ELSEVIER

Physica D 166 (2002) 43–62

PHYSICA D

www.elsevier.com/locate/physd

Information transfer in continuous processes

A. Kaiser*, T. Schreiber

Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38, D-01187 Dresden, Germany

Received 13 November 2001; accepted 14 March 2002
Communicated by H. Müller-Krumbhaar

Abstract

We discuss a recently proposed quantity, called transfer entropy, which uses time series data to measure the amount of information transferred from one process to another. In order to understand its foundation, merits, and limitations, we review some aspects of information theoretic functionals. While for symbol sequences these measures have an intuitive interpretation, their application to continuous state processes and, in particular, their estimation from finite data sets is problematic. For mutual information, finite length scale estimates converge from below and can thus be used to reject the assumption that the observed processes are independent. However, mutual information does not provide any directional information. Conversely, transfer entropy does resolve the directionality of information exchange but no similar monotonic convergence seems to hold. Thus, only in the case of zero transfer entropy in one direction we can reliably infer an asymmetry of the information exchange.

PACS: 05.45.+b; 02.50.Wp; 89.70.+c; 07.05.Kf

Keywords: Information transfer; Mutual information; Information theory; Non-parametric estimation; Stochastic dependence

1. Introduction

Much recent work [1–4] is devoted to the problem of measuring from time series recordings the strength and direction of the coupling between simultaneously observed systems. This is a question that arises naturally, e.g. in the context of synchronisation [5]. For certain particular systems, like phase oscillators [1], a reliable answer can be given. Other more heuristic approaches seem to give the expected results, but not always, and not provably so. In a nonlinear framework, it is not even straightforward to define the term *coupling strength*, or the concept of a *driving* and a *responding* system. Most authors assume that there are coupling terms in the underlying equations of motion. Either the coefficient or the time averaged magnitude of this term then yields the coupling strength. However, the assignment of these coupling terms is non-unique and the resulting coupling strength usually depends on the details of the observation. When one of the present authors in a recent paper [6] suggested a measure for information transport, one of the motivations was to define a concept that has a solid foundation in information theory and is less arbitrary with respect to the choice of variables. Rather than studying the magnitude of a coupling component, its information content was studied—an approach we will also follow in the present paper.

* Corresponding author. Tel.: +49-351-871-1210; fax: +49-351-871-1999.
E-mail address: kaiser@mpipks-dresden.mpg.de (A. Kaiser).

0167-2789/02/$ – see front matter
PII: S0167-2789(02)00432-3

Information theoretical techniques [7–10] like Shannon and Kolmogorov entropies are widely used to analyse nonlinear systems. Statistical dependence between signals is often quantified by their mutual information [12–16]. Often, one also wants to determine the predominant direction of information flow. Mutual information cannot be applied for this purpose because it is a static, symmetric property. In order to analyse dynamical properties such as driving and responding, quantities based on transition probabilities have to be considered. An appropriate relative entropy was introduced in [6]. For discrete state processes, the transfer entropy seems to give the desired answers. For continuous states, however, we have to face serious estimation problems which were only hinted at in [6]. The discussion of these difficulties forms the core of the present paper.

We will start by defining the relevant concepts for processes on a discrete, countable alphabet. For continuous state processes, most definitions and results appear to be quite similar. Only the probabilities have to be replaced by densities. Unfortunately, however, some cases which are of practical importance form pathological exceptions in this framework. In particular, systems with deterministic time evolution and/or deterministic coupling in general do not possess continuous (joint) densities.

1.1. Stationarity

In this paper, we will try to avoid assuming stationarity as long as possible. Time-dependent probabilities cannot be inferred from a single measurement. Thus, little can be learned from a single time series without making further assumptions. In rare but nevertheless important cases, we can make several recordings of the same phenomenon, like, e.g. evoked brain potentials. In that case, we can sample time-dependent probabilities and estimate time-dependent entropies etc. More commonly, only a single time series recording is available. Then, we *have to* assume stationarity, which means that all probabilities and transition probabilities for the process exist and are invariant under time shifts. For certain applications, a slow time dependence of the process can be tolerated [17]. Assuming stationarity, we can take multiple sections from the same series to be independent samples for the estimation of probabilities.

2. Discrete processes

Let X_i and Y_i be two discrete random processes with states x_i, y_i from a countable alphabet A. The probability distribution of a process X_i is given by P^{X_i}. The states may be multivariate without any need for special notation. Nevertheless, we need to define processes consisting of words, or embedding vectors. These are then composite processes $\mathbf{X}_i^{(k)} = X_i \otimes \cdots \otimes X_{i-k+1}$ of k variables. We write $P^{\mathbf{X}_i^{(k)}}$ for the probability distribution of the k-dimensional process $\mathbf{X}_i^{(k)}$. Finally, the joint probability distribution of the $k+l$-dimensional process $\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}$ is $P^{\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}}$. Note that, in general, $P^{X_i} = P^{X_{i+\tau}}$ etc. does not hold. Only if stationarity is assumed, we will drop the absolute time index.

Using $\mathbf{x}_i^{(k)} = (x_i, \ldots, x_{i-k+1})$ as the k-dimensional state of a process $\mathbf{X}_i^{(k)}$ etc., we write

$$p(x_i) = P^{X_i}(\{x_i\}), \qquad p(\mathbf{x}_i^{(k)}) = P^{\mathbf{X}_i^{(k)}}(\{\mathbf{x}_i^{(k)}\}), \qquad p(\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) = P^{\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}}(\{(\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)})\}) \tag{1}$$

The argument has a threefold significance. For example, $p(x_i)$ is the probability at time i for X_i to be in state $x_i \in A$. Analogously, $p(y_j)$ is the probability for Y_j to be in state $y_j \in A$ at time j. In general, these $p(\cdot)$ are different functions. Later, we will use the relation $\sum_{x_i \in A} p(\ldots, x_{i+1}, x_i, x_{i-1}, \ldots) = p(\ldots, x_{i+1}, x_{i-1}, \ldots)$.

Transition probabilities are written

$$p(x_{i+1}|\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) = P^{X_{i+1}|\mathbf{X}_i^{(k)}=\mathbf{x}_i^{(k)}, \mathbf{Y}_j^{(l)}=\mathbf{y}_j^{(l)}}(\{x_{i+1}\})$$

denoting the probability of finding X_{i+1} in x_{i+1} when $\mathbf{X}_i^{(k)}$ and $\mathbf{Y}_j^{(l)}$ were in $\mathbf{x}_i^{(k)}$ and $\mathbf{y}_j^{(l)}$, respectively. We will frequently use the relation $p(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) = p(x_{i+1}|\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) p(\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)})$.

2.1. Entropy and mutual information

The occupation of the state x_i by the process X_i carries the uncertainty $\log 1/p(x_i)$. Averaging over all states results in the Shannon entropy

$$H(X_i) = \sum_{x_i \in A} p(x_i) \log \frac{1}{p(x_i)}, \tag{2}$$

which is a functional that maps the probability distribution onto a scalar. The concept is readily generalised to words $\mathbf{x}_i^{(k)}$ of length k. When Shannon and Weaver [8] introduced this quantity, he also showed that it is the only functional, up to a pre-factor, that is continuous in the $p(x_i)$, increasing with the size of the alphabet, and additive under composition of random variables.

Suppose we have some (possibly erroneous) a priori guess of the probability distribution of a process X_i which we denote by $q(x_i) = Q^{X_i}(\{x_i\})$. We can quantify the error which is made when using $q(x_i)$ instead of the true $p(x_i)$ by averaging the resulting difference in uncertainty, $\log 1/q(x_i) - \log 1/p(x_i)$, which leads to the Kullback entropy [8–11]

$$K_{p|q}(X_i) = \sum_{x_i \in A} p(x_i) \log \frac{p(x_i)}{q(x_i)}. \tag{3}$$

The Kullback entropy is always non-negative [10], which follows from the *log sum inequality*, which is a result of Jensen's inequality [10,18]. It says that for arbitrary non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{4}$$

holds with equality if and only if $a_i/b_i = \text{constant}$. From this we have

$$K_{p|q}(X_i) \geq \left(\sum p(x_i) \right) \log \frac{\sum p(x_i)}{\sum q(x_i)} = 0. \tag{5}$$

Suppose two processes X_i and Y_j are observed. When X_i and Y_j are independent, then the joint probability of $X_i \otimes Y_j$ factorises. The Kullback entropy (Eq. (3)) that takes this a priori assumption $q(x_i, y_j) = p(x_i)p(y_j)$ reads

$$M(X_i, Y_j) = \sum_{x_i \in A} \sum_{y_j \in A} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{6}$$

which is the well known mutual information between the two processes X_i and Y_j [8–10]. Often, one takes $j = i - \tau$, where τ is called the time lag, in order to study time-delayed dependences [12,13].

Using the identity for joint probabilities, the mutual information can be expressed through Shannon entropies

$$M(X_i, Y_j) = H(X_i) + H(Y_j) - H(X_i, Y_j). \tag{7}$$

Evidently, the mutual information is symmetric in its arguments.

