# Chapter 8
# Causality in Time Series: Its Detection and Quantification by Means of Information Theory

Kateřina Hlaváčková-Schindler

**Abstract** While studying complex systems, one of the fundamental questions is to identify causal relationships (i. e. which system drives which) between relevant subsystems. In this paper, we focus on information-theoretic approaches for causality detection by means of directionality index based on mutual information estimation. We briefly review the current methods for mutual information estimation from the point of view of their consistency. We also present some arguments from recent literature, supporting the usefulness of the information-theoretic tools for causality detection.

## 8.1 Introduction

During the history of most natural and social sciences, detection and clarification of cause-effect relationships among variables, events or objects have been the fundamental questions. Despite some philosophers of mathematics like B. Russel (75)(1872-1970) tried to deny the existence of the phenomenon ”causality” in mathematics and physics, saying that causal relationships and physical equations are incompatible, calling causality to be ’a word relic” (see i.e. (66)), the language of all sciences, including mathematics and physics, has been using this term actively until now. To advocate the Russell’s view, any exact and sufficiently comprehensive formulation of what is causality is difficult. Causality can be understood in terms of a ”flow” among processes and expressed in mathematical language and mathematically analysed.

The general philosophical definition of causality from the Wikipedia Encyclopedia (90) states: ”The philosophical concept of causality or causation refers to the set of all particular ”causal” or ’cause-and-effect’ relations. Most generally, causation is a relationship that holds between events, objects, variables, or states of affairs. It is usually presumed that the cause chronologically precedes the effect.” Causality expresses a kind of a ’law’ necessity, while probabilities express uncertainty, a lack of regularity. Probability theory seems to be the most used ”mathematical language” of most scientific disciplines using causal modeling, but it seems not to be able to grasp all related questions. In most disciplines, adopting the above definition, the aim is not only to detect a causal relationship but also to measure or quantify the relative

K
Commission for Scientific Visualization, Austrian Academy of Sciences and Donau-City Str. 1, A-1220 Vienna, Austria and Institute of Information Theory and Automation of the Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 18208 Praha 8, Czech Republic e-mail: **katerina.schindler@assoc.oeaw.ac.at**

strengths of these relationships. This can be done by information theory tools. In (42) we provided a detailed overview of the information-theoretic approaches for measuring of a causal influence in multi-variate time series. Here we mainly focus on the methods using mutual information for the computation of the causal directional index and entropy estimation methods. The methods are discussed from the point of view of their consistency properties and for their detailed description we refer the reader to (42). The outline of the paper is the following. Section 8.1.1 presents measures for causality detection. In Section 8.2 we define the basic information-theoretic functionals and from them derived causality measurements. Sections 8.3 and 8.4 briefly present the non-parametric and parametric methods including their consistency properties. Granger causality is discussed in Section 8.5 and conclusion is in Section 8.6.

### *8.1.1 Causality and causal measures*

Most of the earlier research literature attempts to discuss unique causes in deterministic situations, and two conditions are important for deterministic causation (33): (i) necessity: if $X$ occurs, then $Y$ must occur, and (ii) sufficiency: if $Y$ occurs, then $X$ must have occurred. However, deterministic formulation, albeit appealing and analytically tractable, is not in accordance with reality, as no real-life system is strictly deterministic (i.e. its outcomes cannot be predicted with complete certainty). So, it is more realistic if one modifies the earlier formulation in terms of likelihood (i.e. if $X$ occurs, then the likelihood of $Y$ occurring increases). This can be illustrated by a simple statement such as if the oil price increases, the carbon emission does not necessarily decrease, but there is a good likelihood that it will decrease. The probabilistic notion of causality is nicely described by Suppes (1970) as follows: An event $X$ is a cause to the event $Y$ if (i) $X$ occurs before $Y$, (ii) likelihood of $X$ is non zero, and (iii) likelihood of occurring $Y$ given $X$ is more than the likelihood of $Y$ occurring alone. Although this formulation is logically appealing, there are some arbitrariness in practice in categorizing an event (33). Till 1970, the causal modeling was mostly used in social sciences. This was primarily due to a pioneering work by Selltiz et al. (1959) (81) who specified three conditions for the existence of causality:

1. There must be a concomitant covariation between X and Y.
2. There should be a temporal asymmetry or time ordering between the two observed sequences.
3. The covariance between X and Y should not disappear when the effects of any confounding variables (i.e. those variables which are causally prior to both X and Y) are removed.

