Partial Mutual Information for Coupling Analysis of Multivariate Time Series

Stefan Frenzel and Bernd Pompe
Institute of Physics, University of Greifswald, D-17487 Greifswald, Germany
(Received 23 May 2007; published 14 November 2007)

We propose a method to discover couplings in multivariate time series, based on *partial mutual information*, an information-theoretic generalization of partial correlation. It represents the part of mutual information of two random quantities that is not contained in a third one. By suitable choice of the latter, we can differentiate between direct and indirect interactions and derive an appropriate graphical model. An efficient estimator for partial mutual information is presented as well.

DOI: 10.1103/PhysRevLett.99.204101 PACS numbers: 05.45.Tp, 02.50.−r, 89.70.+c

The analysis of coupling between two or more systems is of general interest, for instance in the analysis of spatial-temporal systems (lattices of coupled oscillator, earth atmosphere, electrical brain activity *et al.*). If we have no possibility to manipulate these systems, we have to restrict our analysis on some observations (time series). In a somehow naive way coupling could be measured by cross correlation or mutual information (MI), where the latter refers also to nonlinear statistical dependencies. However, this may be misleading in coupling and causality analysis.

Consider, for example, three systems forming a causal chain: The first system should operate autonomously and couple to the second, and the latter to the third system. In this case a pairwise mutual information analysis would yield dependencies also between the first and third system, and we could not decide whether this coupling is made directly or mediated by the second one. A well-known method to overcome this problem is to consider partial correlation [1]. Partial mutual information as proposed here has the same intention, but it is a more general approach because it relates also to nonlinear dependencies, and it needs no explicit modeling. It represents the information between two observations that is not contained in a third one. In this way we can discover the real underlying coupling structure as will be shown afterwards in the example of coupled Lorenz systems.

In general, our method works if the underlying processes have a nonvanishing source entropy (be it stochastic or chaotic), and one process should not be a function of the others, to guarantee that at least a part of source entropy is individual to each process. In practice this is typically fulfilled because of omnipresent dynamical noise.

For a recent attempt in this field we refer to [2], which is based on a phase analysis. Closest to our approach are that of Schreiber [3] and Paluš *et al.* [4], who address couplings of bivariate time series. Our proposal is more closely related to the statistical concept of partialization, and it addresses any multivariate time series. Most important, we also present an efficient estimator, Eq. (6), which makes our concept rather practical.

I. Basic definitions.—For a discrete random variable X, with probabilities $\{p_x\}$ of outcomes $\{x\}$, Shannon entropy is defined by $H(X) = -\sum_x p_x \ln p_x$ (e.g., [5]). As we use a natural logarithm, entropies are measured in units of *nit*. MI of two random variables X and Y is given by $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where $H(X, Y)$ is obtained from the joint distribution $\{p_{xy}\}$ of (X, Y). MI is (i) symmetric, $I(X, Y) = I(Y, X)$, (ii) bounded, $0 \leq I(X, Y) \leq \min\{H(X), H(Y)\}$, where $I(X, Y) = 0$ only if X and Y are independent, and (iii) $I(X, Y) = H(Y)$ only if Y is a function of X.

Now, consider a third variable Z, and take the part of $I(X, Y)$ that is *not* in Z (Fig. 1). We call it *partial mutual information* (PMI) [6],

$$I(X, Y|Z) = H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z). \quad (1)$$

In terms of joint probabilities $\{p_{xyz}\}$ of (X, Y, Z), and the corresponding marginal probabilities $p_{x\cdot z} = \sum_y p_{xyz}$, $p_{\cdot yz} = \sum_x p_{xyz}$, $p_{\cdot\cdot z} = \sum_{xy} p_{xyz}$, we get

$$I(X, Y|Z) = \sum_{xyz} p_{xyz} \ln \frac{p_{xyz} p_{\cdot\cdot z}}{p_{x\cdot z} p_{\cdot yz}}.$$

PMI is symmetric under the same condition Z, $I(X, Y|Z) = I(Y, X|Z)$. We have $0 \leq I(X, Y|Z)$, where zero is obtained only if X and Y are independent under