The mutual information is zero for independent processes. The opposite special case occurs when X_i and Y_j are linked by a mapping, e.g. $Y_j = f \circ X_i$. Then it holds

$$p(x_i, y_j) = P^{X_i}(\{x_i\} \cap f^{-1}(\{y_j\})) = p(x_i)\delta_{y_j, f(x_i)}, \qquad p(y_j) = \sum_{\tilde{x}_i \in A} p(\tilde{x}_i)\delta_{y_j, f(\tilde{x}_i)}.$$

Inserting this into Eq. (6), one gets for the mutual information

$$M(X_i, f \circ X_i) = -\sum_{x_i \in A} p(x_i) \log \sum_{\tilde{x}_i \in A} p(\tilde{x}_i)\delta_{f(x_i), f(\tilde{x}_i)}. \tag{8}$$

It follows that $M(X_i, f \circ X_i) \leq H(X_i)$. If and only if f is injective, the second sum reduces to $p(x_i)$ and the inequality becomes an equality.

2.2. Dynamical entropies

The dynamical structure of a random process is reflected by the transition probabilities $p(x_{i+1}|\mathbf{x}_i^{(k)})$. The uncertainty of a transition into a new state, given the past states, is defined by $\log 1/p(x_{i+1}|\mathbf{x}_i^{(k)})$. Inserting this into the expression of the Shannon entropy, Eq. (2), and averaging over all values of the initial points $\mathbf{x}_i^{(k)}$, one gets the conditional Shannon entropy, which reads [9,10]

$$H(X_{i+1}|\mathbf{X}_i^{(k)}) = \sum_{\mathbf{x}_i^{(k)} \in A^k} p(\mathbf{x}_i^{(k)}) \sum_{x_{i+1} \in A} p(x_{i+1}|\mathbf{x}_i^{(k)}) \log \frac{1}{p(x_{i+1}|\mathbf{x}_i^{(k)})} = \sum_{x_{i+1}, \mathbf{x}_i^{(k)}} p(x_{i+1}, \mathbf{x}_i^{(k)}) \log \frac{1}{p(x_{i+1}|\mathbf{x}_i^{(k)})}.$$

When taking the limit $k \to \infty$ one gets the Shannon entropy rate [8,20].

Usually, if a process with unknown underlying dynamics is observed, one has to assume a priori transition probabilities $q(\cdot|\cdot)$ instead of the *true* transition probabilities $p(\cdot|\cdot)$. This results in an increase of uncertainty on average. A measure for the loss of information due to the mistaken assumption can be constructed analogously to Eq. (3), which leads to the conditional Kullback entropy [9,10]

$$K_{p|q}(X_{i+1}|\mathbf{X}_i^{(k)}) = \sum_{x_{i+1}, \mathbf{x}_i^{(k)}} p(x_{i+1}, \mathbf{x}_i^{(k)}) \log \frac{p(x_{i+1}|\mathbf{x}_i^{(k)})}{q(x_{i+1}|\mathbf{x}_i^{(k)})}. \tag{9}$$

Like the Kullback entropy, the conditional version is non-negative.

A very straightforward approach, using conditional Kullback entropy, to quantify the dependence *in the dynamics* between the processes $\mathbf{X}_i^{(k)}$ and $\mathbf{Y}_j^{(l)}$ is the conditional mutual information [9,10]. It is given by

$$M(X_{i+1}, Y_{j+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = \sum_{x_{i+1}, \mathbf{x}_i^{(k)}} \sum_{y_{j+1}, \mathbf{y}_j^{(l)}} p(x_{i+1}, \mathbf{x}_i^{(k)}, y_{j+1}, \mathbf{y}_j^{(l)}) \log \frac{p(x_{i+1}, y_{j+1}|\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)})}{p(x_{i+1}|\mathbf{x}_i^{(k)}) p(y_{j+1}|\mathbf{y}_j^{(l)})}.$$

This definition has the drawback that it is still symmetric under exchange of the processes. Dynamical dependence can be quantified but not the direction of the information exchange. Thus, we do not want to follow this approach but instead introduce another concept in the next section.

2.3. Transfer entropy

Suppose the future state x_{i+1} of X_i depends on the k past states of $\mathbf{X}_i^{(k)}$, but not on the l past states of $\mathbf{Y}_j^{(l)}$. Then the generalised Markov property

$$p(x_{i+1}|\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) = p(x_{i+1}|\mathbf{x}_i^{(k)})$$

holds. If there is any such dependence, it can be quantified by the transfer entropy, which is obtained by inserting $p(x_{i+1}|\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)})$ as underlying transition probability and $p(x_{i+1}|\mathbf{x}_i^{(k)})$ as a priori transition probability in Eq. (9)

$$T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = \sum_{x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}} p(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) \log \frac{p(x_{i+1}|\mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)})}{p(x_{i+1}|\mathbf{x}_i^{(k)})}. \tag{10}$$

In contrast to the mutual information, the transfer entropy T is explicitly non-symmetric under exchange of X and Y. As a special case of the conditional Kullback entropy, T is non-negative. Rewriting T as a difference of conditional Shannon entropies, one gets

$$T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = H(X_{i+1}|\mathbf{X}_i^{(k)}) - H(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = M(X_{i+1}, \mathbf{Y}_j^{(l)}|\mathbf{X}_i^{(k)}).$$

The quantity $M(X_{i+1}, \mathbf{Y}_j^{(l)}|\mathbf{X}_i^{(k)})$ is a special case of conditional transfer information [9,15].

Later, we will need the transfer entropy as a sum of Shannon entropies which reads

$$T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = H(\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}) - H(\mathbf{X}_{i+1}^{(k+1)} \otimes \mathbf{Y}_j^{(l)}) + H(\mathbf{X}_{i+1}^{(k+1)}) - H(\mathbf{X}_i^{(k)}). \tag{11}$$

The transfer entropy can also be expressed as a sum of mutual informations

$$T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = M(X_{i+1} \otimes \mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) - M(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = M(\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}, X_{i+1}) - M(X_{i+1}, \mathbf{X}_i^{(k)}).$$

2.4. A discrete example

Let us have a look at a very simple example where everything can be calculated in order to demonstrate the difference in the behaviour of time-delayed mutual information and transfer entropy. Consider two binary state stationary Markov processes X and Y, where X is autonomous but Y depends on the state of X. Let the transition probabilities for X be such that the state is changed from 0 to 1 and from 1 to 0 with probability 1 at each time step. This process is random only through the initial condition. The dynamics of Y is such that y_{i+1} will be chosen randomly but biased by the state of x_i. The probability that $y_{i+1} = x_i$ will be $(1 + c)/2$ and that $y_{i+1} = 1 - x_i$ will be $(1 - c)/2$. There is no dependence of y_{i+1} on y_i. These rules are sufficient to define all transition probabilities $p(x_{i+1}, y_{i+1}|x_i, y_i)$. By a simple eigenvector calculation one finds that the distribution $p(x_i = 1, y_i = 1) = p(x_i = 0, y_i = 0) = (1 + c)/4$, $p(x_i = 0, y_i = 1) = p(x_i = 1, y_i = 0) = (1 - c)/4$ is left invariant by the transition probabilities. From this, one obtains $p(x_{i+1}, y_{i+1}, x_i, y_i)$ and all the other probabilities. Taking $k = l = 1$ we can then compute

$$M(X_i, Y_i) = M(X_{i+1}, Y_i) = M(Y_{i+1}, X_i) = \tfrac{1}{2}[(1 + c)\log(1 + c) + (1 - c)\log(1 - c)],$$
$$T(X_{i+1}|X_i, Y_i) = 0,$$
$$T(Y_{i+1}|Y_i, X_i) = \tfrac{1}{2}c[(1 + c)\log(1 + c) - (1 - c)\log(1 - c)] - \tfrac{1}{2}(1 + c^2)\log(1 + c^2).$$

These results are shown in Fig. 1. Due to the periodicity of x, simultaneous and time-delayed mutual information are identical for all values of the coupling parameter c. Thus, the interpretation of time-delayed mutual information as an indicator of information transport that is commonly found in the literature [21,22] would indicate that information is exchanged at the same rate in both directions. The causal structure of the process, however, does not allow information to flow from Y to X. This is correctly reported by the transfer entropy which is identically 0 for this direction. Except for couplings $c = 0$ and $c = \pm 1$, the transfer entropy is positive for the direction in which the processes are coupled. Noteworthy is the case of complete synchronisation, $c = \pm 1$, in which no information flow

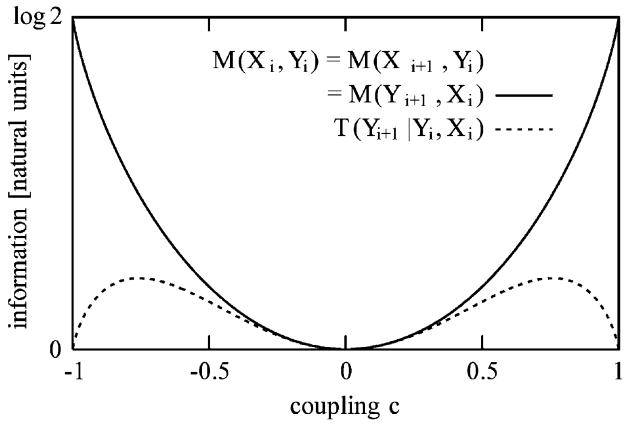

Fig. 1. Mutual information and transfer entropy as a function of a coupling parameter c for two binary processes X and Y, where X is autonomous and Y is probabilistically coupled to X. Note that $T(X_{i+1}|X_i, Y_i) = 0$ for all c.

is found. In fact, there is no way a posteriori to distinguish if Y is autonomously periodic or coupled to X in this case.