The first condition implies a correlation between a cause and its effect, though one should explicitly remember that a perfect correlation between two observed variables in no way implies a causal relationship. The second condition is intuitively based on the arrow of time. The third condition is problematic since it requires that one should rule out all other possible causal factors. Theoretically, there are potentially an infinite number of unobserved confounding variables available, yet the set of measured variables is finite, thus leading to indeterminacy in the causal modeling approach. In order to avoid this, some structure is imposed on the adopted modeling scheme which should help to define the considered model. The way in which the structure is imposed is crucial in defining as well as in quantifying causality.
The first definition of causality which could be quantified and measured computationally, yet very general, was given in 1956 by N. Wiener (89): ”For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.”

The introduction of the concept of causality into the experimental practice, namely into analyses of data observed in consecutive time instants, time series, is due to Clive W. J. Granger, the 2003 Nobel prize winner in economy. In his Nobel lecture (34) he recalled the inspiration by the Wiener's work and identified two components of the statement about causality: 1. The cause occurs before the effect; 2.The cause contains information about the effect that is unique, and is in no other variable.
As Granger put it, a consequence of these statements is that the causal variable can help to forecast the effect variable after other data has been first used (34). This restricted sense of causality, referred to as *Granger causality*, GC thereafter, characterizes the extent to which a process $X_t$ is leading another process $Y_t$, and builds upon the notion of incremental predictability. It is said that the *process $X_t$ Granger causes another process $Y_t$* if future values of $Y_t$ can be better predicted using the past values of $X_t$ and $Y_t$ rather then only past values of $Y_t$. The standard test of GC developed by Granger (31) is based on a linear regression model

$$Y_t = a_o + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \sum_{k=1}^{L} b_{2k} X_{t-k} + \xi_t, \tag{8.1}$$

$\xi_t$ are uncorrelated random variables with zero mean and variance $\sigma^2$, L is the specified number of time lags, and $t = L+1,\ldots,N$. The null hypothesis that $X_t$ does not Granger cause $Y_t$ is supported when $b_{2k} = 0$ for $k = 1,\ldots,L$, reducing Eq. (8.1) to

$$Y_t = a_o + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \tilde{\xi}_t. \tag{8.2}$$

This model leads to two well-known alternative test statistics, the Granger-Sargent and the Granger-Wald test. The Granger-Sargent test is defined as

$$GS = \frac{(R_2 - R_1)/L}{R_1/(N-2L)}, \tag{8.3}$$

where $R_1$ is the residual sum of squares in (8.1) and $R_2$ is the residual sum of squares in (8.2). The GS test statistic has an F-distribution with $L$ and $N-2L$ degrees of freedom. On the other hand, the Granger-Wald test is defined as

$$GW = N \frac{\widehat{\sigma}^2_{\tilde{\xi}_t} - \widehat{\sigma}^2_{\xi_t}}{\widehat{\sigma}^2_{\xi_t}}, \tag{8.4}$$

where $\widehat{\sigma}^2_{\tilde{\xi}_t}$ is the estimate of the variance of $\tilde{\xi}_t$ from model (8.2) and $\widehat{\sigma}^2_{\xi_t}$ is the estimate of the variance of $\xi_t$ from model (8.1). The GW statistic follows the $\chi^2_L$ distribution under the null hypothesis of no causality.

This linear framework for measuring and testing causality has been widely applied not only in economy and finance (see Geweke (30) for a comprehensive survey of the literature), but also in diverse fields of natural sciences such, where specific problems of multichannel electroencephalogram recordings were solved by generalizing the Granger causality concept to multivariate case (13). Nevertheless, the limitation of the present concept to linear relations required further generalizations.

