, which is a well-known problem. However, under the considered problem setting and from a physical viewpoint, it may well occur that x_1 and x_2 definitely belong to the same physical system so that the "hidden-third-system" interpretation is not as appropriate (rather formal) as the "hidden-state-variable" view. Note also that the spurious coupling effect disappears at $c_{12} = 0$ [Fig. 1(a)], which is explained by the absence of the influence of x_2 on $x_1 = u$ so that u is a complete state vector of the subsystem x_1. Then there is no situation of an incomplete state observation for that subsystem, while the influence of x_1 on v is mediated by the hidden variable x_2 and is correctly detected. Thus, the most "dangerous" situation in the sense of spurious couplings is that when (i) the driving system X consists of several components with their own degrees of freedom, (ii) some of those components are hidden, and (iii) the hidden components of X strongly influence both the system Y and the observed components of X. Indeed, in such a case the analysis based on a single variable u is not adequate to a "nonuniform" (more complex) structure of the system X.

All these observations show that the spurious causality due to unobserved state variable of the driving system is a generic (occurring not only at very special parameter values) effect, being often quite strong quantitatively.

Now, consider the case when the system Y is also two-dimensional:

$$\begin{aligned} y_{1,n} &= c_{33}y_{1,n-1} + c_{34}y_{2,n-1} + \xi_{3,n} + c_{32}x_{2,n-1}, \\ y_{2,n} &= c_{43}y_{1,n-1} + c_{44}y_{2,n-1} + \xi_{4,n}, \end{aligned} \tag{10}$$

where Gaussian white noises ξ_3,ξ_4 have variances Σ_{33},Σ_{44} and a covariance Σ_{34} and do not depend on the noise in the system X, i.e., $\Sigma_{ij} = 0, 1 \leqslant i \leqslant 2, 3 \leqslant j \leqslant 4$. The individual structure of the system Y is similar to that of the system X. Both identical and strongly different systems X and Y, both oscillators and relaxation systems, are studied. In all cases, the results appear quite similar to those in Fig. 1. For an illustration, consider only identical systems with the starting set $c_{kk} = 0.9, 1 \leqslant k \leqslant 4$, $c_{12} = c_{34} = 0.1$, $c_{21} = c_{43} = -0.1$, $c_{32} = 0.4$, $\Sigma_{11} = \Sigma_{33} = 0.1$, $\Sigma_{22} = \Sigma_{44} = 1$, $\Sigma_{12} = \Sigma_{34} = 0$, that approximately corresponds to the Euler integration scheme for some continuous-time linear oscillators. Parameters in Fig. 2 are varied such that $c_{11} = c_{33}$, $c_{22} = c_{44}$, $c_{12} = c_{34}$, $c_{21} = c_{43}$, $\Sigma_{11} = \Sigma_{33}$, $\Sigma_{22} = \Sigma_{44}$, and $\Sigma_{12} = \Sigma_{34}$.

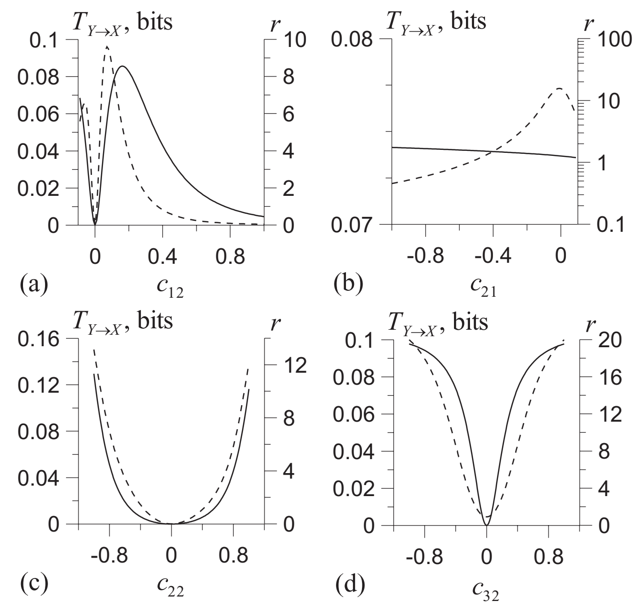

FIG. 2. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the Gaussian process (8) and (10) versus parameters: (a) c_{12}, the influence of the hidden variable x_2 on the observable $u = x_1$; (b) c_{21}, the influence of $u = x_1$ on x_2; (c) c_{22}, controls amplitude and autocorrelations of x_2; (d) c_{32}, the strength of the influence $X \to Y$.

Again, nonzero "spurious" TE values are generic and not small, quantitatively close to the above example (Fig. 1) and even a bit greater. Note that the value of r does not tend to zero even for vanishing coupling strength c_{32} [Fig. 2(d)]. Interestingly, r is often greater than unity and even greater than 10; i.e., it appears much more probable to observe a greater TE in the "spurious" direction for *identical* oscillators! This counterintuitive result relates to the way how coupling is realized: The hidden variable x_2 affects the observed variable y_1; i.e., the coupling is "not symmetric" with respect to variables. In a general case of four nonzero coefficients c_{31}, c_{32}, c_{41}, and c_{42}, the values of r may appear somewhat smaller (not shown), but the spurious coupling effect *per se* remains generic and typically considerable.

2. Transfer entropies for Markov chains with unobserved state variables

Consider a Markov chain with two binary variables which would be an analog of the linear systems described above. First, as a binary analog of the Gaussian white noise, take a sequence of independent random variables taking on the values of 0 or 1 with equal probabilities. Then, an analog of a relaxation system $x_{1,n} = c_{11}x_{1,n-1} + \xi_{1,n}$ is a Markov chain with the transition probabilities matrix of the form

$$\begin{aligned} A_{11} &= \mathbf{P}\{x_{1,n} = 0|x_{1,n-1} = 0\} = 1/2 + p_{11}, \\ A_{12} &= \mathbf{P}\{x_{1,n} = 0|x_{1,n-1} = 1\} = 1/2 - p_{11}, \\ A_{21} &= \mathbf{P}\{x_{1,n} = 1|x_{1,n-1} = 0\} = 1/2 - p_{11}, \\ A_{22} &= \mathbf{P}\{x_{1,n} = 1|x_{1,n-1} = 1\} = 1/2 + p_{11}, \end{aligned} \tag{11}$$

where p_{11} has the same sign as c_{11}. Thus, a positive c_{11} leads to a positive correlation between successive values of x_1. Similarly, a positive p_{11} provides a probability of $x_{1,n} = x_{1,n-1}$ greater than 1/2. These are processes with a tendency to "permanence." Negative c_{11} and p_{11} correspond to the opposite tendency to "alternation."