3. Continuous processes

Now we want to study continuous processes. Because the space of states is not countable any longer, values like Shannon entropy, mutual information, or transfer entropy cannot be expressed directly by the probabilities of specific, single states. Therefore, we have to restrict the analysis to processes which possess a Lebesgue probability density.

Let X_i and Y_j be (possibly multivariate) random processes and x_i and y_j, respectively, the states in which they are found. The process X_i owns a density if the probability distributions P^{X_i} can be written as

$$P^{X_i}(B) = \int_B g(x_i)\,\mathrm{d}x_i$$

for any set $B \in \mathcal{B}(\mathbf{R}^m)$, where $\mathcal{B}(\mathbf{R}^m)$ denotes the Borel σ-algebra of $\mathbf{R}^m$ [18,23,24]. The non-negative real function g is called Lebesgue probability density (of X_i).

As for the probability distributions in the discrete case, we mark the densities only by their arguments in order to distinguish them and to refer to the corresponding probability distribution. For example, $g(x_i)$ is the density of the distribution P^{X_i} of the process X_i and the density $g(x_i, y_j)$ refers to the distribution $P^{X_i \otimes Y_j}$ of the joint process $X_i \otimes Y_j$. We have the obvious but useful identity $\int g(\ldots, x_{i+1}, x_i, x_{i-1}, \ldots)\,\mathrm{d}x_i = g(\ldots, x_{i+1}, x_{i-1}, \ldots)$. Also here, we do not make any restrictions to stationarity yet and the generalisation to embedding vectors $\mathbf{x}_i^{(k)}$ and $\mathbf{y}_i^{(l)}$ of length k and l, respectively, is straightforward. We write $g(x_{i+1}|\mathbf{x}_i^{(k)})$ for the density of the transition probability $P^{X_{i+1}|\mathbf{X}_i^{(k)}=\mathbf{x}_i^{(k)}}$. In addition, for continuous processes we have the relation $g(x_{i+1}, \mathbf{x}_i^{(k)}) = g(x_{i+1}|\mathbf{x}_i^{(k)})g(\mathbf{x}_i^{(k)})$.

With the densities taking over the part of the probabilities, many properties of entropy and mutual information are inherited from the discrete case. However, there are some differences. One is the behaviour under coordinate transformations which will be investigated in more detail below. More critical is the restriction to processes that own densities, because there are important processes that do not. Also, densities represented by Dirac's δ-distribution cannot be considered, since its logarithm is not defined.

3.1. Shannon entropy and mutual information

For continuous processes, the uncertainty that the process X_i is in the state x_i can be defined as $\log 1/g(x_i)$ [8–10]. Again, the continuous Shannon entropy[1] is defined by averaging these uncertainties

$$H(X_i) = \int g(x_i) \log \frac{1}{g(x_i)} \, \mathrm{d}x_i, \tag{12}$$

which characterises the probability distribution P^{X_i}. As above, it still depends on the time index i.

As mentioned above, two processes X_i and Y_j are defined as independent if the joint distributions of $X_i \otimes Y_j$ can be factorised, $P^{X_i \otimes Y_j} = P^{X_i} \otimes P^{Y_j}$, where $P^{X_i} \otimes P^{Y_j}$ denotes the product measure. When the processes own probability densities, this is equivalent to the factorisation of the joint probability density: $g(x_i, y_j) = g(x_i)g(y_j)$ [18,24]. Therefore, in order to obtain a value which measures the information shared between two continuous processes one can pursue the same derivation as for the mutual information in the discrete case (Eq. (6)). This yields the continuous mutual information [8–10], which reads

$$M(X_i, Y_j) = \iint g(x_i, y_j) \log \frac{g(x_i, y_j)}{g(x_i)g(y_j)} \, \mathrm{d}x_i \, \mathrm{d}y_j. \tag{13}$$

It can also be written as a sum of continuous Shannon entropies and yields the analogue of Eq. (7).

In order to calculate the continuous Shannon entropy and mutual information, all the densities involved in the formulae have to exist. For the Shannon entropy, the most important exception is a process where the density contains δ-distributions, e.g. a one-dimensional map with a periodic solution. For mutual information, the two processes may not be coupled through a deterministic function since then, even if the process X_i possesses a smooth density, $X_i \otimes Y_j$ does not. The joint Lebesgue density would then be

$$g(x_i, y_j) = g(x_i)\delta(y_j - f(x_i)). \tag{14}$$

and $g(y_j) = \int g(\tilde{x}_i)\delta(y_j - f(\tilde{x}_i)) \, \mathrm{d}\tilde{x}_i$. When inserting this in Eq. (13), a $\log \delta(\cdot)$ term will still be left and the continuous mutual information does not exist.

3.2. Transfer entropy

The uncertainty in x_{i+1} conditioned on $\mathbf{x}_i^{(k)}$ is given through the density of the transition probability as $\log 1/g(x_{i+1}|\mathbf{x}_i^{(k)})$. The resulting continuous transfer entropy

$$T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = \iiint g(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) \log \frac{g(x_{i+1}|\mathbf{x}^{(k)}, \mathbf{y}_j^{(l)})}{g(x_{i+1}|\mathbf{x}_i^{(k)})} \, \mathrm{d}x_{i+1} \, \mathrm{d}\mathbf{x}_i^{(k)} \, \mathrm{d}\mathbf{y}_j^{(l)}, \tag{15}$$

again quantifies the directed information exchange. The continuous transfer entropy can again be written as a sum of continuous Shannon entropies which yields the same equation as Eq. (11).

3.3. Transformation invariance

For physical processes, an important question is, how quantities like Shannon entropy, mutual information, and transfer entropy behave under coordinate changes. This will be investigated in detail now. Suppose φ is a C^1-diffeomorphism which maps X_i on $\tilde{X}_i$ by $\tilde{X}_i = \varphi \circ X_i$. Inserting the identity for the Lebesgue densities

$$g(x_i) = \tilde{g}(\varphi(x_i))|\det(D\varphi)(x_i)|$$

[1] In certain scientific communities, this entropy is called *differential entropy*, see, e.g. [10].

into the expression for the continuous Shannon entropy and applying the transformation theorem for integrals [23] leads to

$$H(\tilde{X}_i) = H(\varphi \circ X_i) = H(X_i) + \int \tilde{g}(\tilde{x}_i) \log|\det(D\varphi)(\varphi^{-1}(\tilde{x}_i))| \, \mathrm{d}\tilde{x}_i. \tag{16}$$

Thus, when calculating the Shannon entropy for continuous processes, the units of measurement always have to be specified [8–10].

Independence is a property that should not depend on the units of measurement. However, if the mutual information is not 0, we have to investigate if its value will be invariant under coordinate transformations. Consider φ above and another C^1-diffeomorphism ψ acting on Y_j. Then, the map $\varphi \otimes \psi$ is also a C^1-diffeomorphism: $\tilde{X}_i \otimes \tilde{Y}_j = \varphi \otimes \psi \circ (X_i \otimes Y_j) = (\varphi \circ X_i) \otimes (\psi \circ Y_j)$. The Jacobian determinant of $\varphi \otimes \psi$ can be written as a product of those of φ and ψ:

$$\det(D(\varphi \otimes \psi))(x_i, y_j) = \det(D\varphi)(x_i) \det(D\psi)(y_j). \tag{17}$$

If we now write the mutual information as a sum of Shannon entropies (Eq. (7)), and apply the transformation rule for Shannon entropies, Eq. (17) yields the invariance of the mutual information

$$M(\varphi \circ X_i, \psi \circ Y_j) = M(X_i, Y_j).$$

Now, let φ and ψ act on k- and l-dimensional delay vectors and consider a third C^1-diffeomorphism χ which maps X_{i+1} to $\tilde{X}_{i+1} = \chi \circ X_{i+1}$. Then, the Jacobian determinant of $\chi \otimes \varphi \otimes \psi$ factorises

$$\det(D(\chi \otimes \varphi \otimes \psi))(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_j^{(l)}) = \det(D\chi)(x_{i+1}) \det(D\varphi)(\mathbf{x}_i^{(k)}) \det(D\psi)(\mathbf{y}_j^{(l)})$$

and it follows

$$T(\chi \circ X_{i+1} | \varphi \circ \mathbf{X}_i^{(k)}, \psi \circ \mathbf{Y}_j^{(l)}) = T(X_{i+1} | \mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}).$$

Thus, the continuous transfer entropy is a measure of directed dependence between processes which is invariant under coordinate transformations.