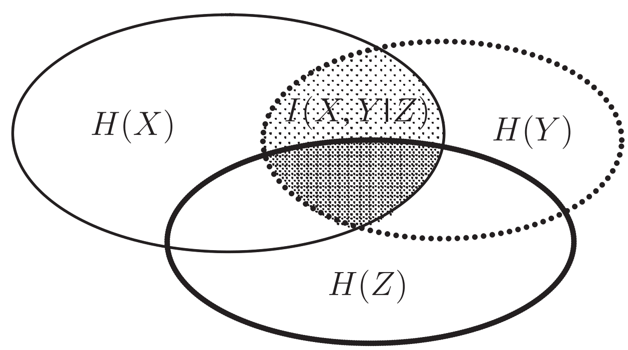

FIG. 1. Partial mutual information $I(X, Y|Z)$ (light shaded) selects the part of mutual information $I(X, Y)$ (all shaded) which is not in Z.

condition Z, i.e., if $(p_{xyz}/p_{\cdot\cdot z}) = (p_{x\cdot z}/p_{\cdot\cdot z})(p_{\cdot yz}/p_{\cdot\cdot z})$ for all x, y, z. This is the case, e.g., if X or Y is a function of Z. We note that it might be that $I(X, Y|Z) > I(X, Y)$ which is not obvious from the diagram Fig. 1. (Partial correlation might also be larger than simple correlation).

In the continuous case, where we have a distribution density $p(x, y, z)$ of (X, Y, Z), with the marginal densities $p(x, \cdot, z)$, $p(\cdot, y, z)$, and $p(\cdot, \cdot, z)$, we get

$$I(X, Y|Z) = \int p(x, y, z) \ln \frac{p(x, y, z)p(\cdot, \cdot, z)}{p(x, \cdot, z)p(\cdot, y, z)} dxdydz. \quad (2)$$

Later on we will consider only continuous random variables. In this case, the term

$$h(X) = -\int p(x) \ln p(x) dx \quad (3)$$

is called differential entropy of any random variable X with density $p(x)$. Then we can write PMI (2) also in terms of differential entropies,

$$I(X, Y|Z) = h(X, Z) + h(Y, Z) - h(Z) - h(X, Y, Z). \quad (4)$$

Continuous PMI is invariant under strictly monotonic transformations which makes it robust against possibly nonlinear distortions of the time series. This property does not hold for the usual correlation.

If the condition Z is irrelevant, i.e., for $p(x, y, z) = p(x, y, \cdot)p(\cdot, \cdot, z)$, PMI (2) is equal to MI,

$$I(X, Y) = h(X) + h(Y) - h(X, Y). \quad (5)$$

II. Estimation.—The derivation of adequate estimators for PMI (2) is a crucial but nontrivial task. For our purposes, the three time series $\{x_t\}$, $\{y_t\}$, and $\{z_t\}$ are considered as finite realizations of underlying stationary ergodic processes $\{X_t\}$, $\{Y_t\}$, and $\{Z_t\}$, respectively. As PMI is a functional of high-dimensional joint probability distributions, we could take the empirical distributions for its estimation, any simple box-counting algorithm [4], or a kernel estimator [3]. However, usually the series have to be extremely long to get good estimates in this way. This causes serious limitations for practical use. An estimator for MI (5) based on kth nearest neighbor statistics has been proposed elsewhere [7]. We now generalize it to an estimator for PMI (2).

For each vector $v_t \equiv (x_t, y_t, z_t)$, $t = 1, 2, \ldots, T$, and a fixed integer k with $1 \leq k \ll T$, we determine the distance $\varepsilon_k(t)$ to the kth nearest neighbor. That means, in $\{v_{t^*}\}$ with $t^* = 1, \ldots, T$, $t^* \neq t$, are just $k - 1$ points at distance strictly less than $\varepsilon_k(t)$ and $N - k - 1$ points at distance strictly larger than $\varepsilon_k(t)$. Here we have to use maximum norm, i.e., $\|\cdot\| = \max\{\|\cdot\|_x, \|\cdot\|_y, \|\cdot\|_z\}$, where $\|\cdot\|_x$, $\|\cdot\|_y$ and $\|\cdot\|_z$ could be any norm, but we use maximum norm as well. Now we switch to the marginal vectors $w_t \equiv (x_t, z_t)$, $t = 1, 2, \ldots, T$, and determine for each t the number of points in $\{w_{t^*}\}$ with distance strictly less than $\varepsilon_k(t)$,