Recent development in nonlinear dynamics (1) evoked lively interactions between statistics and econometrics on one side, and physics and other natural sciences on the other side. In the field of economy, Baek and Brock (9) and Hiemstra and Jones (39) proposed a nonlinear extension of the Granger causality concept. Their non-parametric dependence estimator is based on so-called correlation integral, a probability distribution and entropy estimator, developed by physicists Grassberger and Procaccia in the field of nonlinear dynamics and deterministic chaos as a characterization tool of chaotic attractors (35). A non-parametric

approach to non-linear causality testing, based on non-parametric regression, was proposed by Bell et al. (**?** ). Following Hiemstra and Jones (39), Aparicio and Escribano (6) succinctly suggested an information-theoretic definition of causality which include both linear and nonlinear dependence.

In physics and nonlinear dynamics, a considerable interest recently emerged in studying cooperative behavior of coupled complex systems (68; 14). Synchronization and related phenomena were observed not only in physical, but also in many biological systems. Examples include the cardio-respiratory interaction (76; 62) and the synchronization of neural signals (72; 57). In such physiological systems it is not only important to detect synchronized states, but also to identify drive-response relationships and thus the causality in evolution of the interacting (sub)systems. Schiff et al. (77) and Quyen et al. (72) used ideas similar to those of Granger, however, their cross-prediction models utilize zero-order nonlinear predictors based on mutual nearest neighbors. A careful comparison of these two papers (77; 72) reveals how complex is the problem of inferring causality in nonlinear systems. The authors of the two papers use contradictory assumptions for interpreting the differences in prediction errors of mutual predictions, however, both teams were able to present numerical examples in which their approaches apparently worked.

While the latter two papers use the method of mutual nearest neighbors for mutual prediction, Arnhold et al. (8) proposed asymmetric dependence measures based on averaged relative distances of the (mutual) nearest neighbors. As pointed out by Quian Quiroga et al. and by Schmitz (73; 78), these measures, however, might be influenced by different dynamics of individual signals and different dimensionality of the underlying processes, rather than by asymmetry in coupling.

Another nonlinear extension of the Granger causality approach was proposed by Chen et al. (18) using local linear predictors. An important class of nonlinear predictors are based on so-called radial basis functions (16) which were used for nonlinear parametric extension of the Granger causality concept (4). A nonparametric method for measuring causal information transfer between systems was proposed by Schreiber (80). His *transfer entropy* is designed as a Kullback-Leibler distance (Eq. (8.14) in Sec. 8.2.1) of transition probabilities. This measure is in fact an information-theoretic functional of probability distribution functions.

Paluš et al. (57) proposed to study synchronization phenomena in experimental time series by using the tools of information theory. Mutual information, an information-theoretic functional of probability distribution functions, is a measure of general statistical dependence. For inferring causal relation, conditional mutual information can be used. It was shown that, with proper conditioning, the Schreiber's transfer entropy (80) is equivalent to the conditional mutual information (57). The latter, however, is a standard measure of information theory.

Turning our attention back to econometrics, we can follow further development due to Diks and DeGoede (22). They again applied a nonparametric approach to nonlinear Granger causality using the concept of correlation integrals (35) and pointed out the connection between the correlation integrals and information theory. Diks and Panchenko (23) critically discussed the previous tests of Hiemstra and Jones (39). As the most recent development in economics, Baghli (10) proposes information-theoretic statistics for a model-free characterization of causality, based on an evaluation of conditional entropy. The nonlinear extension of the Granger causality based the information-theoretic formulation has found numerous applications in various fields of natural and social sciences. Let us mention just a few examples. Schreiber's transfer entropy was used in physiology and neurophysiology (45). Paluš et al. (57) applied their conditional mutual information based measures in analyses of electroencephalograms of patients suffering from epilepsy. Other applications of the conditional mutual information in neurophysiology are due to Hinrichs et al. (40). Causality or coupling directions in multimode laser dynamics is another field where the conditional mutual information was applied (56). Having reviewed the relevant literature, we can state that the information-theoretic approach to the Granger causality plays an important, if not a dominant role in analyses of causal relationships in nonlinear systems.