3.4. Gaussian processes

For processes with Gaussian distributions, Shannon entropy, mutual information and transfer entropy can be calculated analytically from the covariance matrices [12]. Let $C_{\mu\nu}(\mathbf{X}_i^{(k)}) = c(X_{i-k+\mu}, X_{i-k+\nu})$, $\mu, \nu = 1, \ldots, k$, where $c(X_m, X_n) = \mathrm{E}[X_m X_n] - \mathrm{E}[X_m]\,\mathrm{E}[X_n]$ are the covariances of X_m and X_n. Here, $\mathrm{E}[X]$ denotes the expected value of X. If we allow delay vectors, the Shannon entropy reads [10,12]

$$H(\mathbf{X}_i^{(k)}) = \tfrac{1}{2} k \log(2\pi e) + \tfrac{1}{2} \log(\det C(\mathbf{X}_i^{(k)})).$$

When $\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}$ is multivariate Gaussian, then it follows from Eq. (7) that

$$M(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = \frac{1}{2} \log \frac{\det C(\mathbf{X}_i^{(k)}) \det C(\mathbf{Y}_j^{(l)})}{\det C(\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)})}. \tag{18}$$

Analogously, the transfer entropy is given by

$$T(X_{i+1} | \mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) = \frac{1}{2} \log \frac{\det C(\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}) \det C(X_{i+1} \otimes \mathbf{X}_i^{(k)})}{\det C(X_{i+1} \otimes \mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}) \det C(\mathbf{X}_i^{(k)})} \tag{19}$$

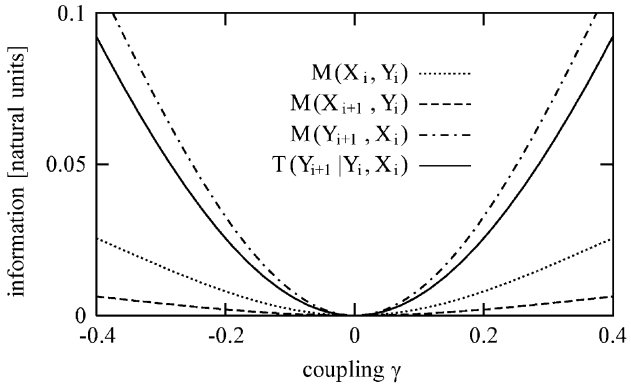

Fig. 2. Mutual information and transfer entropy as a function of a coupling parameter γ for two linear stochastic processes X and Y, where X is autonomous and Y is coupled to X. Note that $T(X_{i+1}|X_i, Y_i) = 0$ for all γ.

if $X_{i+1} \otimes \mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}$ is multivariate Gaussian. Of course, multivariate Gaussian distributions are mostly relevant to linear stochastic processes where information theory is not much called for. We will use the above formulae mostly as a testing ground where exact results are available.

3.5. A continuous example

Let us finish the section with a simple example that admits analytical treatment. Consider two linear autoregressive processes with Gaussian white noise increments, where process X is autonomous and Y is coupled to X

$$X_{i+1} = \alpha X_i + \eta_i^X, \qquad Y_{i+1} = \beta Y_i + \gamma X_i + \eta_i^Y. \tag{20}$$

With η^X and η^X Normal random numbers, all distributions involved are Gaussian or multivariate Gaussian and we can use Eqs. (18) and (19). When we take $k = l = 1$, and define $u = 1/(1-\alpha^2)$, $v = 1/(1-\beta^2)$, and $w = 1/(1-\alpha\beta)$ we obtain

$$c(X_i, X_i) = u, \qquad c(Y_i, Y_i) = v + \gamma^2(1+\alpha\beta)uvw, \qquad c(X_{i+1}, X_i) = \alpha u,$$
$$c(Y_{i+1}, Y_i) = \beta v + \gamma^2(\alpha+\beta)uvw, \qquad c(X_{i+1}, Y_i) = \alpha^2\gamma uw, \qquad c(X_i, Y_{i+1}) = \alpha\beta\gamma uw + \gamma u,$$
$$c(X_i, Y_i) = \alpha\gamma uw.$$

With this, we can compute the mutual information and transfer entropy for both directions. The results for $\alpha = 0.5$ and $\beta = 0.6$ as a function of the coupling parameter γ are shown in Fig. 2. Note in particular, that the non-zero value of $M(X_{i+1}, Y_i)$ could erroneously suggest that there is a dynamical dependence of X on Y. In contrast, the transfer entropy clearly reflects the structure of the model given in Eq. (20).

4. Coarse grained entropies

If we plan to estimate either discrete or continuous entropies from data, we will face very different levels of difficulty in the two cases. Probabilities for discrete states can be obtained with relative ease, whereas the estimation of densities and in particular transition probability densities can be quite tricky and in any case requires lots of data. We will discuss some of the available techniques later in this paper. Many attempts have been made to take an intermediate approach, converting the continuous variable into discrete states by some coarse graining procedure

and then using the formalism for discrete processes. In this section, we will show on the one hand that this procedure is partly justified for mutual information since it converges to the continuous value monotonically from below. For transfer entropy, a similar statement does not necessarily hold. More importantly, transformation invariance as shown in Section 3.3 holds for the continuous entropies only, but is usually broken for coarse grained approximations.

4.1. Convergence under refinement

Consider a continuous random process $\mathbf{X}_i^{(k)}$ with a continuous density $g(\mathbf{x}_i^{(k)})$. The most widespread method of coarse graining is to partition k-dimensional space into a set $\{I\}$ of arbitrarily chosen non-overlapping cubes $I(m)$. Then, the cube index $m \in A$ can be used as a discrete state from an alphabet A. For the discrete process $\tilde{X}_i^{(k)}$, the probability of state m is given by $p(m) = \int_{I(m)} g(\mathbf{x}_i^{(k)})\,\mathrm{d}\mathbf{x}_i^{(k)}$. We denote the Shannon entropy $H(\tilde{X}_i^{(k)})$ of the discrete process $\tilde{X}_i^{(k)}$ with $H_{\{I\}}(\mathbf{X}_i^{(k)})$ and call it Shannon entropy of $\mathbf{X}_i^{(k)}$ on the partition $\{I\}$.

Now consider a series of partitions $\{I\}_n$ with $\|\{I\}_n\| \to 0$. The norm $\|\{I\}_n\|$ is defined as the largest edge size $\Delta_j(m_n)$ of all partition elements. Using the mean value theorem, one can prove the convergence

$$H_{\{I\}_n}(\mathbf{X}_i^{(k)}) + \left\langle \sum_{j=i-k+1}^{i} \log \Delta_j(m_n) \right\rangle_{\{I\}_n} \xrightarrow{\|\{I\}_n\|\to 0} H(\mathbf{X}_i^{(k)}), \tag{21}$$

which is only guaranteed if $H(\mathbf{X}_i^{(k)})$ exist as a Riemann integral, see also [10]. Thus, coarse grained versions of Shannon entropy in the limit of infinitely fine partitions differ from the continuous Shannon entropy by the expected value of the logarithm of the cube volumes $\prod_j \Delta_j(m_n)$.

For the mutual information and transfer entropy, we may partition the spaces of $X_{i+1}, \mathbf{X}_i^{(k)}$, and $\mathbf{Y}_j^{(l)}$ individually. The partitioning of $\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}$ and $X_{i+1} \otimes \mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}$ follows immediately by taking the Cartesian product. Applying Eq. (21) and inserting in Eqs. (7) and (11), we get the convergence toward the continuous mutual information or continuous transfer entropy, respectively

$$M_{\{I\}_n}(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) \xrightarrow{\|\{I\}_n\|\to 0} M(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}), \tag{22}$$

$$T_{\{I\}_n}(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) \xrightarrow{\|\{I\}_n\|\to 0} T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}). \tag{23}$$

Multidimensional cubes are the natural choice as elements of the partition due to the Cartesian structure of the underlying space. When using elements with other shape one gets an additional geometrical factor. For deterministically coupled systems, we find that the mutual information on a partition $M_{I(A)}(\mathbf{X}_i^{(k)}, f \circ \mathbf{X}_i^{(k)})$ diverges like $\log 1/\|\{I\}\|$.

The convergence of the mutual information in Eq. (22) can be generalised to other types of partitions. One can partition the space of $\mathbf{X}_i^{(k)} \otimes \mathbf{Y}_j^{(l)}$ into arbitrarily non-overlapping cubes and then simply follow the derivation of Eq. (21), see also [14]. For the transfer entropy, this seems not to be possible without further assumptions about the densities.

Let us investigate the convergence behaviour more closely. First, we show that the Kullback entropy on a series of refined partitions increases monotonically from below. Consider a k-dimensional continuous process $\mathbf{Z}$ and a partition $\{I\}$ on it consisting of non-overlapping cubes $I(m)$ with volume $V(I(m))$. This defines a discrete process by $\tilde{Z} = m$ if $\mathbf{Z} \in I(m)$ with the distribution given through $p(m) = \int_{I(m)} g(\mathbf{z})\,\mathrm{d}\mathbf{z}$. It is sufficient to show that the Kullback entropy $K_{\{I'\}}(\mathbf{Z}) \geq K_{\{I\}}(\mathbf{Z})$ if the partition $\{I'\}$ is the simplest possible refinement of $\{I\}$ (written $\{I'\} > \{I\}$) constructed by dividing one of the cubes in $\{I\}$, say $I(\hat{m})$, into two cubes $I'(\hat{m}'_1)$ and $I'(\hat{m}'_2)$. Thus, all the probabilities $p(m)$ and a priori probability $q(m)$ remain the same, except for $p'(\hat{m}'_1)$, $p'(\hat{m}'_2)$ for which

$p'(\hat{m}'_1) + p(\hat{m}'_2) = p(\hat{m})$, and $q'(\hat{m}'_1), q'(\hat{m}'_2)$ for which $q'(\hat{m}'_1) + q(\hat{m}'_2) = q(\hat{m})$. Using the sum log inequality (Eq. (4)), it follows:

$$K_{\{I\}}(\mathbf{Z}) = p(\hat{m}) \log \frac{p(\hat{m})}{q(\hat{m})} + \sum_{m \neq \hat{m}} p(m) \log \frac{p(m)}{q(m)}$$

$$\leq p'(\hat{m}'_1) \log \frac{p'(\hat{m}'_1)}{q'(\hat{m}'_1)} + p'(\hat{m}'_2) \log \frac{p'(\hat{m}'_2)}{q'(\hat{m}'_2)} + \sum_{m \neq \hat{m}} p(m) \log \frac{p(m)}{q(m)} = K_{\{I'\}}(\mathbf{Z}).$$

Because mutual information is a special case of Kullback entropy, it follows immediately that

$$\{I'\} > \{I\} \Rightarrow M_{\{I'\}}(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}) \geq M_{\{I\}}(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)}). \tag{24}$$

Consequently, for any series of refined partitions, the mutual information increases monotonically and with Eq. (22) converges from below toward the continuous mutual information.

