Measuring Information Transfer

Thomas Schreiber
Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Strasse 38, 01187 Dresden, Germany

(Received 19 January 2000)

An information theoretic measure is derived that quantifies the statistical coherence between systems evolving in time. The standard time delayed mutual information fails to distinguish information that is actually exchanged from shared information due to common history and input signals. In our new approach, these influences are excluded by appropriate conditioning of transition probabilities. The resulting *transfer entropy* is able to distinguish effectively driving and responding elements and to detect asymmetry in the interaction of subsystems.

PACS numbers: 89.70.+c, 05.20.Gg, 05.45.Tp

The time evolution of a system may be called irregular if it generates information at a nonzero rate. For stochastic or deterministically chaotic systems, this is quantified by the entropy. For a system consisting of more than one component, important information on its structure can be obtained by measuring to which extent the individual components contribute to information production and at what rate they exchange information among each other. This paper proposes a method to answer the latter question on the basis of time series observations.

Many authors have used *mutual information* [1] to quantify the overlap of the information content of two (sub)systems. Unfortunately, mutual information neither contains dynamical nor directional information. Introducing a time delay in one of the observations is an important, if somewhat arbitrary, improvement in this respect, but still does not explicitly distinguish information that is actually exchanged from that due to the response to a common input signal or history.

The purpose of this paper is to motivate and derive an alternative information theoretic measure, to be called *transfer entropy,* that shares some of the desired properties of mutual information but takes the dynamics of information transport into account. With minimal assumptions about the dynamics of the system and the nature of their coupling one will be able to quantify the exchange of information between two systems, separately for both directions, and, if desired, conditional to common input signals.

This work augments recent studies [2] of the nonlinear coherence of signals, most notably in physiological systems. While these measures are often very powerful for a specific set of applications, it is also important to aim at an understanding of the underlying theoretical concepts. In the generic case that neither one of the systems, nor their coupling, may be assumed to be deterministic, information theory seems to be an appropriate starting point.

Let us briefly recall the most basic concepts of information theory [3]. The average number of bits needed to optimally encode independent draws of the discrete variable I following a probability distribution $p(i)$ is given by the Shannon entropy [1] $H_I = -\sum_i p(i)\log_2 p(i)$, where the sum extends over all states i the process can assume. The base of the logarithm determines only the units used for measuring information and will be dropped henceforth.

In order to construct an optimal encoding that uses just as many bits as given by the entropy, it is necessary to know the probability distribution $p(i)$. The excess number of bits that will be coded if a different distribution $q(i)$ is used is given by the Kullback entropy [4] $K_I = \sum_i p(i)\log p(i)/q(i)$. We will later also need the Kullback entropy for conditional probabilities $p(i \mid j)$. For a single state j we have $K_j = \sum_i p(i \mid j)\log p(i \mid j)/q(i \mid j)$. Summation over j with respect to $p(j)$ yields

$$K_{I \mid J} = \sum_{i,j} p(i,j)\log\frac{p(i \mid j)}{q(i \mid j)} . \tag{1}$$

The *mutual information* of two processes I and J with joint probability $p_{IJ}(i,j)$ can be seen as the excess amount of code produced by erroneously assuming that the two systems are independent, i.e., using $q_{IJ}(i,j) = p_I(i)p_J(j)$ instead of $p_{IJ}(i,j)$. The corresponding Kullback entropy is

$$M_{IJ} = \sum p(i,j)\log\frac{p(i,j)}{p(i)p(j)} , \tag{2}$$

which is the well known formula for the mutual information. Here and in the following, we omitted the summation index and the subscript of the probabilities specifying the process. This derivation shows that mutual information is a natural way to quantify the deviation from independence of two processes. We have $M_{IJ} = H_I + H_J - H_{IJ} \geq 0$. Note that M_{IJ} is symmetric under the exchange of I and J and therefore does not contain any directional sense.

A related, nonsymmetric quantity is the conditional entropy $H_{I \mid J} = -\sum p(i,j)\log p(i \mid j) = H_{IJ} - H_J$. However, since $H_{I \mid J} - H_{J \mid I} = H_I - H_J$, it is nonsymmetric only due to the different individual entropies and not due to information flow. Mutual information can be given a directional sense in a somewhat *ad hoc* way by introducing a time lag in either one of the variables and compute, e.g.,

$$M_{IJ}(\tau) = \sum p(i_n, j_{n-\tau})\log\frac{p(i_n, j_{n-\tau})}{p(i)p(j)} .$$

© 2000 The American Physical Society

As we will see below, considering the two systems at different times occurs naturally as soon as transition probabilities are introduced.