## 8.2 Information theory as a tool for causality detection

### 8.2.1 Definitions of basic information theoretic functionals

We begin with the definition of differential entropy for a continuous random variable as it was introduced in 1948 by Shannon (82). Let $X$ be a random vector taking values in $R^d$ with probability density function (pdf) $p(x)$, then its **differential entropy** is defined by

$$H(x) = -\int p(x)\ln p(x)dx, \tag{8.5}$$

where ln is natural logarithm. We assume that $H(x)$ is well-defined and finite. Let $S$ be a discrete random variable having possible values $s_1,\ldots,s_m$, each with corresponding probability $p_i = p(s_i), i = 1,\ldots m$. The average amount of information gained from a measurement that specifies one particular value $s_i$ is given by **entropy** $H(S)$:

$$H(S) = -\sum_{i=1}^{m} p_i \ln p_i. \tag{8.6}$$

More general term of entropy for which is Shannon differential entropy a special case, is **Rényi entropy**, defined for a continuous case as (74)

$$H_\alpha(x) = \frac{1}{1-\alpha}\int \ln^\alpha p(x)dx \tag{8.7}$$

and for the discrete case

$$H_\alpha(S) = \frac{1}{1-\alpha}\ln\sum_{i=1}^{n} p_i^\alpha, \tag{8.8}$$

where $\alpha > 0$, $\alpha \neq 1$. As $\alpha \to 1$, $H_\alpha(x)$ converges to $H(x)$ (or $H_\alpha(S)$ converges to $H(S)$), which is Shannon entropy. The **joint entropy** $H(X,Y)$ of two discrete random variables $X$ and $Y$ is

$$H(X,Y) = -\sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} p(x_i,y_j)\ln p(x_i,y_j) \tag{8.9}$$

where $p(x_i,y_j)$ denotes the joint probability that $X$ is in state $x_i$ and $Y$ in state $y_j$. In general, the joint entropy may be expressed in terms of **conditional entropy** $H(X|Y)$ as follows

$$H(X,Y) = H(X|Y) + H(Y), \text{ where} \tag{8.10}$$

$$H(X|Y) = -\sum_{i=1}^{m_X}\sum_{j=1}^{m_Y} p(x_i,y_j)\ln p(x_i|y_j) \tag{8.11}$$

and $p(x_i|y_j)$ denotes the conditional probability. The **mutual information** $I(X,Y)$ between two random variables $X$ and $Y$ is then defined as (82)

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{8.12}$$

It reflects the mutual reduction in uncertainty of one by knowing the other variable. This measure is nonnegative since $H(X,Y) \leq H(X)+H(Y)$. The equality holds if and only if $X$ and $Y$ are statistically independent. The **conditional mutual information** (82) between random variables $X$ and $Y$ given $Z$ is defined as

$$I(X,Y|Z) = H(X|Z)+H(Y|Z)-H(X,Y|Z). \tag{8.13}$$

For $Z$ independent of $X$ and $Y$ we have $I(X,Y|Z) = I(X,Y)$. The **Kullback-Leibler divergence** (KLD, also called relative entropy or cross-entropy), introduced by Kullback and Leibler (50), is an alternative approach to mutual information. $K(p,p^0)$ between two probability distributions $\{p\}$ and $p^0$ is

$$K(p,p^0) = \sum_{i=1}^{m} p_i \ln(\frac{p_i}{p_i^0}). \tag{8.14}$$

It can be interpreted as the information gain when an initial probability distribution $p^0$ is replaced by a final distribution $p$. This entropy is however not symmetric and therefore not a distance in the mathematical sense. The KLD is always nonnegative and is zero iff the distributions $p$ and $p^0$ are identical. Mutual information is the Kullback-Leibler divergence of the product $P(X)P(Y)$ of two marginal probability distributions from the joint probability distribution $P(X,Y)$, see i.e. (29). So we can look at the results about Kullback-Leibler entropy as if they were applied to mutual information.