Unfortunately, a similar result cannot be shown for the conditional Kullback entropy Eq. (9). The reason is that in general the transfer probabilities are not additive in their argument of condition: $P(A|B) + P(A|C) \neq P(A|B \cup C)$. Thus, monotonic convergence holds for mutual information, but not necessarily for transfer entropy. In the examples we have studied so far, transfer entropy does seem to grow monotonically under refinement of the partitions until finite sample effects set in.

4.2. No invariance without convergence

The need to take the limit of small length scales in quantities like fractal dimension and Lyapunov exponents has been discussed thoroughly in the literature. As argued in [19], the desirable invariance property of these concepts are broken for finite length scale estimates. Exactly the same problem arises for coarse grained versions of mutual information and transfer entropy. The invariance under diffeomorphisms shown above holds for the continuous quantities only. If we work in a regime of length scales where convergence can be established to a certain accuracy because estimates do not change significantly under further refinement of the partitions, it is reasonable to assume that invariance holds. Often, coarse grained entropies are used with the justification that they are not to be taken as numerical estimates but only for the relative comparison of signals and states. But even that is dangerous. Consider the case that we want to measure the information exchange between two processes. Even if the information transport between them is symmetric, we can easily measure a significant difference by rescaling one of the signals. This is particularly problematic when the two signals are of different physical nature and no natural relative scaling is available. As an example for this problem, we will study heart and breath rate later in this paper.

5. Estimation from data

5.1. Multiple realisations and stationarity

Suppose, we have N realisations of the stochastic process $X_i(n)$, $i = 0, 1, \ldots, T$ and $n = 1, 2, \ldots, N$. Here, the index i stands for time and n marks the realisation. If N is large enough, then the probability distributions $p(x_i)$ or densities $g(x_i)$ etc. can be estimated using the available samples of $X_i(n)$. Thus, we do not have to assume stationarity in order to form averages. If desired, we can estimate Shannon entropy, mutual information, transfer entropy, etc. as functions of time and study their evolution. To our knowledge, this has not been done in the literature, but it might give new insights for the study of transient phenomena like evoked brain potentials.

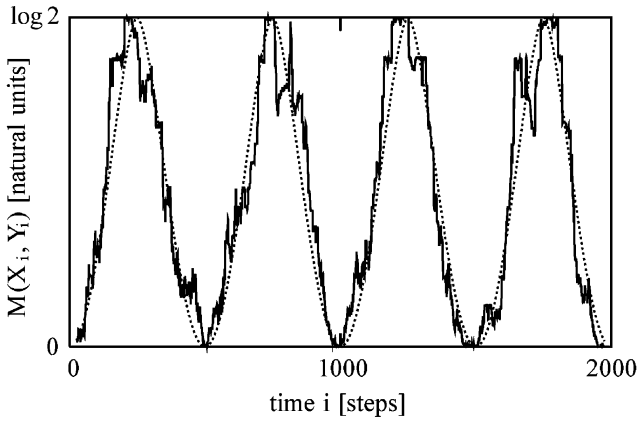

Fig. 3. Mutual information for a non-stationary process versus time. Estimating probabilities by averages over 51 time steps already gives the qualitative behaviour of the theoretical result (dotted). Assuming stationarity would yield $M(\bar{X}, \bar{Y}) = 0$.

On the other hand, if only a single realisation $X_i, i = 0, 1, \ldots, T$ is given, probability distributions cannot be estimated, unless we assume that $P^{X_i} \approx P^{X_{i+\tau}}$ for a certain range of values of τ and then use all the samples from X_i to $X_{i+\tau}$ in order to estimate $p(x_i)$. Most commonly, it is assumed that $P^{X_i} = P^{X_{i+\tau}}\ \forall\tau$, i.e., that the process is stationary. In this case, all the time dependence of averages like entropy is lost. If the assumption is not actually true, we should not find convergence for increasing sample size and the estimated values will not reflect the information content of the actual process.

An intermediate possibility is to assume slow variation of P^{X_i} as a function of time, such that $P^{X_i} \approx P^{X_{i+\tau}}$ holds for small τ. Then, averages over time windows centred around i can give meaningful results despite statistical errors.

Let us illustrate the preceding discussion with a simple example, two binary state non-stationary Markov processes X and Y, where X is autonomous with $p(x_i = 0) = p(x_i = 1) = 1/2$. The probability that y_i is equal to x_i oscillates slowly in time: $p(y_i = 0|x_i = 0) = p(y_i = 1|x_i = 1) = (1 + \sin i\omega)/2$. Then the mutual information $M(X_i, Y_i)$ oscillates between 0 and $\log 2$ (or 1 bit) with twice the frequency ω. If we assume stationarity, we get an averaged $\bar{p}(y|x) = 1/2$ and $M(\bar{X}, \bar{Y}) = 0$. Fig. 3 shows an example with $\omega = 2\pi/1000$ and $N = 2000$ points. Time-dependent mutual information has been estimated by averaging $p(x_i, y_i)$ over centred windows of length 51 time steps. The theoretical behaviour of $M(X_i, Y_i)$ (dotted) is already discernible despite the statistical fluctuations. Of course, averaging over such short times gives stable results only for discrete processes with few states.

5.2. *Estimating probability densities and their ratios*

For discrete processes, probabilities can be obtained with relative ease by the visitation frequency: $p(x_i) \approx n(x_i)/N$, where $n(x_i)$ is the number of observations found in state x_i and N is the total number of observations. Finite sample corrections for this estimator have been given, e.g. in [25–27]. Also, statistical errors can be computed under some assumptions [28,29]. For mutual information and transfer entropy, however, some of the assumptions made are problematic. In particular, using Eqs. (7) and (11) and computing bias and variance individually for each of them implies the assumption that the finite sample fluctuations of these terms are independent, which is usually not the case.

For continuous processes, in principle the density at an infinite number of states has to be estimated with a finite number of samples. Obviously, this is a hopeless task unless we are willing to make some assumptions on the densities. Basically, one can distinguish parametric and non-parametric approaches. Both will be discussed below.

For mutual information and transfer entropy, we need to estimate ratios of densities. The typical case is that the ratios, like $g(x_i, y_j)/g(x_i)g(y_j)$ in Eq. (13) are close to 1, whereas the individual terms can be of arbitrary magnitude. This subtle balance is of course very sensitive to statistical fluctuations. These can become very pronounced for differences of Shannon entropies, like in Eqs. (7) and (11), unless product partitions or coverings are used.

5.2.1. Parametric distributions

If the process is assumed to follow a distribution from a specific parametric family, it is sufficient to estimate these parameters and then calculate mutual information, transfer entropy, etc. analytically or numerically from the known functional form of the density. In Section 3.4, we have already given the analytical expressions for multivariate Gaussian distributions. For a random process with an arbitrary parametric distribution, expressions for $M(\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)})$ and $T(X_{i+1}|\mathbf{X}_i^{(k)}, \mathbf{Y}_j^{(l)})$ are not usually available analytically. However, once the distribution parameters have been estimated, these values can be calculated by numerical integration.

To our knowledge, there is no general scheme that yields systematic approximations in more than one dimensions that has proven useful for the study of the interrelation of processes. One would need to expand the densities in a controlled way that allows to assess the convergence to the continuous entropies. Most signals which are currently under study with respect to their coupling structure, like, e.g. electroencephalogram (EEG) channels or time variable hormone concentrations etc. are neither of Gaussian distribution nor even approximately so. Therefore, parametric representations have not been used extensively.

5.2.2. Uniform and adaptive partitions

The theoretical basis to estimate entropies for continuous processes using discretisation was already given in Section 4. The application is conceptually straightforward but presents a number of practical problems. When reducing the size of the partition elements, too few sample points fall into each element and convergence often cannot be established, with the consequences discussed above.

The simplest implementations use a regular mesh of boxes and refinements are done uniformly by decreasing the mesh spacing. For inhomogeneous distributions, this is quite wasteful. The mesh has to be fine enough to resolve all the details of the densities, but in many places we will then find very few points in each box. An attractive alternative is given by adaptive partitions where boxes are subdivided only locally in places where substructure is statistically significant. Such an algorithm for mutual information was already given by Fraser and Swinney [13], who subdivide boxes recursively by splitting them along each phase space dimension in turn. Schreiber [30] describes a variant using balanced trees. Recently, Darbellay [14] has described a very similar algorithm which, however, divides along all coordinate axes simultaneously. This avoids the problem of the previous methods which break the symmetry of mutual information under exchange of the arguments. In any case, an objective criterion is needed that tells us when to stop dividing further. References [13,14] use a χ^2-test for independence for this purpose.