$$N_{xz}(t) = \#\{t^* \neq t: \|w_{t^*} - w_t\| < \varepsilon_k(t)\}.$$

Note that $k \leq N_{xz} + 1$. In the same way we determine $N_{yz}(t)$ and $N_z(t)$ using marginal vectors (y_t, z_t) and z_t, respectively. Now, our PMI estimator is given by

$$\hat{I}(X, Y|Z) = \langle h_{N_{xz}(t)} + h_{N_{yz}(t)} - h_{N_z(t)} \rangle - h_{k-1}, \quad (6)$$

with the negative Nth harmonic number $h_N \equiv -\sum_{n=1}^{N} n^{-1}$, and the time average $\langle \ldots \rangle \equiv \frac{1}{T}\sum_{t=1}^{T} \ldots$ Note the beautiful analogy to (4). If we disregard the condition, that means $Z = \varnothing$, we have $N_z = T - 1$, $N_{xz} = N_x$ and $N_{yz} = N_y$. This yields the estimator for MI, $I(X, Y) = I(X, Y|\varnothing)$, as derived in [7],

$$\hat{I}(X, Y) = \langle h_{N_x(t)} + h_{N_y(t)} \rangle - h_{T-1} - h_{k-1}. \quad (7)$$

A detailed derivation of formula (6) is given in [8]. We have numerical evidences that this PMI estimator is much more efficient than any estimator based on box-counting. We note that in [9] an estimator of differential entropy (3) was proven to be asymptotically unbiased and consistent if (i) the observations are statistically independent, and (ii) some rather general conditions on the density $p(x)$ are fulfilled. Our PMI estimator (6) is based on this differential entropy estimator. Hence it fulfils at least the same properties as each estimator for differential entropy on the right side of (4). However, in practice of time series analysis, these conditions cannot be checked because the densities are unknown.

III. Example: Correlated Gaussian Distribution.—In the following we want to demonstrate the properties of the PMI estimator (6) in the case of Gaussian distributions, where we are able to calculate PMI analytically. For a Gaussian distributed random variable $X \sim \mathcal{N}_d(m, C)$, with the mean values m and covariance matrix C, it applies

$$h(X) = \frac{d}{2}(1 + \ln 2\pi) + \frac{1}{2} \ln \det C. \quad (8)$$

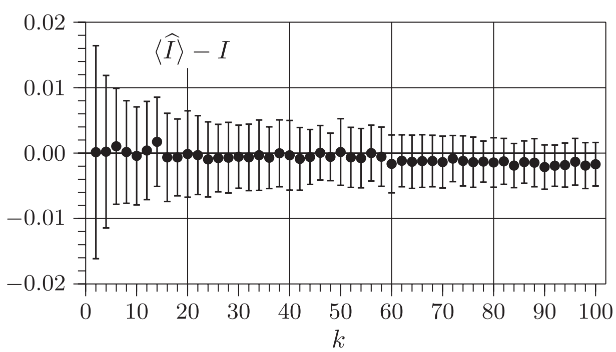

FIG. 2. Bias of the partial mutual information estimator (6) in dependence of parameter k for $T = 1000$ samples of the six-dimensional Gaussian distribution with covariance $c = 0.9$, averaged over 100 trials. The exact value is given analytically by $I = 0.0195, \ldots$, error bars represent the standard deviations of the trials.

We now consider the random variable $(X, Y, Z_3, \dots, Z_d) \sim \mathcal{N}_d(0, C)$ with covariance matrix $C_{ij} = 1$ for $i = j$, $C_{ij} = c$ for $i \neq j$, $|c| < 1$, and $i, j = 1, \dots, d$. The PMI $I \equiv I(X, Y|Z_3, \dots, Z_d)$ can easily be calculated by using Eqs. (4) and (8). We estimate PMI using Eq. (6) for $T = 1000$ samples. Figure 2 shows the corresponding bias $\langle \hat{I} \rangle - I$, averaged over 100 trials for varied k. In general the bias increases with k, while the standard deviation decreases. Hence for $T = 1000$ a well balanced choice would be $k = 20 \dots 30$.