Now, consider a Markov chain with two state variables which, hence, has four states: $(x_1,x_2) = (0,0),(0,1),(1,0),(1,1)$. To make it analogous to the two coupled relaxation systems (8), let the value $x_{2,n-1}$ change the probabilities of $x_{1,n}$ by a certain value p_{12}:

$$\begin{aligned} &\mathbf{P}\{x_{1,n} = 0|x_{1,n-1} = 0,x_{2,n-1} = 0\} = 1/2 + p_{11} + p_{12}, \\ &\mathbf{P}\{x_{1,n} = 0|x_{1,n-1} = 0,x_{2,n-1} = 1\} = 1/2 + p_{11} - p_{12}, \\ &\mathbf{P}\{x_{1,n} = 0|x_{1,n-1} = 1,x_{2,n-1} = 0\} = 1/2 - p_{11} + p_{12}, \\ &\mathbf{P}\{x_{1,n} = 0|x_{1,n-1} = 1,x_{2,n-1} = 1\} = 1/2 - p_{11} - p_{12}. \end{aligned} \tag{12}$$

Everything is similar for the conditional probabilities of $x_{2,n}$ with probability shifts p_{22} and p_{21} instead of p_{11} and p_{12}. Let $x_{1,n}$ and $x_{2,n}$ be independent of each other, given $x_{1,n-1}$ and $x_{2,n-1}$, analogously to mutually independent noises ξ_1 and ξ_2 in Eq. (8):

$$\begin{aligned} &P(x_{1,n},x_{2,n}|x_{1,n-1},x_{2,n-1}) \\ &\quad = P(x_{1,n}|x_{1,n-1},x_{2,n-1})P(x_{2,n}|x_{1,n-1},x_{2,n-1}). \end{aligned} \tag{13}$$

From Eqs. (12) and (13), all 16 elements of the entire transition probabilities matrix can be computed through similar multiplications, e.g.,

$$\begin{aligned} A_{12} &= P\{x_{1,n} = 0,x_{2,n} = 0|x_{1,n-1} = 0,x_{2,n-1} = 1\} \\ &= (1/2 + p_{11} - p_{12})(1/2 - p_{22} + p_{21}). \end{aligned} \tag{14}$$

The obtained system X is a binary analog of Eq. (8), where the signs of p_{ij} correspond to the signs of c_{ij}. Again, an observed variable is $u = x_1$, a hidden one is x_2.

Let the system X drive a relaxation system Y which is described by a single binary variable y_1 similarly to Eq. (11) with a probability shift p_{33} instead of p_{11}. Let the driving be mediated through the variable x_2, which shifts probabilities of $y_{1,n}$ by the value of p_{32}:

$$\begin{aligned} &\mathbf{P}\{y_{1,n} = 0|y_{1,n-1} = 0, \quad x_{2,n-1} = 0\} = 1/2 + p_{33} + p_{32}, \\ &\mathbf{P}\{y_{1,n} = 0|y_{1,n-1} = 0, \quad x_{2,n-1} = 1\} = 1/2 + p_{33} - p_{32}, \\ &\mathbf{P}\{y_{1,n} = 0|y_{1,n-1} = 1, \quad x_{2,n-1} = 0\} = 1/2 - p_{33} + p_{32}, \\ &\mathbf{P}\{y_{1,n} = 0|y_{1,n-1} = 1, \quad x_{2,n-1} = 1\} = 1/2 - p_{33} - p_{32}, \end{aligned} \tag{15}$$

where $y_{1,n}$ is independent of $x_{1,n-1}$, given $y_{1,n-1}$ and $x_{2,n-1}$. Due to the unidirectional coupling character, $x_{1,n}$ and $x_{2,n}$ are determined by $x_{1,n-1}$ and $x_{2,n-1}$ (12), being conditionally independent of $y_{1,n-1}$. Similarly to Eq. (13), the three quantities $x_{1,n},x_{2,n},y_{1,n}$ are mutually independent given $x_{1,n-1},x_{2,n-1},y_{1,n-1}$. All this specifies completely a Markov chain with eight states, whose transition probabilities matrix is found (similarly to the matrix of the isolated system X) through multiplications of the three conditional probabilities.

As a starting set of the parameters, take $p_{11} = p_{22} = 1/4$ (sufficiently strong "permanence" tendency of the system X), $p_{12} = p_{21} = 1/8$ (moderately strong symmetric coupling between the variables x_1 and x_2), $p_{33} = 0$ (an isolated system Y would generate white noise), and $p_{32} = 1/2$ (one-to-one relationship between $x_{2,n-1}$ and $y_{1,n}$ so that a precise restoration of $x_{2,n-2}$ from $y_{1,n-1}$ takes place and, hence, a relatively good restoration of $x_{2,n-1}$ desirable to predict u_n may occur). Figure 3 shows generically nonzero "spurious" TE versus different parameters and resembles Figs. 1 and 2 in many respects. In particular, Fig. 3(c) is qualitatively similar to Figs. 1(c) or 2(c), which show TE versus c_{22} in the systems (8) and (9) or (10). The explanation of the nonzero values of $T_{T\to X}$ is the same as that for the Gaussian processes. Numerical values of $T_{T\to X} \approx 0.03$ bits achieved at certain p_{22} correspond to the Shannon entropies for the distribution of u, conditioned by the past of u, equal to 0.31 bits. Thus, 0.03 bits corresponds to the relative reduction of uncertainty (entropy) about 10%; i.e., the spurious couplings are rather high and as considerable as in the previous example.

All the results are similar for a two-variable driven system Y in analogy to what is demonstrated in Sec. IV A1 for the two-dimensional driven Gaussian process (10). They are not shown for brevity since the corresponding Markov chain with 16 states is specified by quite a cumbersome expression for its transition probabilities matrix.