Our main concern against adaptive partitioning is that convergence is not usually checked. Instead, authors rely on the χ^2-test alone. However, since convergence must be established in order to interpret the information theoretic quantities properly, this is a potential source of error. Due to its monotonic behaviour, the mutual information may be less problematic in this respect, but we refrain from using a modification of an adaptive algorithm for transfer entropy. If convergence is tested for, however, this could be a worthwhile possibility. The fixed mass method which we will discuss in Section 5.2.3 can be seen as an adaptive version of kernel estimation.

Another problem with partition-based density estimators is that a bias may arise due to serial correlations in the points that count in each bin. (For a detailed discussion of this problem, see [19].) The only known correction is to bin only points which are de-correlated in time, which is very wasteful in the amount of data needed. For kernel estimators (see below) there is a much more economic solution.

If only a test for independence of the continuous processes X_i and Y_i is desired, it is sufficient to check if the mutual information on an arbitrarily chosen partition is 0 within statistical fluctuations because mutual information converges monotonically from below. The simplest partition is a binary one, setting $\tilde{X}_i = 1$ if $X_i \leq x_c$ and $\tilde{X}_i = 0$ otherwise. In a statistical test, one could use the error formula derived in [14], or a surrogate data test [31].

5.2.3. Density estimation with kernel methods

Non-parametric estimation using kernel techniques is an attractive alternative to binning a distribution. It is discussed thoroughly in the literature, e.g. [32]. The main assumption is that the probability density is smooth enough such that structure below a certain kernel band width may be ignored. The simplest possibility is to estimate the density at a point $\tilde{x}$ by the number of points in a box of size ε divided by its volume. One can either choose a fixed box size ε (or band width), or one can require a fixed number of points for each box. Both methods are discussed below. Rather than simply counting points, one can give them distance-dependent weights by some kernel function.

In order to reconstruct the density of X_i at an arbitrary point $\tilde{x}_i$, the general form of a kernel estimator is given by:

$$\hat{g}(\tilde{x}_i) = \frac{1}{N\varepsilon} \sum_{n=1}^{N} K\left(\frac{\tilde{x}_i - x_i(n)}{\varepsilon}\right). \tag{25}$$

Here, $x_i(n), n = 1, \ldots, N$ are the N observed states of the process X_i. For stationary processes, $x_i(n)$ is given by x_n. The non-negative function K is called the kernel and determines the distance-dependent weight of each point. Examples for commonly used kernels are given in Table 1. The parameter ε is called the band width and determines the scale below which structure is ignored. If K satisfies

$$\int K(u)\,\mathrm{d}u = 1, \qquad \int uK(u)\,\mathrm{d}u = 0, \qquad \int u^2 K(u)\,\mathrm{d}u < \infty \tag{26}$$

and g is twice differentiable, then at each point x_i the convergence $\hat{g}(\tilde{x}_i) \to g(\tilde{x}_i)$ for $\varepsilon \to 0$ holds on average [32]. With a finite number of points, the convergence of course may not be seen since for small ε, statistical fluctuations become important.

The concept of kernel estimators can be generalised for higher dimensions. Unfortunately, the construction of an efficient kernel with fast convergence and small variance, depends strongly on the underlying structure and dynamics of the process [32]. For unknown dynamics all we can use is the Cartesian structure of state space. Therefore, we assume the multidimensional kernel to be a product

$$\hat{g}(\tilde{\mathbf{x}}_i^{(k)}) = \frac{1}{N\prod_{j=1}^{k}\varepsilon_j} \sum_{n=1}^{N}\prod_{j=1}^{k} K_j\left(\frac{\tilde{x}_{i-j+1} - x_{i-j+1}(n)}{\varepsilon_j}\right), \tag{27}$$

Table 1
Kernels and their definitions

Epanechnikov	$K(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right), \quad \lvert u\rvert < \sqrt{5}$
Biweight	$K(u) = \frac{15}{16}(1-u^2)^2, \quad \lvert u\rvert < 1$
Triangular	$K(u) = 1 - \lvert u\rvert, \quad \lvert u\rvert < 1$
Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(1/2)u^2}$
Rectangular	$K(u) = \frac{1}{2}, \quad \lvert u\rvert < 1$

where K_j is the kernel and ε_j the band width which may differ for each coordinate. In the case of the rectangular kernel, the sum is proportional to the number of sample points found within a distance ε_j in each direction.

Once the density can be estimated, we have to perform the integral over space, like, e.g. in Eq. (15). If multiple realisations are used, this is done by averaging over all samples. In time series work, the average is usually replaced by a sum in time over the available data points and $x_{i-j+1}(n)$ is given by x_{n-j+1}. Since the data are supposed to be distributed according to the density $g(\mathbf{x}_i^{(k)})$, all we have to do is to average the logarithm term over the data set. As proposed by Theiler [33] in the context of fractal dimension estimation, it is necessary to correct this scheme for a bias due to serial correlations when using time series data. One simply excludes terms from the sum in Eq. (27) which are too close in time to the reference point when taking the time average. The final formula for the transfer entropy, e.g. then reads

$$\hat{T}(X_{+1}|\mathbf{X}^{(k)}, \mathbf{Y}^{(l)}) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\hat{g}(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})\hat{g}(\mathbf{x}_i^{(k)})}{\hat{g}(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})\hat{g}(x_{i+1}, \mathbf{x}_i^{(k)})} \tag{28}$$

with

$$\hat{g}(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)}) = \alpha^{-1} \sum_{|n-i|>t} \frac{1}{\varepsilon_x^k \varepsilon_y^l} \prod_{j=1}^{k} K_j\left(\frac{x_{i-j+1} - x_{n-j+1}}{\varepsilon_x}\right) \prod_{j=1}^{l} K_j\left(\frac{y_{i-j+1} - y_{n-j+1}}{\varepsilon_y}\right),$$

etc. where $\alpha = N - 1 - \min(i + t, N) + \max(i - t, 1)$ is a normalisation constant and the de-correlation time t has been specified. We have made the natural choice of equal band width ε_x for each component of X, and ε_y for each component of Y. If the physical nature or the dynamical behaviour of X and Y are different, it will not always be easy to select the ratio of ε_x and ε_y properly.

5.2.3.1. Fixed band width. In the notation so far, we have implied that the kernel band widths ε_j are chosen constant. Local variations of the density are then reflected by the value of the sums in Eq. (27) etc. This approach has the advantage that the resolution is uniform in state space and can be specified in physical units. Furthermore, this approach is computationally the most attractive. Its main drawback is that the statistical error of the local density estimates may be highly non-uniform. For kernels with finite support, the sums may locally have very few non-zero terms, or none at all, resulting in statistical as well as systematic errors. For the rectangular kernel, a bias correction formula for entropies was derived by Grassberger [25]. In leading order, it replaces the logarithm of the point count by the digamma function $\psi(x) = \Gamma'(x)/\Gamma(x)$, which is regular at the value 0, $\psi(0) = 0$. For mutual information and transfer entropy, one is tempted to apply Eq. (11) and then use the same correction formulae. This was done in [6], but also Roulston [28,29] has used the same rationale with a different bias formula. This approach is not viable, though, since the finite sample fluctuations in, e.g. $\hat{g}(x_i, y_j)$ and $\hat{g}(x_i)$ are not independent and we cannot correct their bias separately. It turns out that the adverse effect of the bias correction is much worse than the bias itself. Therefore, since a proper correction is not available, we propose to use the uncorrected kernel estimator and set all terms in the sum in Eq. (28) to 0 that would otherwise become irregular. This means that where less then $m_{\min}$ points are available, we assume that the probabilities factorise locally. This results in a bias towards 0 for small band widths. We obtained the most stable results with $m_{\min} = 5$.

As an example of fixed band width kernel estimation of mutual information and transfer entropy, let us revisit the coupled linear stochastic process equation (20). We fix $\gamma = 0.4$ and use rectangular kernels (see Table 1) to approximate the densities using a time series consisting of 10 000 points. The same band width is chosen here for both signals and all components. In Fig. 4, the mutual information $M(X_{i+1}, Y_i)$ and $M(Y_{i+1}, X_i)$, as well as the transfer entropy $T(X_{i+1}|X_i, Y_i)$ and $T(Y_{i+1}|Y_i, X_i)$ are plotted versus the band width ε (log scale). For decreasing band width, all the curves reach plateau values close to the theoretical values. At the lower end of the plateau, finite

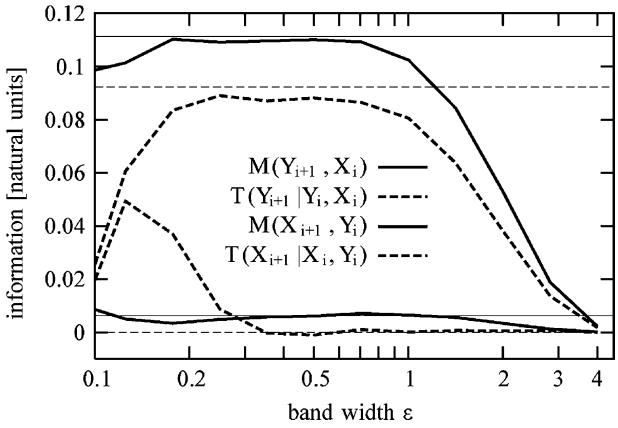

Fig. 4. Mutual information (solid lines) and transfer entropy (dashed lines) versus band width for the unidirectionally coupled linear process equation (20). The two upper curves denote $M(Y_{i+1}, X_i)$ and $T(Y_{i+1}|Y_i, X_i)$, the two lower curves $M(X_{i+1}, Y_i)$ and $T(X_{i+1}|X_i, Y_i)$. Exact values of mutual information and transfer entropy are given as solid and dashed horizontal lines, respectively.

sample fluctuations set in, which are more pronounced for the transfer entropy, which does not actually reach the theoretical value of 0.0923 (dashed line). For $T(Y_{i+1}|Y_i, X_i)$, we find a value 0 within statistical errors over a large range of length scales.