One can incorporate dynamical structure by studying transition probabilities rather than static probabilities. Consider a system that may be approximated by a stationary Markov process of order k, that is, the conditional probability to find I in state i_{n+1} at time $n + 1$ is independent of the state i_{n-k}: $p(i_{n+1} \,|\, i_n, \dots, i_{n-k+1}) = p(i_{n+1} \,|\, i_n, \dots, i_{n-k+1}, i_{n-k})$. Henceforth we will use the shorthand notation $i_n^{(k)} = (i_n, \dots, i_{n-k+1})$ for words of length k.

The average number of bits needed to encode one additional state of the system if all previous states are known is given by the *entropy rate*

$$h_I = -\sum p(i_{n+1}, i_n^{(k)}) \log p(i_{n+1} \,|\, i_n^{(k)}) \,. \tag{3}$$

Since $p(i_{n+1} \,|\, i_n^{(k)}) = p(i_{n+1}^{(k+1)})/p(i_n^{(k)})$, this is just the difference between the Shannon entropies of the processes given by $k + 1$ and k dimensional delay vectors [5] constructed from I: $h_I = H_{I^{(k+1)}} - H_{I^{(k)}}$.

If I is obtained by coarse graining a continuous system X at resolution r, the entropy $H_X(r)$ and entropy rate $h_X(r)$ will depend on the partitioning and in general diverge like $-\log r$ when $r \to 0$. However, for the special case of a deterministic dynamical system, $\lim_{r\to 0} h_X(r) = h_{\mathrm{KS}}$ may exist and is then called the Kolmogorov-Sinai entropy [6]. (For non-Markov systems, also the limit $k \to \infty$ needs to be taken.) Confusingly, the opposite is true for the mutual information: For generic noisy interdependence, $\lim_{r\to 0} M_{XY}(r)$ is finite and independent of the partition, but for deterministically coupled processes, $M_{XY}(r)$ will diverge as $r \to 0$.

For the study of the dynamics of shared information between processes it is desirable to generalize the entropy rate, rather than Shannon entropy, to more than one system, since the dynamics of the processes is contained in the transition probabilities. The most straightforward way to construct a mutual information rate by generalizing h_I to two processes (I, J) is again by measuring the deviation from independence. The corresponding Kullback entropy is still symmetric under the exchange of I and J. It is therefore preferable to measuring the deviation from the generalized Markov property,

$$p(i_{n+1} \,|\, i_n^{(k)}) = p(i_{n+1} \,|\, i_n^{(k)}, j_n^{(l)}) \,.$$

In the absence of information flow from J to I, the state of J has no influence on the transition probabilities on system I. The incorrectness of this assumption can again be quantified by a Kullback entropy (1) by which we define the *transfer entropy*:

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{p(i_{n+1} \,|\, i_n^{(k)}, j_n^{(l)})}{p(i_{n+1} \,|\, i_n^{(k)})} \,. \tag{4}$$

This is the central concept of this paper. The most natural choices for l are $l = k$ or $l = 1$. Usually, the latter is preferable for computational reasons. $T_{J\to I}$ is now explicitly nonsymmetric since it measures the degree of dependence of I on J and not vice versa. Similar quantities have been discussed outside a dynamical framework; see, e.g., the section on *conditional transinformation* in Ref. [7], or the discussion in Ref. [8].

For coarse grained states (I, J) of continuous systems (X, Y), the limit $\lim_{r\to 0} T_{Y\to X}(r)$ is finite and independent of the partition, except for the case of deterministic coupling, when $T_{Y\to X}(r)$ diverges as $r \to 0$. In this respect, transfer entropy behaves like mutual information. If computationally feasible, the influence of a known common driving force Z may be excluded by conditioning the probabilities under the logarithm to z_n as well.

For practical applications, the limit $r \to 0$ is not obtainable and has to be replaced appropriately. Either one can study transfer entropy as a function of the resolution or one can fix a resolution for the scope of a study. Furthermore, there are several methods of coarse graining. A partition consisting of a fixed mesh of boxes is suitable only when data can be produced with little effort.