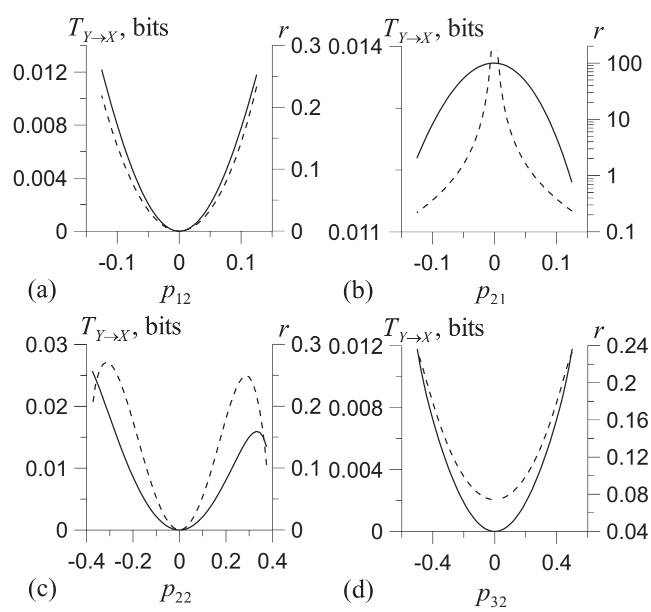

FIG. 3. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the Markov chain (12), (13), and (15) versus parameters: (a) p_{12}, the influence of the hidden variable x_2 on the observable $u = x_1$; (b) p_{21}, the influence of $u = x_1$ on x_2; (c) p_{22}, controls individual properties (autocorrelations) of x_2; (d) p_{32}, the strength of the influence $X \to Y$.

Note that Markov chains can be appropriate models of continuous-state nonlinear systems (see, e.g., Refs. [40,43,44]), which is especially clear if one considers a higher resolution (the number of possible values of state variables B is greater than 2) to make a Markov chain a good piecewise-constant stochastic analog of a nonlinear chaotic map. It follows from the above consideration that all the results are expected to be qualitatively the same for large B, and, hence, the spurious coupling effect should be typical of diverse nonlinear systems.

B. Low temporal resolution

This subsection deals with another situation, where observed variables would contain complete information about the systems states if sampled properly, but insufficient temporal resolution of the data leads to spurious causalities.

Consider a discrete-time model of a stochastic linear dissipative oscillator in the form of a second-order autoregression process

$$x_n = a_x x_{n-1} + b_x x_{n-2} + \xi_{x,n}, \tag{16}$$

where Gaussian white noise ξ_x has a variance σ_x^2, while a quasiperiod of the decaying oscillations T_x and their relaxation time τ_x are given by $a_x = 2\cos(2\pi/T_x)\exp(-1/\tau_x)$ and $b_x = -\exp(-2/\tau_x)$ [45]. An observable is $u_n = x_{2n}$; i.e., a sampling interval is equal to two time units so that every second value of x is lost, which is a kind of downsampling. This case reduces to the formalism of Sec. III A if one introduces a two-dimensional state vector with coordinates $x_{1,n} = x_{2n}$ (an observed variable) and $x_{2,n} = x_{2n-1}$ (a hidden variable). Then Eq. (16) can be rewritten in the form (8) where $c_{11} = a_x^2 + b_x, c_{12} = a_x b_x, c_{21} = a_x, c_{22} = b_x, \Sigma_{11} = \sigma_x^2(1 + a_x^2), \Sigma_{22} = \sigma_x^2, \Sigma_{12} = \sigma_x^2 a_x$, and a single time step in Eq. (8) corresponds to two time steps in Eq. (16). Note the nonzero noise covariance Σ_{12} arising in this example.

Let the driven system Y be a similar oscillator

$$y_n = a_y y_{n-1} + b_y y_{n-2} + c x_{n-1} + \xi_{y,n}, \tag{17}$$

where ξ_y is Gaussian white noise with variance σ_y^2 which is independent of ξ_x, c is a coupling coefficient, $a_y = 2\cos(2\pi/T_y)\exp(-1/\tau_y)$, $b_y = -\exp(-2/\tau_y)$, and an observable is $v_n = y_{2n}$. By denoting $y_{1,n} = y_{2n}$ and $y_{2,n} = y_{2n-1}$, one gets the system Y in the form (10) with $c_{33} = a_y^2 + b_y, c_{34} = a_y b_y, c_{43} = a_y, c_{44} = b_y, \Sigma_{33} = \sigma_y^2(1 + a_y^2), \Sigma_{44} = \sigma_y^2, \Sigma_{34} = \sigma_y^2 a_y$, coupling coefficients $c_{31} = (a_x + a_y)c, c_{32} = b_x c, c_{41} = c, c_{42} = 0$, and noise covariances $\Sigma_{13} = a_x \sigma_x^2 c, \Sigma_{14} = \Sigma_{24} = 0, \Sigma_{23} = \sigma_x^2 c$. Since the system (16) and (17) is reduced to the form (8) and (10), it can be studied with the same formalism. Thus, low temporal resolution appears mathematically equivalent to the situation of unobserved state variable: Imperfect observations of the states of X occur here because a complete state of X would consist of the two adjacent values (x_{n-1}, x_n), but one of them is lost due to the downsampling.

For a more diverse analysis, TE is studied below versus such physically meaningful parameters as basic oscillation periods and relaxation times of the systems. As a starting set of values, specify $T_x = T_y = 5$, $\tau_x = 10$, $\tau_y = 1$, $\sigma_x^2 = 1$,

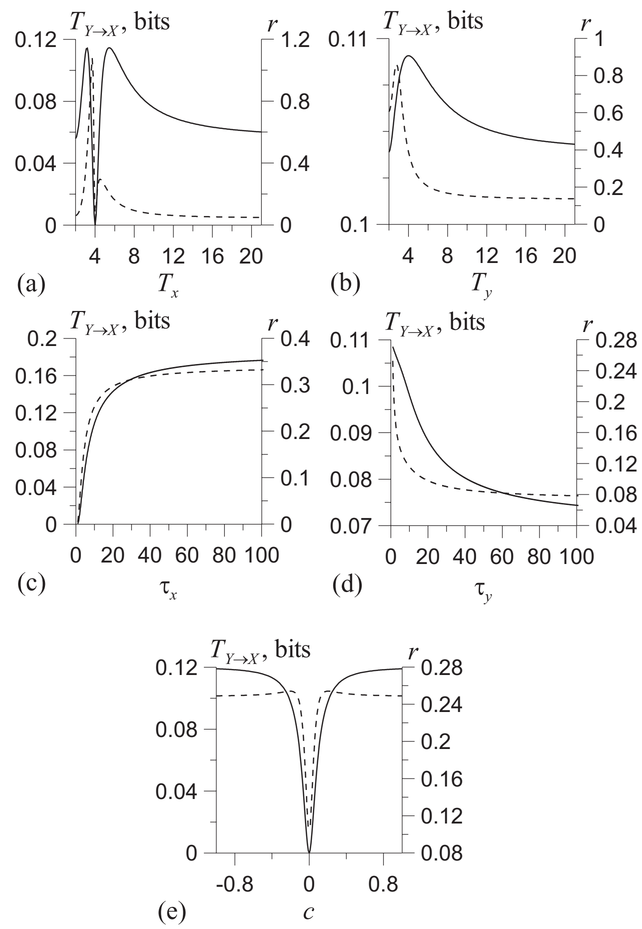

FIG. 4. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the linear oscillators (16) and (17) in the case of downsampling versus parameters: (a, b) and T_x and T_y, individual periods of the oscillators; (c, d) τ_x and τ_y, their relaxation times; (e) c, the strength of the influence $X \to Y$.