5.2.3.2. Fixed mass. In order to avoid the inhomogeneous statistical precision of the kernel estimator of fixed band width, one can select ε adaptively such that the sums in Eq. (27) etc. assume a fixed value. It is natural to keep the ratio of the ε_j for different j fixed for all points $\tilde{x}$. Although in principle this approach works with any of the kernel functions, its use has been reported only with the rectangular kernel. In that case, we have, e.g.

$$\hat{g}(\tilde{\mathbf{x}}_i^{(k)}) = \alpha^{-1} \frac{M(\tilde{\mathbf{x}}_i^{(k)})}{V(\tilde{\mathbf{x}}_i^{(k)})}, \tag{29}$$

where $M(\tilde{\mathbf{x}}_i^{(k)})$ denotes the number of sample points within a small box, centred at $\tilde{\mathbf{x}}_i^{(k)}$ and $V(\tilde{\mathbf{x}}_i^{(k)}) = \prod_{j=1}^{k} 2\varepsilon_j$ its volume. Again, points closer in time than a de-correlation time t are omitted, which is taken care of by the normalisation constant $\alpha = N - 1 - \min(i + t, N) + \max(i - t, 1)$. With this, we get

$$\hat{T}(X_{+1}|\mathbf{X}^{(k)}, \mathbf{Y}^{(l)}) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})}{V(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})} + \log \frac{M(\mathbf{x}_i^{(k)})}{V(\mathbf{x}_i^{(k)})} - \log \frac{M(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})}{V(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})} - \log \frac{M(x_{i+1}, \mathbf{x}_i^{(k)})}{V(x_{i+1}, \mathbf{x}_i^{(k)})}. \tag{30}$$

Now, there are two distinct ways to go. Either one can select the box size separately for the four log terms, such that all the $M(\cdot)$ equal a fixed number m. This is identical to using Eq. (11) with fixed mass estimators for Shannon entropy. With this approach, the resolution will be finer for the lower dimensional spaces: at the same m, boxes will need to be larger for $M(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)}) = m$ than for $M(\mathbf{x}_i^{(k)}) = m$. Unfortunately, this possibility suffers most from statistical fluctuations and converges very poorly to the continuous transfer entropy.

The second option is to select the edge lengths $\varepsilon_j, j = 1, \ldots, k, k+1, \ldots, k+1+l$ such that $M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)}) = m$ holds, where m is fixed. In order to obtain $M(\mathbf{x}_i^{(k)})$, $M(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$, and $M(x_{i+1}, \mathbf{x}_i^{(k)})$, the box $V(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$ is projected onto the lower dimensional space in order to obtain the corresponding boxes, e.g. $V(\mathbf{x}_i^{(k)})$. Since $M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$ is defined in the highest dimensional space, we have $M(\cdot) \geq m$ for the projections.

Fixed mass formulae are usually used together with their known finite sample correction terms [25,26]. A box that is just large enough to contain m points is systematically smaller than one that contains the probability m/N. This effect is corrected by replacing the $\log M$ terms by the digamma function $\psi(M) = \Gamma'(M)/\Gamma(M)$. However, for mutual information and transfer entropy, the derivation of this term is no longer valid, as pointed out already. We were unable to provide a correction formula that takes the correlations between the counts $M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$, $M(\mathbf{x}_i^{(k)})$, $M(\mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$, and $M(x_{i+1}, \mathbf{x}_i^{(k)})$ into account. If the boxes are selected such that $M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)}) = m$ and the others by projection, we suggest to replace $\log M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$ by $\psi(M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)}))$ and leave the other terms uncorrected.

Empirical results for different possible correction schemes are shown in Fig. 5 for the coupled linear stochastic process equation (20). Again, 10 000 points have been used. We find that without correction (upper panel, solid line), there is a strong positive bias for small masses M/N. If Eq. (11) is used (upper panel, dashed lines), convergence is not very convincing. If all $\log M$ terms are replaced by $\psi(M)$ (lower panel, solid lines, correction (a)), a small positive bias remains and still no plateau is found. If only $M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$ is replaced, the bias is even smaller and negative. We now find a plateau region at the correct values.

For this particular example, we find that the fixed mass and fixed band width approaches yield results of comparable quality. Although it is good to have a choice of methods, we usually prefer a fixed band width algorithm because it is computationally less demanding.

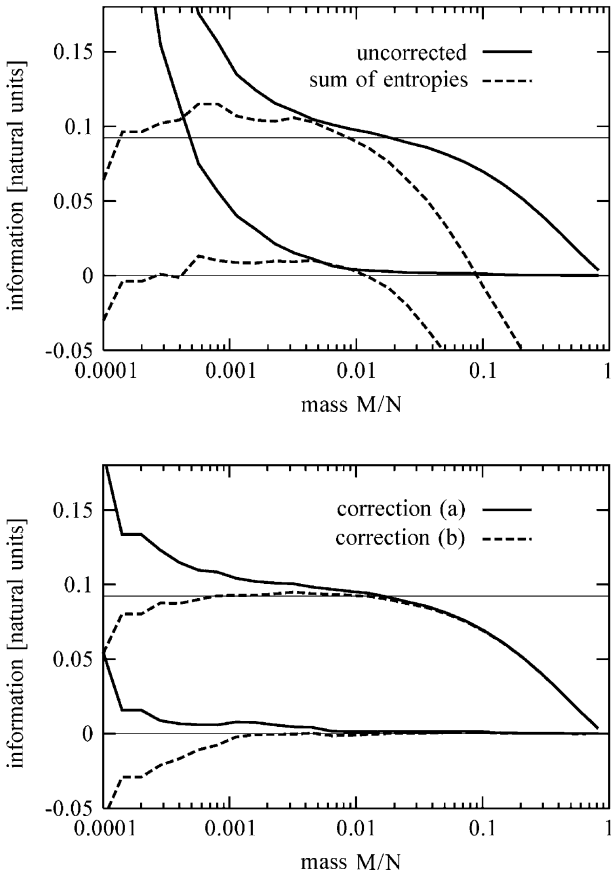

Fig. 5. Fixed mass estimates of transfer entropy $T(Y_{i+1}|Y_i, X_i)$ (upper lines in both panels) and $T(X_{i+1}|X_i, Y_i)$ (lower lines) for the unidirectionally coupled linear process equation (20). Upper panel: results without finite sample correction (solid lines) and using sums of Shannon entropies (dashed lines). Lower panel: results using individual corrections, ignoring correlations ((a), solid lines) and correcting only $\log M(x_{i+1}, \mathbf{x}_i^{(k)}, \mathbf{y}_i^{(l)})$ ((b), dashed lines). Exact values are given as horizontal lines.

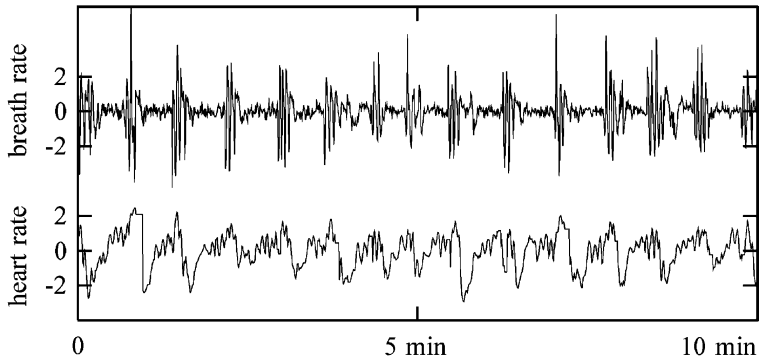

Fig. 6. Bi-variate time series of the breath rate (upper) and instantaneous heart rate (lower) of a sleeping human. The data is sampled at 2 Hz. Both traces have been normalised to zero mean and unit variance.