### *8.2.2 Coarse-grained entropy and information rates*

A considerable amount of approaches to inferring causality from experimental time series have their roots in studies of synchronization of chaotic systems. A. N. Kolmogorov, who introduced the theoretical concept of classification of dynamical system by information rates (46), was inspired by information theory and together with Y.G. Sinai generalized the notion of entropy of an information source (46; 84). Paluš (59) concentrated on attributes of dynamical systems studied in the ergodic theory, such as mixing and generating partitions, and demonstrated how they were reflected in the behaviour of information-theoretic functionals estimated from chaotic data. In order to obtain an asymptotic entropy estimate of an $m-$ dimensional dynamical system, large amounts of data are necessary (59). To avoid this, Paluš (59) proposed to compute "coarse-grained entropy rates" (CER's) as relative measures of "information creation" and of regularity and predictability of studied processes.

Let $\{x(t)\}$ be a time series considered as a realization of a stationary and ergodic stochastic process $\{X(t)\}$, $t = 1,2,3,\ldots$. We denote $x(t)$ as $x$ and $x(t+\tau)$ as $x_\tau$ for simplicity. To define the simplest form of CER, we compute the mutual information $I(x;x_\tau)$ for all analyzed datasets and find such $\tau_{max}$ that for $\tau' \geq \tau_{max}$: $I(x;x_{\tau'}) \approx 0$ for all the data sets. Then we define a **norm of the mutual information**

$$||I(x;x_\tau)|| = \frac{\Delta\tau}{\tau_{max}-\tau_{min}+\Delta\tau} \sum_{\tau=\tau_{min}}^{\tau_{max}} I(x;x_\tau) \tag{8.15}$$

with $\tau_{min} = \Delta\tau = 1$ sample as a usual choice. The CER $h^1$ is then defined as $h^1 = I(x,x_{\tau_0}) - ||I(x;x_\tau)||$. Since usually $\tau_0 = 0$ and $I(x;x) = H(X)$ which is given by the marginal probability distribution $p(x)$, the sole quantitative descriptor of the underlying dynamics is the mutual information norm (8.15). Paluš et al. (57)

called this descriptor the **coarse-grained information rate** (CIR) of the process $\{X(t)\}$ and denoted by $i(X)$.

Now, consider two time series $\{x(t)\}$ and $\{y(t)\}$ regarded as realizations of two processes $\{X(t)\}$ and $\{Y(t)\}$ which represent two possibly linked (sub) systems. These two systems can be characterized by their respective CIR's $i(X)$ and $i(Y)$. In order to characterize an interaction of the two systems, in analogy with the above CIR, Paluš et al. (57) defined their **mutual coarse-grained information rate** (MCIR) by

$$i(X,Y) = \frac{1}{2\tau_{max}} \sum_{\tau=-\tau_{max}}^{\tau_{max};\tau\neq 0} I(x;y_\tau). \tag{8.16}$$

Due to the symmetry properties of $I(x;y_\tau)$ is the mutual CIR $i(X,Y)$ symmetric, i.e., $i(X,Y) = i(Y,X)$. Assessing the direction of coupling between the two systems, i.e., causality in their evolution, we ask how is the dynamics of one of the processes, say $\{X\}$, influenced by the other process $\{Y\}$. For the quantitative answer to this question, Paluš et al. (57) proposed to evaluate the **conditional coarse-grained information rate** CCIR $i_0(X|Y)$ of $\{X\}$ given $\{Y\}$:

$$i_0(X|Y) = \frac{1}{\tau_{max}} \sum_{\tau=1}^{\tau_{max}} I(x;x_\tau|y), \tag{8.17}$$

considering the usual choice $\tau_{min} = \Delta\tau = 1$ sample. For independent variables we have $i_0(X|Y) = i(X)$ for $\{X\}$ independent of $\{Y\}$, i.e., when the two systems are uncoupled. In order to have a measure which vanishes for an uncoupled system (although then it can acquire both positive and negative values), Paluš et al. (57) define