$\sigma_y^2 = 0.01$, and $c = 0.3$. Figure 4 evidences that the "spurious" TE is generically nonzero and takes on considerable values (up to 0.1–0.2 bits) and the values of r can be rather large (about 0.1–0.3 and sometimes up to unity and even greater) in analogy with Sec. IV A1. In particular, $T_{Y\to X}$ is maximal at $T_x = 3.2$ or $T_x = 5.4$ [Fig. 4(a)] and at $T_y = 4$ [Fig. 4(b)]. It rises with τ_x [Fig. 4(c)] and falls with τ_y [Fig. 4(d)]. The plots versus the coupling strength c [Fig. 4(e)] are rather similar to the previous examples [Figs. 2(d) and 3(d)]. Note that the largest values of $T_{Y\to X}$ and r take place for nonidentical oscillators; i.e., when their oscillation periods as well as relaxation times differ from each other. As the oscillators become "more identical" with respect to those parameters, the values of $T_{Y\to X}$ and r decrease. Thus, the spurious coupling effect is not so strong here for identical oscillators, unlike Sec. IV A1 where the "spurious" TE is large for identical systems. This distinction is due to the different forms of coupling in the representation (8) and (10): Here the driving is performed via both variables x_1 and x_2 (not only via the hidden variable x_2) on both variables of the system Y (rather than only on y_1). Still, this example illustrates further the general character of the spurious coupling effect, which is shown to be considerable for the coupling character different from that of Sec. IV A1. Variations in the coupling form lead to some quantitative differences, while the qualitative picture of the spurious couplings remains.

For a Markov chain analogous to the above oscillators, an effect of temporal resolution on TEs can be considered like in Sec. IV A2. This is done in Appendix A, and the results appear quite similar to those presented above confirming their general character. Thus, low temporal resolution can induce considerable spurious couplings. Such a situation may well be widespread in practice since a sampling interval is determined by the conditions of observations or measurements and may not agree with intrinsic time scales of the systems under study. Low temporal resolution represents another particular case of imperfect observations of the driving system states, in addition to the case of hidden state variables.

C. Observational noise

This subsection deals with the third factor, which can induce spurious causalities as was shown earlier for the linear Granger causality [39] and for TE and nonlinear maps of a special form [40]. Here it is placed into a broader context so that the present study complements the results of Refs. [39,40] and shows that their common cause can again be formulated as imperfect observations of the driving system states.

Consider a relaxation system X specified by the equation

$$x_n = a_x x_{n-1} + \xi_{x,n}, \tag{18}$$

where the noise ξ_x has a variance $\sigma^2_{\xi,x}$ and an observable is $u_n = x_n + \eta_{x,n}$, where the observational noise η_x is Gaussian, white, independent of ξ_x, and with a variance $\sigma^2_{\eta,x}$. The observable u does not completely determine a state of the system X since the latter is specified by the hidden variable x. This situation can again be reduced to the formalism of Sec. III A. If one denotes $x_{1,n} = x_n + \eta_{x,n}$ (an observed variable) and $x_{2,n} = x_n$ (a hidden variable), the system X takes the form (8) with $c_{11} = c_{21} = 0$, $c_{12} = c_{22} = a_x$, $\Sigma_{11} = \sigma^2_{\xi,x} + \sigma^2_{\eta,x}$, and $\Sigma_{12} = \Sigma_{22} = \sigma^2_{\xi,x}$. Similarly to Sec. IV B, the noise covariance in Eq. (8) is nonzero. As the driven system Y, take a relaxation system

$$y_n = a_y y_{n-1} + c x_n + \xi_{y,n}, \tag{19}$$

where the noise ξ_y has a variance $\sigma^2_{\xi,y}$ and an observable is $v_n = y_n + \eta_{y,n}$ with the observational noise η_y Gaussian, white, independent of ξ_y, and possessing a variance $\sigma^2_{\eta,y}$, c is a coupling coefficient. Here ξ_y and η_y are independent of ξ_x and η_x. Analogously, by denoting $y_{1,n} = y_n + \eta_{y,n}$ (an observed variable) and $y_{2,n} = y_n$, one reduces the system Y to the form (10) with $c_{33} = c_{43} = 0$, $c_{34} = c_{44} = a_y$, $c_{31} = c_{41} = 0$, $c_{32} = c_{42} = c$, $\Sigma_{33} = \sigma^2_{\xi,y} + \sigma^2_{\eta,y}$, and $\Sigma_{43} = \Sigma_{44} = \sigma^2_{\xi,y}$. Thus, the full system (18) and (19) is rewritten in the form (8) and (10) and can be studied exactly as in Secs. IV A1 and IV B.