For time series applications, an alternative implementation using generalized correlation integrals is preferable. Mutual information and redundancies have been generalized for their estimation by order q correlation integrals [9]. It is possible to follow the same arguments in generalizing transfer entropy. However, for the computationally most attractive case $q = 2$, we would have to give up positivity of $T_{I\to J}$. Instead, we propose an implementation of the definition (4) where the probability measure $p(i_{n+1}, i_n^{(k)}, j_n^{(l)})$ is realized by a sum over all available realizations of $(x_{n+1}, x_n^{(k)}, y_n^{(l)})$ in a time series. The transition probabilities are expressed by joint probabilities and then obtained by kernel estimation, e.g.,

$$\hat{p}_r(x_{n+1}, x_n, y_n) = \frac{1}{N} \sum_{n'} \Theta\left(\left| \begin{pmatrix} x_{n+1} - x_{n'+1} \\ x_n - x_{n'} \\ y_n - y_{n'} \end{pmatrix} \right| - r \right).$$

We use the step kernel $\Theta(x > 0) = 1$; $\Theta(x \le 0) = 0$. The norm $|\cdot|$ can be simply the maximum distance but other norms and kernels can be considered. In particular, different overall scales of X and Y can be accounted for by using appropriate weights. Similar to standard dimension and entropy calculations, fast neighbor search strategies are advisable for all but the smallest data sets. Dynamically correlated pairs should be excluded as usual. Since these technical issues are the same as in many nonlinear time series methods, the reader is referred to the discussion in the literature [5].

In order to demonstrate the use of transfer entropy, let us study three examples, two spatiotemporal systems and a bivariate physiological time series. In a one dimensional lattice of unidirectionally coupled maps

$$x_{n+1}^m = f(\epsilon x_n^{m-1} + (1 - \epsilon) x_n^m) \,, \tag{5}$$

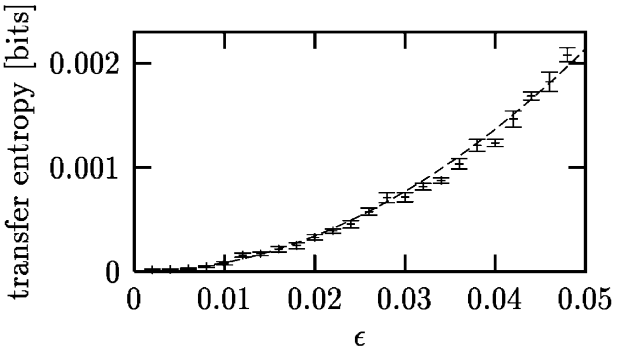

FIG. 1. Transfer entropy $T_{I^{m-1}\to I^m}$ as a function of the coupling strength ϵ in a tent map lattice (binary partition). Error bars: error of the mean of 10 runs of 100 000 iterates. Line: theoretical curve $\alpha^2\epsilon^2/\ln(2)$ with fitted $\alpha = 0.77$.

information can be transported only in the direction of increasing m. One of the simplest cases is given by the tent map, $f(x < 0.5) = 2x$; $f(x \geq 0.5) = 2 - 2x$. Let us study coarse grained states I^m with i_n^m defined by a partition at $x_0 = 0.5$. At zero coupling, all static and transfer probabilities are equal to $1/2$, $M(\tau) = 0$ for all values of τ, and also $T_{I^{m-1}\to I^m} = T_{I^m\to I^{m-1}} = 0$. For nonzero coupling, we still have $T_{I^m\to I^{m-1}} = 0$, but $T_{I^{m-1}\to I^m}$ becomes positive. For small coupling, it can be assumed that the invariant density at a single site is essentially unchanged whence the transition probabilities $p(I_{n+1}^m \mid I_n^m, I_n^{m-1})$ are changed by an amount proportional to ϵ. In particular, $p(0\mid 0,0)$, $p(0\mid 1,1)$, $p(1\mid 0,1)$, and $p(1\mid 1,0)$ are increased by a factor of $1 + \alpha\epsilon$ with $\alpha = O(1)$. All others are decreased by that amount. Evaluating (4) in lowest order of ϵ with $k = l = 1$, we obtain $T_{I^{m-1}\to I^m} = \alpha^2\epsilon^2/\ln(2) + O(\epsilon^4)$. For this particular case, the changes in $p(i_{n+1}^m, i_n^{m-1})$ exactly cancel out and the mutual information is zero. Figure 1 shows a numerical verification of these results for a spatially periodic lattice of 100 maps. Averages of 10 runs of 10^5 iterates after 10^5 transients are shown. The transfer entropy $T_{I^m\to I^{m-1}}$ and both directions of $M(\tau = 1)$ were found to be consistent with zero and are therefore not shown.