IV. Example: Coupled Lorenz Systems.—In the following we test our method in the case of three coupled Lorenz systems Σ_i: $(\dot{X}_i(t), \dot{Y}_i(t), \dot{Z}_i(t))$, described by the differential equations

$$\dot{X}_i(t) = \sigma(Y_i(t) - X_i(t)),$$
$$\dot{Y}_i(t) = rX_i(t) - Y_i(t) - X_i(t)Z_i(t) + \sum_{j \neq i} K_{ij} Y_j^2(t - \tau_{ij}),$$
$$\dot{Z}_i(t) = X_i(t)Y_i(t) - bZ_i(t), \quad (9)$$

with $i, j = 1, 2, 3$. We integrated these equations numerically, applying a fourth order Runge-Kutta method with integration step 0.003, but we recorded only every 100th point leading to time step $\Delta t = 0.3$. We cut away transient dynamics at the beginning, and take standard parameters $\sigma = 10$, $r = 28$, $b = 8/3$. In the uncoupled case $K_{ij} = 0$, the systems are autonomic and perform chaotic motions near the well-known Lorenz attractor. Coupling is realized via the quadratic Y_j components at delays τ_{ij}, controlled via K_{ij}. Here we consider the case of a causal chain: $K_{12} = K_{23} = 1$ and $K_{ij} = 0$ otherwise. The corresponding delays are set to $\tau_{12} = 10$ and $\tau_{23} = 15$ time steps. We consider the three time series $\{y_{i,t}\} \equiv \{Y_i(t \cdot \Delta t)\}$, $t = 1, 2, \dots, T = 1000$. However, for our following analysis it is only important to have at least one series of each subsystem Σ_i. As the autonomic systems are chaotic and numerical round off errors act like dynamical noise, we can suppose that at least a part of source entropy is individual to each series.

The question now is, can we derive the complete coupling structure from an analysis of the corresponding time series alone? We start our analysis with the pairwise cross MI functions,

$$I_{ij}(\tau) \equiv I(Y_{i,t}, Y_{j,t+\tau}), \qquad i \neq j, \quad (10)$$

applying the estimator (7) for $k = 20$ (see comment [10]). They suggest some very significant dependencies with delays $\tau_{12} \approx 10$, $\tau_{13} \approx 27$, and $\tau_{23} \approx 15$ [Figs. 3(a)–3(c)], respectively. These delays are all positive. Our idea of causality is that there is first the cause and then the effect. Hence, from this analysis we could conclude that there are couplings $\Sigma_1 \to \Sigma_2$, $\Sigma_1 \to \Sigma_3$, and $\Sigma_2 \to \Sigma_3$. There might be also a weak backward coupling $\Sigma_2 \to \Sigma_1$, because there are also significant dependencies for some negative τ [Fig. 3(a)]. Graphical models which are consistent by pairwise MI analysis are summarized in Fig. 4. Auto dependencies in the series lead to a broadening of the peaks in the cross MI functions, which indicate dependencies also for negative τ as in the case of Fig. 3(a). Having in mind that dependencies between uncoupled systems can be, e.g., due to a common driver system, we cannot surely conclude the true global structure, Fig. 4(d), of coupling from simple pairwise MI analysis.

In order to derive the true one we will now *partialize out* the influence of the third system when analyzing the dependencies between the others. Instead of cross MI function $I_{ij}(\tau)$ we ask now for the MI of $Y_{i,t}$ and $Y_{j,t+\tau}$ which is not in $\{\dots, Y_{l,t-1}, Y_{l,t}, Y_{l,t+1}, \dots\}$, where $i, j, l = 1, 2, 3$, all different. This is our PMI (2), for condition Z being ∞-dimensional. In practice, we have to restrict to a finite dimensional delay embedding vector

$$Y_{l,t+\vartheta_l} \equiv (Y_{l,t+\vartheta_{l,1}}, Y_{l,t+\vartheta_{l,2}}, \dots, Y_{l,t+\vartheta_{l,D}}),$$