As starting values for the analysis, take $a_x = 0.99$, $\sigma^2_{\xi,x} = 0.1$, $\sigma^2_{\eta,x} = 1$, $a_y = 0.5$, $c = 0.2$, $\sigma^2_{\xi,y} = 0.0001$, and $\sigma^2_{\eta,y} = 0$. Figure 5 shows generically nonzero $T_{Y\to X}$ and the values of r, which can often exceed unity similarly to the examples of Secs. IV A1 and IV B. In particular, the value of the "spurious" $T_{Y\to X}$ is greatest (about 0.1 bits) at big absolute values of a_x (close to unity) and intermediate $\sigma^2_{\xi,x}$ (around 0.1, i.e., by the order of magnitude less than $\sigma^2_{\eta,x} = 1$) and $\sigma^2_{\eta,x}$ (around

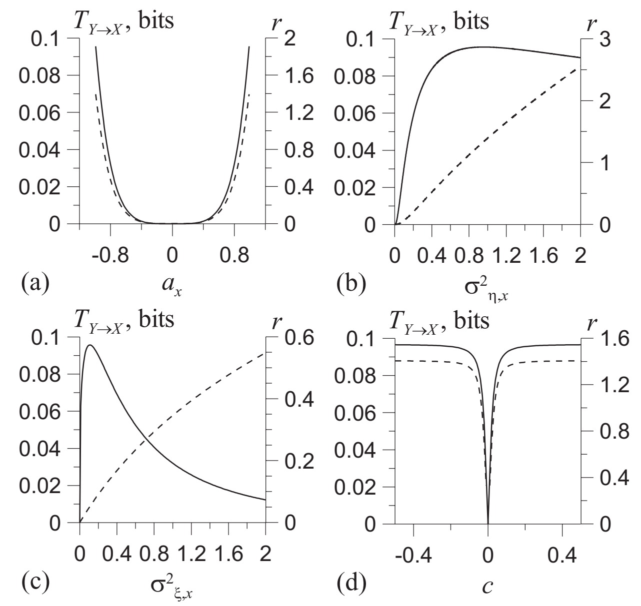

FIG. 5. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the Gaussian process (18) and (19) in the case of observational noise versus parameters: (a) a_x, an individual parameter of the system X controlling autocorrelations of x; (b) $\sigma^2_{\eta,x}$, an observational noise variance in the system X; (c) $\sigma^2_{\xi,x}$, a dynamical noise variance in the system X controlling the amplitude of x; (d) c, the strength of the influence $X \to Y$.

1.0, i.e., by the order of magnitude greater than $\sigma^2_{\xi,x} = 0.1$). It rises with the coupling strength c, but saturates already at $c \approx 0.2$. The character of all those dependencies can be explained via comparison of the contributions of the different terms entering the right-hand side of Eqs. (18) and (19) to the variances of u and v. In particular, small spurious couplings at weak observational noise in the system X [small $\sigma^2_{\eta,x}$ in Fig. 5(b)] are expected because such a case corresponds to accurate observations of the states of X so that the necessary reason for the spurious causalities disappears.

The effect of observational noise on TEs for Markov chains appears quite analogous to the above results as discussed in Appendix B, where TEs are analyzed for Markov chains without such a specific adjustment of their form as in Ref. [40]. Thus, observational noise is shown to be the third factor leading to spurious causalities (similarly to hidden state variables and low temporal resolution) by making observations of the driving system states imperfect.

V. DISCUSSION

Though the above results are presented for two-variable driving systems whose equations can be written in such a form that one of the state variables $u = x_1$ is observed and another one, x_2, is hidden, all qualitative conclusions hold true in a general case where the observables $\mathbf{u}$ and $\mathbf{v}$ are vectors of dimensions d_u and d_v, being arbitrary single-valued functions of the respective state vectors $\mathbf{u} = \mathbf{h}_u(\mathbf{x})$ and $\mathbf{v} = \mathbf{h}_v(\mathbf{y})$. Similarly to the above consideration, nonzero values of the "spurious" $T_{Y\to X}$ are possible if $d_u < d_X$ so that an incomplete state of the driving system X is observed. The spurious coupling effect is impossible if the observation function $\mathbf{h}_u$ is one-to-one, i.e., gives a complete state of X.

To summarize the above results from a physical and a technical point of view, the following main circumstances leading to spurious causalities can be formulated. The first one is more physical and relates to a situation where the driving system X consists of several components with their own degrees of freedom (i.e., X has a relatively complex structure), some of which are unobserved and strongly influence both the observed X components and the system Y. In particular, such a situation might easily arise in studies of various modes of climate variability. Indeed, if one focuses on some basic large-scale modes (e.g., in the Pacific and Atlantic oceans) as is usually done [18,20], their interdependence may well be determined by some neglected processes occurring at other spatial and temporal scales. Additional physical justifications then seem necessary to select appropriate variables for the analysis and make the results reliable. The other two reasons for spurious causalities, insufficient temporal resolution and observational noise, belong to more technical conditions of observations. Whether they are considerable in a concrete practical situation should also be specially analyzed on the basis of entire substantial information about the systems under study and intrinsic time scales of their dynamics as well as about the measurement process.

For an additional mathematical discussion, note that the spurious coupling effect is theoretically possible only for stochastic systems. If X and Y are deterministic, then TE in both directions is zero; more precisely, the conditional distribution of an observable, given its individual past, is a Dirac δ function at a sufficient embedding dimension d as follows from Takens's theorems [46]. The fact that in practice it appears possible to determine coupling direction for deterministic systems based on TE estimation [29,30] takes its roots, seemingly, in TE estimation errors. It may appear simpler sometimes to predict the future of a driven system with a given accuracy based on the past of both systems than only on its own past due to more flexible coping with the "curse of dimensionality" problem (easier approximation of nonlinear predictors under a parametric approach or coarser binning under nonparametric conditional distribution estimation). However, the nontrivial problem of an evolution operator approximation lies apart from the principal circumstance of incomplete state observations and seems, in a sense, more technical.

For an additional statistical consideration, note that the spurious coupling effect may occur in various linear and nonlinear systems, but in practice it may be masked by the TE estimation errors; i.e., a TE estimate can be obtained with so broad confidence interval that a theoretically nonzero but small "spurious" TE would appear statistically indistinguishable from zero. Based on some asymptotic distributions of the TE estimators (see, e.g., Refs. [28,31,47]), one can estimate the length of a time series at which a given small TE becomes detectable at a given significance level. A sufficiently short time series prevents false detections of directional couplings based on nonzero TEs. However, correct detections may appear difficult as well, if the existing coupling is not strong enough.

Large estimation uncertainties do not exclude the risk of false coupling detections in general. In particular, the "spurious" TE may even exceed the "correct" one, strongly distorting a researcher's impression about a coupling character, which may occur in a wide range of situations described above with the examples of oscillators and relaxation systems [48]. To avoid errors, one needs a special test for BC, i.e., against a null hypothesis of UC (indeed, a null hypothesis is always the simpler one [49]). As follows from the above consideration, such a test might be based on checking whether an observed time series is sufficiently likely to be generated by a certain mathematical model with UC or only by a system with BC. To do such a check, one should perform an identification of mathematical models from a certain class, which should be selected on the basis of substantial information about possible physical data-generating mechanisms. Such a test has already been developed to cope with the downsampling problem [38]. A test suitable in a more general situation could be implemented similarly. This idea seems to be a useful practical consequence of the results presented above.