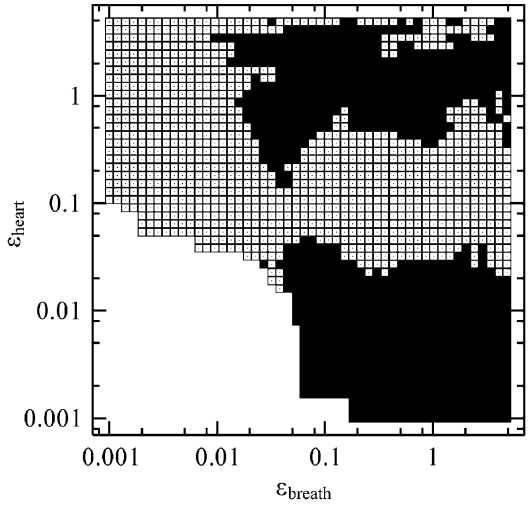

Fig. 7. Alleged dominant direction of information flow as a function of scale. Different band widths $\varepsilon_{\text{heart}}$ and $\varepsilon_{\text{breath}}$ were used in comparing the transfer entropy between the heart and breath rate signals shown in Fig. 6. Scales at which $T_{\text{heart}\to\text{breath}} > T_{\text{breath}\to\text{heart}}$ are marked in black, scales where $T_{\text{breath}\to\text{heart}} > T_{\text{heart}\to\text{breath}}$ are shown as open squares.

5.2.4. An example

As a final example that also illustrates the difficulties of this kind of analysis, let us revisit a problem that was studied in [6] already. There, a bi-variate time series (see Fig. 6) of the breath rate and instantaneous heart rate of a sleeping human suffering from sleep apnoea (samples 2350–3550 of data set B of the Santa Fe Institute time series contest held in 1991 [34]) was used. In [6], time-delayed mutual information was found almost symmetric between both series while the transfer entropy seemed to indicate a stronger flow of information from the heart rate to the breath rate than vice versa over a certain range of length scales ε. For this, both channels had been normalised to unit variance and the same band width was used for both. However, it was pointed out that this is not an entirely satisfactory procedure since the two signals are of different physical nature and cannot be easily compared. Here, we revisit the data set in the light of the results of the previous sections.

In [6], no convergence of either mutual information or transfer entropy was found and the discussion had to be based on finite length scale estimates.[2] In that case, transformation invariance cannot be assumed and the results

[2] Also, the incorrect finite sample correction ignoring correlations was used.

may depend on the scaling of the data sets. This we will demonstrate in the following. Relative scaling of the two signals is equivalent to varying the band widths ε_j between the channels. We have computed the transfer entropy for a range of band widths $\varepsilon_{\text{heart}}$ and $\varepsilon_{\text{breath}}$. Fig. 7 shows regions where $T(X_{i+1}|X_i, Y_i) > T(Y_{i+1}|Y_i, X_i)$ in black and regions where $T(X_{i+1}|X_i, Y_i) < T(Y_{i+1}|Y_i, X_i)$ as open boxes. White regions are those where both values are estimated as 0 due to lack of points.

In our point of view, the case is totally inconclusive due to the lack of convergence of the transfer entropies. An answer to the question which system is following which must not depend on the relative scaling of the measured signals. Therefore, we cannot give such an answer based on the analysis made.

6. Conclusions

We have discussed the recently introduced concept of transfer entropy which is constructed to determine the direction of the information flow between two coupled processes. We have demonstrated that the new quantity is more adequate for this purpose than mutual information or time-delayed mutual information.

The main focus of this paper was on the properties of transfer entropy for continuous processes. Like the mutual information, transfer entropy is invariant against transformations by C^1-diffeomorphisms. This very attractive property is only valid, however, if a convergent estimator of continuous entropy is used. It is usually broken if convergence is not reached.

The most common approach to calculate mutual information is to use a sequence of successively refined partitions. The same can be done for transfer entropy. For a series of refined partitions, coarse grained results converge toward the *true* continuous value for processes with a continuous Lebesgue density. Mutual information converges monotonically from below. We were unable to show the latter property for transfer entropy and we are not sure if it holds. In any case, convergence is not guaranteed when analysing real data. In order to see convergence, small partition elements have to be chosen. But with finite time series, these elements may contain too few data points and estimates of the probabilities are dominated by fluctuations.

Partition-based estimates suffer from a bias due to the serial correlations present in time series data. This bias can be removed easily if one uses an alternative way to estimate mutual information and transfer entropy, the non-parametric density estimation by kernel estimators. Also in this case, convergence requires taking the limit of small band width, which is hampered by the finite number of available data points. Numerically, we find that the results obtained at a fixed band with are of comparable quality to those of the fixed mass approach.

Finally, transfer entropy and mutual information can be calculated if the processes have a known parametric distribution. Then, only the unknown parameters have to be estimated from data, which is usually straightforward. In the case of Gaussian processes, an analytical expression is available which uses the expected value and the covariance matrix. In that case, however, the information theoretic functionals do not reveal any structure beyond the usual linear correlations.

For time series analysis, the major problem is that convergence to the continuous entropies often cannot be achieved. For coarse grained entropies, however, the intuitive interpretations like “common information content”, “amount of information transferred”, or “direction of information flow” have to be used with great care since we cannot allow such concepts to depend on the absolute or relative scales or band widths used in the analysis.

References

[1] M.G. Rosenblum, A.S. Pikovsky, Detecting direction of coupling in interacting oscillators, Phys. Rev. E 64 (2001) 045202(R) .
[2] J. Arnhold, P. Grassberger, K. Lehnertz, C.E. Elger, A robust method for detecting interdependences: application to intracranially recorded EEG, Physica D 134 (1999) 419.

[3] M.G. Rosenblum, A.S. Pikovsky, J. Kurths, Phase synchronisation of chaotic attractors, Phys. Rev. Lett. 76 (1996) 1804.
[4] M. Le Van Quyen, C. Adam, M. Baulac, J. Martinerie, F.J. Varela, Non-linear interdependences of EEG signals during intracranial ictal activities, Brain Res. 792 (1998) 24.
[5] A. Pikovsky, M. Rosenblum, J. Kurths, Synchronization, Cambridge University Press, Cambridge, 2001.
[6] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2000) 461.
[7] P. Billingsley, Ergodic theory and information, Wiley, New York, 1965.
[8] C.E. Shannon, W. Weaver, The Mathematical Theory of Information, University of Illinois Press, Urbana, IL, 1949.
[9] G. Jumarie, Relative Information: Theories and Applications, Springer Series in Synergetics, Vol. 47, Springer, Berlin, 1990.
[10] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Series in Telecommunications, Wiley, New York, 1991.
[11] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[12] D. Prichard, J. Theiler, Generalized redundancies for time series analysis, Physica D 84 (1995) 476.
[13] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134.
[14] G.A. Darbellay, An estimator of the mutual information based on a criterion for independence, Comput. Stat. Data Anal. 32 (1999) 1.
[15] M. Palus, Detecting nonlinearity in multivariate time series, Phys. Lett. A 213 (1996) 138.
[16] B. Pompe, Measuring statistical dependences in a time series, J. Stat. Phys. 73 (1993) 587.
[17] R. Hegger, H. Kantz, L. Matassini, T. Schreiber, Coping with non-stationarity by over-embedding, Phys. Rev. Lett. 84 (2000) 4092.
[18] H. Bauer, Wahrscheinlichkeitstheorie, 4th Edition, Walter de Gruyter, Berlin, 1991.
[19] H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, Cambridge University Press, Cambridge, MA, 1997.
[20] A.N. Kolmogorov, Information theory and the theory of algorithms, Selected Works, Vol. 3, Kluwer, Dordrecht, 1993.
[21] K. Kaneko, Lyapunov analysis and information flow in coupled map lattices, Physica D 23 (1986) 436.
[22] J.A. Vastano, H.L. Swinney, Information transport in spatiotemporal systems, Phys. Rev. Lett. 60 (1988) 1773.
[23] H. Bauer, Maß- und Integrationstheorie, 2nd Edition, Walter de Gruyter, Berlin, 1992.
[24] P. Billingsley, Probability and Measure, Wiley Series in Probability and Mathematical Statistics, 3rd Edition, Wiley, New York, 1995.
[25] P. Grassberger, Finite sample corrections to entropy and dimension estimates, Phys. Lett. A 128 (1988) 369.
[26] P. Grassberger, Generalizations of the Hausdorff dimension of fractal measures, Phys. Lett. A 107 (1985) 101.
[27] H. Herzel, A.O. Schmitt, W. Ebeling, Finite sample effects in sequence analysis, Chaos Solitons Fract. 4 (1994) 97.
[28] M.S. Roulston, Estimating the errors on measured entropy and mutual information, Physica D 125 (1999) 285.
[29] M.S. Roulston, Significance testing of information theoretic functionals, Physica D 110 (1997) 62.
[30] T. Schreiber, Spatio-temporal structure in coupled map lattices: two-point correlations versus mutual information, J. Phys. A 23 (1990) L393.
[31] T. Schreiber, A. Schmitz, Surrogate time series, Physica D 142 (2000) 346.
[32] B.W. Silverman, Density estimation for statistics and data analysis, Monographs on Statistics and Applied Probability, Vol. 26, Chapman & Hall, London, 1986.
[33] J. Theiler, Spurious dimension from correlation algorithms applied to limited time series data, Phys. Rev. A 34 (1986) 2427.
[34] D.R. Rigney, A.L. Goldberger, W. Ocasio, Y. Ichimaru, G.B. Moody, R. Mark, Multi-channel physiological data: description and analysis, in: A.S. Weigend, N.A. Gershenfeld (Eds.), Time Series Prediction: Forecasting the Future and Understanding the Past, Santa Fe Institute Studies in the Science of Complexity, Proc. Vol. XV, Addison-Wesley, Reading, MA, 1993.