The situation is more complicated for the Ulam map $f(x) = 2 - x^2$ and nonsmall coupling. For each coupling, a bivariate time series was generated using a lattice of 100 points (random initial conditions) and recording 10 000 iterates of x_n^1 and x_n^2 after 10^5 steps of transients. Correlation sums at $r = 0.2$ were used to compute mutual information in both directions, $M_{X^1,X^2}(\tau = 1)$ and $M_{X^2,X^1}(\tau = 1)$, as well as transfer entropies $T_{X^1\to X^2}$ and $T_{X^2\to X^1}$ with $k = l = 1$. Neighbors closer in time than 100 iterates were excluded from the kernel estimation.

Figure 2 shows M and T as functions of the coupling strength. Both M and T are able to detect the anisotropy since the information is consistently larger in the positive direction. The lattice undergoes a number of bifurcations when the coupling is changed. Around $\epsilon = 0.18$, the asymptotic state is of temporal and spatial period two. For this case, the mutual information is found to be 1 bit. This is correct although information is neither produced nor exchanged and reflects the static correlation between the sites. The transfer entropy finds a zero rate of information transport, as desired. Around this periodic window, the mutual information is nonzero in both directions and the signature of the unidirectional coupling is less pronounced. Around $\epsilon = 0.82$, the lattice settles to a (spatially inhomogeneous) fixed point state. Here both measures correctly show zero information transfer. The most important finding, however, is that the transfer entropy for the negative direction remains consistent with zero for all couplings, reflecting the causality in the system.

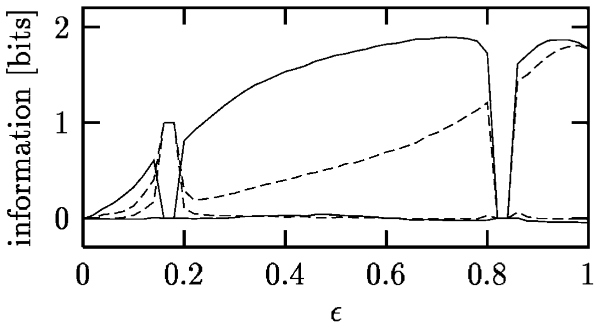

FIG. 2. Transfer entropies $T_{X^1\to X^2}$ and $T_{X^2\to X^1}$ (solid lines) and time delayed mutual information $M_{X^1,X^2}(\tau = 1)$ and $M_{X^2,X^1}(\tau = 1)$ (dashed lines) as functions of the coupling strength ϵ for a unidirectionally coupled Ulam lattice. For both quantities, the upper line denotes the direction $X^{m-1} \to X^m$ while the lower line shows $X^{m+1} \to X^m$. Although the lattice undergoes a sequence of bifurcations, the transfer entropy T clearly reflects the unidirectional character of the coupling. It also consistently outperforms the time delayed mutual information in this respect. See text for further details.

As a last example, take a bivariate time series (see Fig. 3) of the breath rate and instantaneous heart rate of a sleeping human suffering from sleep apnea (samples 2350–3550 of data set B of the Santa Fe Institute time series contest held in 1991 [10]). Figure 4

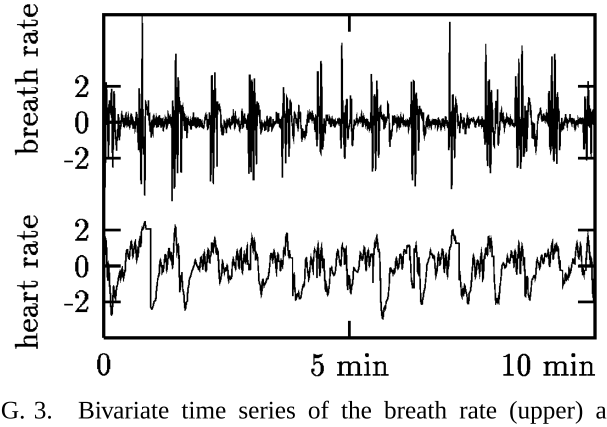

FIG. 3. Bivariate time series of the breath rate (upper) and instantaneous heart rate (lower) of a sleeping human. The data is sampled at 2 Hz. Both traces have been normalized to zero mean and unit variance.