VI. CONCLUSIONS

This work shows that a widely used characteristic of directional couplings, the transfer entropy, may lead to spurious causality inference in a wide range of situations. Namely, in the case of a unidirectional coupling, wrong conclusions about a bidirectional coupling can often be made based on nonzero values of TE in both directions. Within the framework of stochastic systems with finite-dimensional state vectors, it is shown that a common cause of false coupling detections is an incomplete observation of the driving system state, which includes the cases of unobserved state variables, low temporal resolution, and observational noise. For a reliable quantitative analysis of those situations, mathematical benchmark systems (Gaussian processes and Markov chains) are selected here so to provide computation of exact values of TEs, rather than their statistical estimates from simulated time series.

A practical consequence of the obtained results is the necessity to perform special tests for bidirectional coupling if such a conclusion is of importance, which is often the case. An idea behind such a test follows directly from the performed analysis of the different sources of false conclusions and consists in checking whether all appropriate properties of an observed time series can be reproduced by a mathematical model with a unidirectional coupling. However, a practical implementation of such a test may not always appear straightforward and efficient. In particular, it is highly desirable to make the class of trial models as narrow as possible due to the use of prior substantial information about the systems under study.

ACKNOWLEDGMENT

The work is supported by the Russian Foundation for Basic Research and a program of the Russian Academy of Sciences.

APPENDIX A: TRANSFER ENTROPY FOR MARKOV CHAINS IN THE CASE OF DOWNSAMPLING

To get a Markov chain analogous to the oscillators (16) and (17), consider the system X described by a binary variable whose current distribution depends on its two previous values:

$$\begin{aligned}
&\mathbf{P}\{x_n = 0 | x_{n-1} = 0, x_{n-2} = 0\} = 1/2 + p_1 + p_2,\\
&\mathbf{P}\{x_n = 0 | x_{n-1} = 0, x_{n-2} = 1\} = 1/2 + p_1 - p_2,\\
&\mathbf{P}\{x_n = 0 | x_{n-1} = 1, x_{n-2} = 0\} = 1/2 - p_1 + p_2,\\
&\mathbf{P}\{x_n = 0 | x_{n-1} = 1, x_{n-2} = 1\} = 1/2 - p_1 - p_2.
\end{aligned} \tag{A1}$$

An observable is again $u_n = x_{2n}$. By introducing $x_{1,n} = x_{2n}$ and $x_{2,n} = x_{2n-1}$, one gets a chain with two binary state variables and the following transition probabilities:

$$\begin{aligned}
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 0 | x_{1,n-1} = 0, x_{2,n-1} = 0\}\\
&= \mathbf{P}\{x_{2,n} = 0 | x_{1,n-1} = 0, x_{2,n-1} = 0\}\\
&\quad \times \mathbf{P}\{x_{1,n} = 0 | x_{2n} = 0, x_{1,n-1} = 0\}\\
&= (1 + p_1 + p_2)(1 + p_1 + p_2),
\end{aligned} \tag{A2}$$

and similarly for all the other transitions:

$$\begin{aligned}
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 0 | x_{1,n-1} = 0, x_{2,n-1} = 1\}\\
&= (1 + p_1 - p_2)(1 + p_1 + p_2),\\
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 0 | x_{1,n-1} = 1, x_{2,n-1} = 0\}\\
&= (1 - p_1 + p_2)(1 + p_1 - p_2),\\
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 0 | x_{1,n-1} = 1, x_{2,n-1} = 1\}\\
&= (1 - p_1 - p_2)(1 + p_1 - p_2),\\
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 1 | x_{1,n-1} = 0, x_{2,n-1} = 0\}\\
&= (1 - p_1 - p_2)(1 - p_1 + p_2),\\
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 1 | x_{1,n-1} = 0, x_{2,n-1} = 1\}\\
&= (1 - p_1 + p_2)(1 - p_1 + p_2),\\
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 1 | x_{1,n-1} = 1, x_{2,n-1} = 0\}\\
&= (1 + p_1 - p_2)(1 - p_1 - p_2),\\
\mathbf{P}\{x_{1,n} &= 0, x_{2,n} = 1 | x_{1,n-1} = 1, x_{2,n-1} = 1\}\\
&= (1 + p_1 + p_2)(1 - p_1 - p_2),
\end{aligned} \tag{A3}$$

with $P(x_{1,n}, x_{2,n} | x_{1,n-1}, x_{2,n-1}) = P(1 - x_{1,n}, 1 - x_{2,n} | 1 - x_{1,n-1}, 1 - x_{2,n-1})$. Here $x_{1,n}$ and $x_{2,n}$ are no longer conditionally independent, i.e., do not satisfy (13), similarly to the nonzero noise covariance in Sec. IV B.

Let Y be a single-variable relaxation system similar to (11) where the driving changes the transition probabilities of y by the value of c:

$$\begin{aligned}
&\mathbf{P}\{y_n = 0 | y_{n-1} = 0, x_{n-1} = 0\} = 1/2 + p_3 + c,\\
&\mathbf{P}\{y_n = 0 | y_{n-1} = 0, x_{n-1} = 1\} = 1/2 + p_3 - c,\\
&\mathbf{P}\{y_n = 0 | y_{n-1} = 1, x_{n-1} = 0\} = 1/2 - p_3 + c,\\
&\mathbf{P}\{y_n = 0 | y_{n-1} = 1, x_{n-1} = 1\} = 1/2 - p_3 - c.
\end{aligned} \tag{A4}$$

An observable is $v_n = y_{2n}$. Denote a new state variable $y_{1,n} = y_{2n}$. To apply the formalism of Sec. III B, one should find the transition probabilities from the previous state of the three-variable Markov chain to the current one. For $p_3 = 0$ used here similarly to Sec. IV A2, the formulas for those transition probabilities simplify and read

$$\begin{aligned}
&\mathbf{P}(x_{1,n}, x_{2,n}, y_{1,n} | x_{1,n-1}, x_{2,n-1}, y_{1,n-1})\\
&\quad = \mathbf{P}(x_{1,n}, x_{2,n} | x_{1,n-1}, x_{2,n-1})(1/2 \pm c),
\end{aligned} \tag{A5}$$