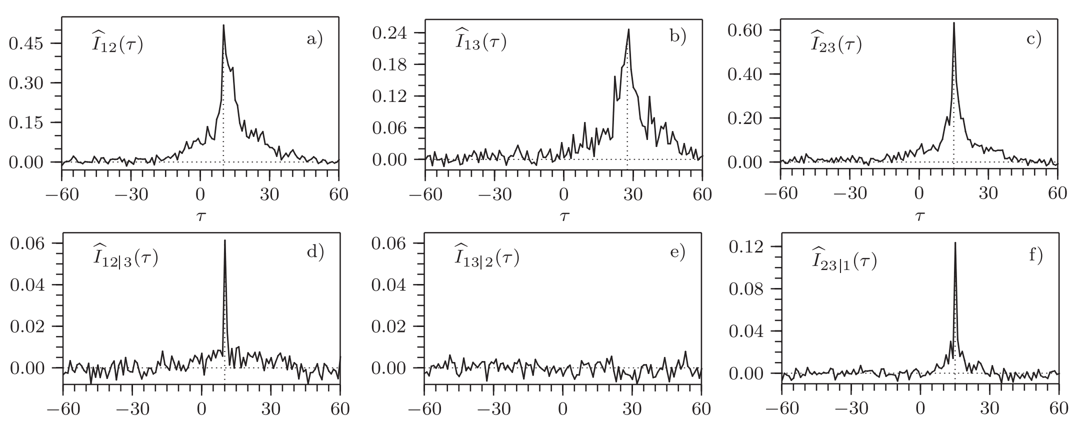

FIG. 3. From pairwise mutual information we would conclude the Lorenz systems to be pairwise unidirectionally coupled with delays $\tau_{12} \approx 10$, $\tau_{13} \approx 27$ and $\tau_{23} \approx 15$. Indeed the interdependence between Σ_1 and Σ_3 is indirect due to they are both coupled to Σ_2. This can be seen from the vanishing partial mutual information function $\hat{I}_{13|2}(\tau)$.

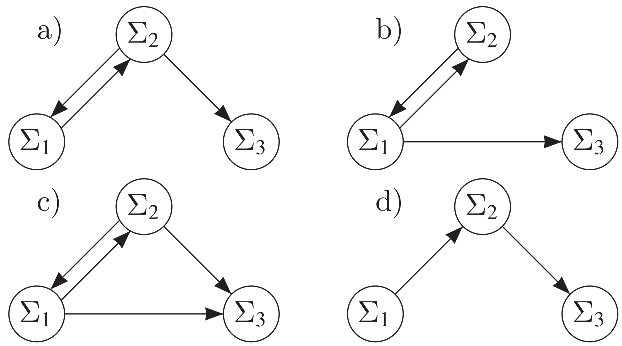

FIG. 4. Coupling structures which are consistent with the pairwise MI analysis of the Lorenz chain. The true structure (d) can be derived by considering partial mutual information.

with time comb $\vartheta_l = (\vartheta_{l,1}, \ldots, \vartheta_{l,D})$. We therefore consider PMI functions

$$I_{ij|l}(\tau) \equiv I(Y_{i,t}, Y_{j,t+\tau}|Y_{l,t+\vartheta_l}). \tag{11}$$

The comb ϑ_l is chosen such that (i) $Y_{l,t+\vartheta_l}$ is maximally statistically related to $Y_{i,t}$, and (ii) the components of $Y_{l,t+\vartheta_l}$ are minimally redundant. A relevant comb fulfilling the first condition is given by the τ values where the MI function $I_{il}(\tau)$ significantly differs from zero.

We start with considering the MI function between Σ_1 and Σ_2, $I_{12}(\tau)$ [Fig. 3(a)]. It shows significant dependencies at $\tau_{12} \approx 10$. From $I_{13}(\tau)$ [Fig. 3(b)] we read off dependencies at $\tau_{13} \approx 27$. Hence we choose $\vartheta_3 = (24, 25, \ldots, 30)$, symmetrically around τ_{13}. From the resulting PMI function $I_{12|3}(\tau)$ [Fig. 3(d)] we now clearly detect a unidirectional coupling $\Sigma_1 \rightarrow \Sigma_2$ with a sharp peak at delay $\tau_{12} = 10$. So there cannot be an edge from node Σ_2 to Σ_1 in the corresponding graphical model [Figs. 4(a)–4(c)].