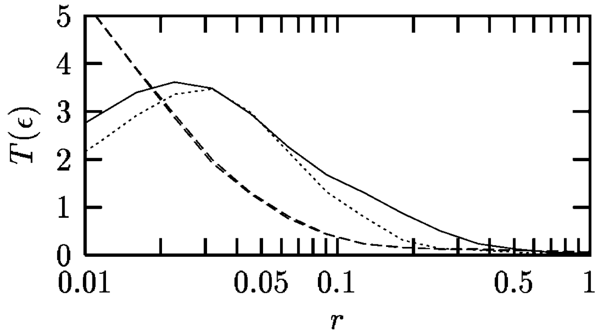

FIG. 4. Transfer entropies T(heart → breath) (solid line), T(breath → heart) (dotted line), and time delayed mutual information $M(\tau = 0.5\ \mathrm{s})$ (directions indistinguishable, dashed line) for the physiological time series shown in Fig. 3.

shows that time delayed mutual information is almost symmetric between both series. The transfer entropy also finds information transport in both directions but indicates a stronger flow of information from the heart rate to the breath rate than vice versa over a significant range of length scales r. Note that for small r the curves deflect down to zero due to the finite sample size.

In contrast to the previous examples, the two channels studied here differ in their individual information contents. In such a situation, unless we find zero information transfer in one of the directions, too rash conclusions about the nature of the interaction have to be avoided. Different rates of information production and transport between length scales will naturally cause some asymmetry in the rate of information transfer, as measured by T. Reducing the analysis to the identification of a "drive" and a "response" may not be useful and could even be misleading. In this particular data set, the dominant direction of information flow from the heart to the breath signal is consistent with the observation that the patient breathes in bursts which seem to occur whenever the heart rate crosses some threshold. Finally, note that the findings could also be explained with a coupling of both signals to a common external trigger.

In conclusion, the new *transfer entropy* is able to detect the directed exchange of information between two systems. Unlike mutual information, it is designed to ignore static correlations due to the common history or common input signals. Most prominent applications include multivariate analysis of time series and the study of spatially extended systems.

Several authors [11] have proposed to use time delayed mutual information $M(\Delta l, \tau)$ as a function of spatial distance Δl and temporal delay τ to define a velocity of information transport in spatiotemporal systems. Often, one finds that $M(\Delta l, \tau)$ for fixed Δi reaches a local maximum at some lag τ^*. Hence a velocity can be defined by the ratio $\Delta i/\tau^*$, in particular if that ratio is fairly constant over the resolvable range of values for Δi. This reasoning has been challenged [12] by giving an example where the above interpretation implies superluminal communication. In fact, much of the common information is due to the common history that allows the lattice to partially synchronize. Preliminary results indicate that appropriate conditioning for the common history by replacing time delayed mutual information by a variant of Eq. (4) resolves this apparent paradox. However, conditioning with respect to a large number of variables poses immense numerical problems whence this study will be concluded at a later time.

Part of this work has been supported by the SFB 237 of the Deutsche Forschungsgemeinschaft.

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Information* (University of Illinois Press, Urbana, IL, 1949).

[2] J. Arnhold, P. Grassberger, K. Lehnertz, and C. E. Elger, Physica (Amsterdam) **134D**, 419 (1999); M. G. Rosenblum, A. S. Pikovsky, and J. Kurths, Phys. Rev. Lett. **76**, 1804 (1996); M. Le Van Quyen, C. Adam, M. Baulac, J. Martinerie, and F. J. Varela, Brain Res. **792**, 24 (1998).

[3] P. Billingsley, *Ergodic Theory and Information* (Wiley, New York, 1965).

[4] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[5] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, MA, 1997).

[6] A. N. Kolmogorov, *Information Theory and The Theory of Algorithms* (Kluwer, Dordrecht, 1993), selected works, Vol. 3.

[7] G. Jumarie, *Relative Information: Theories and Applications*, Springer Series in Synergetics Vol. 47 (Springer, Berlin, 1990).

[8] M. Palus, Phys. Lett. A **213**, 138 (1996).

[9] P. Grassberger, T. Schreiber, and C. Schaffrath, Int. J. Bifurcation Chaos Appl. Sci. Eng. **1**, 521 (1991); B. Pompe, J. Stat. Phys. **73**, 587 (1993); D. Prichard and J. Theiler, Physica (Amsterdam) **84D**, 476 (1995).

[10] D. R. Rigney, A. L. Goldberger, W. Ocasio, Y. Ichimaru, G. B. Moody, and R. Mark, in *Time Series Prediction: Forecasting the Future and Understanding the Past*, edited by A. S. Weigend and N. A. Gershenfeld (Addison-Wesley, Reading, MA, 1993).

[11] K. Kaneko, Physica (Amsterdam) **23D**, 436 (1986); J. A. Vastano and H. L. Swinney, Phys. Rev. Lett. **60**, 1773 (1988).

[12] T. Schreiber, J. Phys. A **23**, L393 (1990).