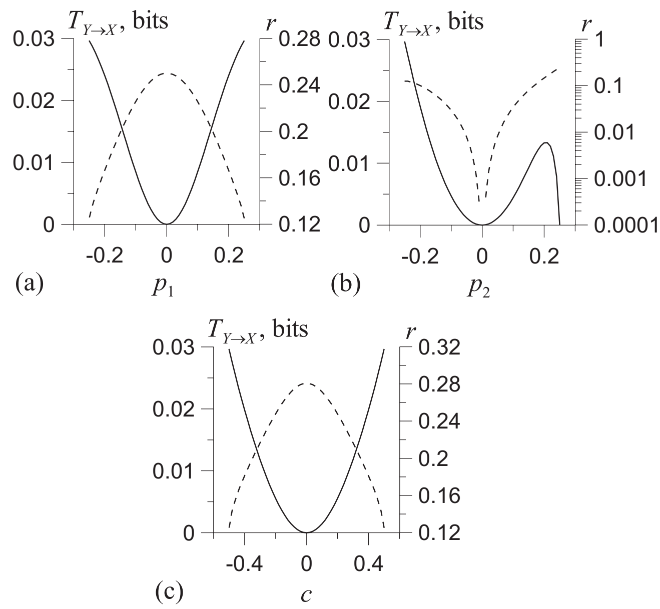

FIG. 6. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the Markov chain (A1) and (A4) in the case of downsampling versus different parameters: (a) and (b) p_1 and p_2, individual parameters of X controlling autocorrelations of u; (c) c, the strength of the influence $X \to Y$.

where $\pm c$ correspond to equal/different values in the pair $(x_{2,n}, y_{1,n})$. Thereby, the full transition probabilities matrix for the Markov chain with three binary variables is specified.

Figure 6 presents the "spurious" TE for this Markov chain with the following starting set of parameters: $p_1 = 1/4$,

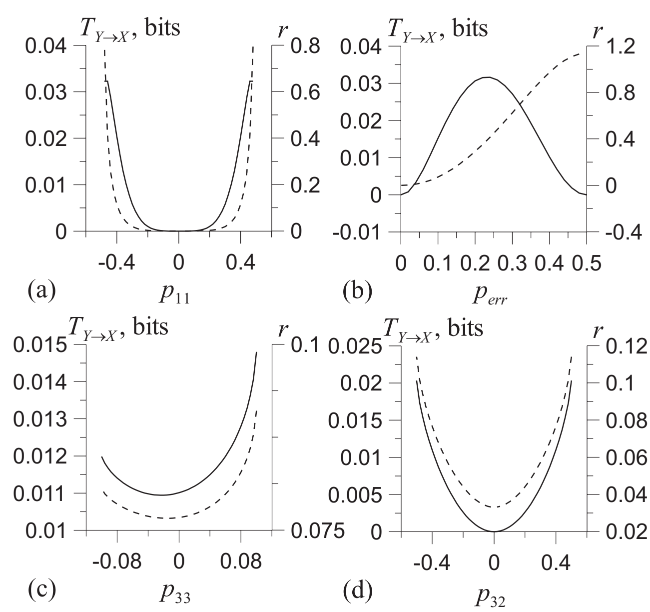

FIG. 7. "Spurious" TE $T_{Y\to X}$ (solid lines) and the relative spurious coupling measure r (dashed lines) for the Markov chains (B1) and (15) in the case of observational noise versus parameters: (a) p_{11}, a "permanence" parameter of the system X controlling autocorrelations of x; (b) p_{err}, an observational noise intensity in the system X; (c) p_{33}, a "permanence" parameter of the system Y controlling autocorrelations of v; (d) p_{32}, the strength of the influence $X \to Y$.

$p_2 = -1/4$ similarly to the negative b_x in Eq. (16), $p_3 = 0$, $c = 1/2$. The results are quite similar to the previous examples in that they evidence generically nonzero $T_{Y\to X}$ and its quite considerable values. Note that, as compared with Fig. 3, this example of an "oscillatory" Markov chain exhibits even greater values of the "spurious" TE.

APPENDIX B: TRANSFER ENTROPY FOR MARKOV CHAINS IN THE CASE OF OBSERVATIONAL NOISE

To get a Markov chain with observation errors, consider a single-variable chain (11) and let an observed value u_n differ from x_n with a fixed probability $p_{\rm err}$. This is a kind of observational noise, where $p_{\rm err}$ characterizes the noise intensity. By introducing $x_{1,n} = u_n$ and $x_{2,n} = x_n$, one gets a Markov chain with four states and transition probabilities

$$P(x_{1,n},x_{2,n}|x_{1,n-1},x_{2,n-1}) = P(x_{2,n}|x_{2,n-1})P(x_{1,n}|x_{2,n}), \tag{B1}$$

where the first multiplier is given by Eq. (11) and the second one is equal to $p_{\rm err}$, if $x_{1,n} \neq x_{2,n}$, and $1 - p_{\rm err}$, otherwise. Let Y be a single-variable chain (15), i.e., to be driven by the variable x_2, whose previous value $x_{2,n-1}$ changes the conditional probabilities of $y_{1,n}$ by the value of p_{32}. Given $y_{1,n-1}$ and $x_{2,n-1}$, the variable $y_{1,n}$ is independent of $x_{1,n}$ and $x_{2,n}$ which allows us to compute the transition probabilities matrix for the full three-variable Markov chain through multiplication of the conditional probabilities. An observable is again $v_n = y_{1,n}$.

Figure 7 presents the results for this Markov chain at a starting set of parameters $p_{11} = 0.4$, $p_{\rm err} = 1/8$, $p_{33} = 0$, $p_{32} = 1/2$ [$p_{32} = 0.4$ only in Fig. 7(c) to provide some freedom in variations of p_{33} which must satisfy $|p_{33}| + |p_{32}| \leqslant 1/2$]. They are similar to Fig. 5 in that they show generically nonzero "spurious" $T_{Y\to X}$, while r can take on rather large values, especially for strong observational noise [$p_{\rm err} \approx 1/2$ in Fig. 7(b)], a strong coupling [$p_{32} \approx 1/2$ in Fig. 7(d)], and a large "permanence" parameter of the driving system X [$p_{11} \approx 1/2$ in Fig. 7(a)]. The value of TE weakly depends on the "permanence" parameter p_{33} of the driven system Y [Fig. 7(c)]. Again, small values of $T_{Y\to X}$ for small $p_{\rm err}$ [Fig. 7(b)] agree with the general expectation that accurate observations of the driving system states prevent the spurious coupling effect.