In a similar way we analyze the coupling between Σ_2 and Σ_3. For PMI $I_{23|1}(\tau)$ [Fig. 3(f)] we take time comb $\vartheta_1 = (-7, \ldots, -13)$ read from [Fig. 3(a)], having in mind $I_{21}(\tau) = I_{12}(-\tau)$. Thus we detect a unidirectional coupling $\Sigma_2 \rightarrow \Sigma_3$ from the sharp peak of $I_{23|1}(\tau)$ at delay $\tau_{23} = 15$. So there must be an edge between the corresponding nodes; i.e., graph (b) in Fig. 4 must be wrong.

Finally we analyze the coupling between Σ_1 and Σ_3. For PMI $I_{13|2}(\tau)$ [Fig. 3(e)] we take time comb $\vartheta_2 = (7, \ldots, 13)$ read from [Fig. 3(a)]. $I_{13|2}(\tau)$ vanishes in contrast to $I_{13}(\tau)$ [Fig. 3(b)]. This shows, that there is no direct interaction Σ_1 to Σ_3. The dependence between these systems shown in Fig. 3(b) must be due to *indirect* interaction, mediated by Σ_2. So there is no edge between the corresponding nodes; i.e., the graph Fig. 4(c) must be wrong as well.

Summarizing all these results we thus have detected the right graphical model for coupling: it is Fig. 4(d). We note that linear correlation analysis does not show any significant dependencies between systems Σ_1 and Σ_2 as well as between Σ_1 and Σ_3, leading to wrong conclusions about coupling of the underlying systems. To overcome this problem, specific higher moment correlations could be considered. However, MI analysis performs this at once, without any model assumptions.

V. Final remarks.—We have introduced partial mutual information (4) and an efficient estimator (6) for the analysis of couplings between systems. Thus we applied the concept of partialization to mutual information analysis of multivariate time series. Next steps would be applications to real data.

The transfer entropy introduced in [3] can be considered as a PMI for detecting couplings between two time series: It measures the MI of the present of a process and the past of another one, which is not contained in the past of the first one. Also in this case of bivariate time series the analysis would profit (i) from a nontrivial choice of time combs, especially if the systems operate on different time scales, and (ii) from the usage of the estimator (6).

We acknowledge useful comments of C. Bandt, as well as the support of DFG Project No. SFB-TR 24, B8.

[1] H. Blalock, *Causal Inferences in Nonexperimental Research* (University of North Carolina, Chapel Hill, NC, 1961).

[2] B. Schelter, M. Winterhalder, R. Dahlhaus, J. Kurths, and J. Timmer, Phys. Rev. Lett. **96**, 208103 (2006).

[3] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[4] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, Phys. Rev. E **63**, 046211 (2001).

[5] R. W. Hamming, *Coding and Information Theory* (Prentice-Hall, Englewood Cliffs, NJ, 1986).

[6] What we call *partial* is sometimes called *conditional* MI (e.g.: J. Rissanen, Lectures on Statistical Modeling Theory, IBM Almaden Research Center, DPE-B2/802). However, in [5], conditional MI denotes a different quantity. We prefer *partial* as it best matches our interpretation, and it is analogue to partial correlation.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).

[8] S. Frenzel, Diploma thesis, University of Greifswald, 2007.

[9] L. F. Kozachenko and N. N. Leonenko, Probl. Inf. Transm. **23**, 95 (1985).

[10] We note here that the invariant measure of an ergodic dynamical system is often not absolutely continuous with respect to Lebesgue measure, especially not for our (coupled) Lorenz systems. Hence there is no probability density and MI is infinite. Nevertheless, if we have a finite length of time series, $T < \infty$, then k/T plays the role of a coarse graining parameter, and the magnitude of the estimated MI resp. PMI depends on this quotient. Thus we should be very carefully when interpreting the estimated magnitude of MI, we rather look at the relative magnitude for varying delay τ.