[1] A. S. Pikovsky, M. G. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, 2001).
[2] S. Boccaletti, J. Kurths, G. Osipov, D. Valladares, and C. Zhou, Phys. Rep. **366**, 1 (2002).
[3] V. S. Anishchenko, V. V. Astakhov, A. B. Neiman, N. E. Vadivasova, and L. Schimanski-Geier, *Nonlinear Dynamics of Chaotic and Stochastic Systems* (Springer, Berlin, 2002).
[4] E. Mosekilde, Yu. Maistrenko, and D. Postnov, *Chaotic Synchronization: Applications to Living Systems* (World Scientific, Singapore, 2002).
[5] G. V. Osipov, J. Kurths, and C. Zhou, *Synchronization in Oscillatory Networks* (Springer, Berlin, 2007).
[6] A. Balanov, N. Janson, D. Postnov, and O. Sosnovtseva, *Synchronization: From Simple to Complex* (Springer, Berlin, 2008).
[7] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, 2000).
[8] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search* (MIT Press, Cambridge, MA, 2001).
[9] B. P. Bezruchko and D. A. Smirnov, *Extracting Knowledge from Time Series: An Introduction to Nonlinear Empirical Modeling* (Springer, Berlin, 2010).
[10] K. J. Friston, L. Harrison, and W. Penny, NeuroImage **19**, 1273 (2003).
[11] E. Pereda, R. Quian Quiroga, and J. Bhattacharya, Progr. Neurobiol. **77**, 1 (2005).
[12] B. Schelter, M. Winterhalder, M. Eichler, M. Peifer, B. Hellwig, B. Guschlbauer, C. Luecking, R. Dahlhaus, and J. Timmer, J. Neurosci. Methods **152**, 210 (2006).
[13] D. A. Smirnov, U. B. Barnikol, T. T. Barnikol, B. P. Bezruchko, C. Hauptmann, C. Buehrle, M. Maarouf, V. Sturm, H.-J. Freund, and P. A. Tass, Europhys. Lett. **83**, 20003 (2008).
[14] M. Staniek and K. Lehnertz, Phys. Rev. Lett. **100**, 158101 (2008).
[15] R. G. Andrzejak and T. Kreuz, Europhys. Lett. **96**, 50012 (2011).
[16] S. Stramaglia, G. R. Wu, M. Pellicoro, and D. Marinazzo, Phys. Rev. E **86**, 066211 (2012).
[17] W. Wang, B. T. Anderson, R. K. Kaufmann, and R. B. Myneni, J. Climate **17**, 4752 (2004).
[18] I. I. Mokhov and D. A. Smirnov, Geophys. Res. Lett. **33**, L03708 (2006).
[19] D. A. Smirnov and I. I. Mokhov, Phys. Rev. E **80**, 016208 (2009).
[20] I. I. Mokhov, D. A. Smirnov, P. V. Nakonechny, S. S. Kozlenko, E. P. Seleznev, and J. Kurths, Geophys. Res. Lett. **38**, L00F04 (2011).
[21] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, Phys. Rev. Lett. **108**, 258701 (2012).
[22] M. Timme, Phys. Rev. Lett. **98**, 224101 (2007).
[23] Z. Levnajic and A. Pikovsky, Phys. Rev. Lett. **107**, 034101 (2011).
[24] C. W. J. Granger, Econometrica **37**, 424 (1969).
[25] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
[26] C. W. J. Granger, J. Econ. Dyn. Control **2**, 329 (1980).
[27] A. Kaiser and T. Schreiber, Physica D **166**, 43 (2002).
[28] K. Hlavackova-Schindler, M. Palus, M. Vejmelka, and J. Bhattacharya, Phys. Rep. **441**, 1 (2007).
[29] I. Vlachos and D. Kugiumtzis, Phys. Rev. E **82**, 016207 (2010).
[30] L. Faes, G. Nollo, and A. Porta, Phys. Rev. E **83**, 051112 (2011).
[31] L. Barnett and T. Bossomaier, Phys. Rev. Lett. **109**, 138105 (2012).
[32] A. Porta, G. Baselli, F. Lombardi, N. Montano, A. Malliani, and S. Cerutti, Biol. Cybern. **81**, 119 (1999).
[33] M. Palus, V. Komarek, Z. Hrncir, and K. Sterbova, Phys. Rev. E **63**, 046211 (2001).
[34] M. Palus and A. Stefanovska, Phys. Rev. E **67**, 055201 (2003).

[35] M. Palus and M. Vejmelka, Phys. Rev. E **75**, 056211 (2007).
[36] M. Vejmelka and M. Palus, Phys. Rev. E **77**, 026214 (2008).
[37] C. A. Sims, Econometrica **39**, 545 (1971).
[38] D. A. Smirnov and B. P. Bezruchko, Europhys. Lett. **100**, 10005 (2012).
[39] H. Nalatore, M. Ding, and G. Rangarajan, Phys. Rev. E **75**, 031123 (2007).
[40] D. W. Hahs and S. D. Pethel, Phys. Rev. Lett. **107**, 128701 (2011).
[41] L. Barnett, A. B. Barrett, and A. K. Seth, Phys. Rev. Lett. **103**, 238701 (2009).
[42] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control* (Holden-Day, San Francisco, 1970).
[43] R. James, C. Ellison, and J. P. Crutchfield, Chaos **21**, 037109 (2011).
[44] N. T. Schmandt and R. F. Galan, Phys. Rev. Lett. **109**, 118101 (2012).
[45] J. Timmer, M. Lauk, W. Pfleger, and G. Deuschl, Biol. Cybern. **78**, 349 (1998).
[46] F. Takens, Lect. Notes Math. **898**, 366 (1981).
[47] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York, 1972), p. 946.
[48] As a simple practical guide, the “spurious” TE may well appear large if correlations between subsequent values of observables are small. At that, missed (unobserved) data may well exhibit nontrivial behavior, which cannot be restored from observations, and determine the system state to a good deal so that their hiddenness could lead to an essential loss of information. Then, spurious detections of directional couplings become more probable. However, relatively “smooth” realizations do not guarantee correct conclusions in practice either, because the smoothness can be the result of an additional “superimposed” filtering rather than a feature of the dynamics of the systems under study. In such a case, TEs in both directions can be small, and their ratio may well misrepresent the coupling directionality.
[49] M. Palus, Contemp. Phys. **48**, 307 (2007).