$$i(X|Y) = i_0(X|Y) - i(X). \tag{8.18}$$

For another approach to a directional information rate, let us consider the mutual information $I(y;x_\tau)$ measuring the average amount of information contained in the process $\{Y\}$ about the process $\{X\}$ in its future $\tau$ time units ahead ($\tau$-future thereafter). This measure, however, could also contain an information about the $\tau$-future of the process $\{X\}$ contained in this process itself if the processes $\{X\}$ and $\{Y\}$ are not independent, i.e., if $I(x;y) > 0$. In order to obtain the "net" information about the $\tau$-future of the process $\{X\}$ contained in the process $\{Y\}$, we need the conditional mutual information $I(y;x_\tau|x)$. Next, we sum $I(y;x_\tau|x)$ over $\tau$ as above

$$i_1(X,Y|X) = \frac{1}{\tau_{max}} \sum_{\tau=1}^{\tau_{max}} I(y;x_\tau|x); \tag{8.19}$$

In order to obtain the "net asymmetric" information measure, we subtract the symmetric MCIR (8.16):

$$i_2(X,Y|X) = i_1(X,Y|X) - i(X,Y). \tag{8.20}$$

Using a simple manipulation, we find that $i_2(X,Y|X)$ is equal to $i(X|Y)$ defined in Eq. (8.18). By using two different ways for definition of a directional information rate, Paluš et al. (57) arrived to the same measure which they denoted by $i(X|Y)$ and called the **coarse-grained transinformation rate** (CTIR) of $\{X\}$ given $\{Y\}$. It is the average rate of the net amount of information "transferred" from the process $\{Y\}$ to the process $\{X\}$ or, in other words, the average rate of the net information flow by which the process $\{Y\}$ influences the process $\{X\}$.

Using several numerical examples of coupled chaotic systems, Paluš et al. (57) demonstrated that the CTIR is able to identify the coupling directionality from time series measured in coupled, but not yet fully

synchronized systems. As a practical application, CTIR was used in analyses of electroencephalograms of patients suffering from epilepsy. Causal relations between EEG signals measured in different parts of the brain were identified. Paluš et al. demonstrated suitability of the conditional mutual information approach for analyzing causality in cardio-respiratory interaction (57).

### *8.2.3 Conditional mutual information and transfer entropy*

The principal measure, used by Paluš et al. (57) for inferring causality relations, i.e., the directionality of coupling between the processes $\{X(t)\}$ and $\{Y(t)\}$, is the conditional mutual information $I(y;x_\tau|x)$ and $I(x;y_\tau|y)$. If the processes $\{X(t)\}$ and $\{Y(t)\}$ are substituted by dynamical systems evolving in measurable spaces of dimensions $m$ and $n$, respectively, the variables $x$ and $y$ in $I(y;x_\tau|x)$ and $I(x;y_\tau|y)$ should be considered as $n-$ and $m-$dimensional vectors. In experimental practice, however, usually only one observable is recorded for each system. Then, instead of the original components of the vectors $\mathbf{X}(t)$ and $\mathbf{Y}(t)$, the time delay embedding vectors according to Takens (85) are used. Then, back in time-series representation, we have

$$I\big(\mathbf{Y}(t);\mathbf{X}(t+\tau)|\mathbf{X}(t)\big) = \tag{8.21}$$

$$I\Big(\big(y(t),y(t-\rho),\ldots,y(t-(m-1)\rho)\big);x(t+\tau)|\big(x(t),x(t-\eta),\ldots,x(t-(n-1)\eta)\big)\Big),$$

where $\eta$ and $\rho$ are time lags used for the embedding of systems $\mathbf{X}(t)$ and $\mathbf{Y}(t)$, respectively. For simplicity, only the information about one component $x(t+\tau)$ in the $\tau-$future of the system $\mathbf{X}(t)$ is used. The opposite CMI $I\big(\mathbf{X}(t);\mathbf{Y}(t+\tau)|\mathbf{Y}(t)\big)$ is defined in the full analogy. Exactly the same formulation can be used for Markov processes of finite orders $m$ and $n$. Using the idea of finite-order Markov processes, Schreiber (80) introduced a measure quantifying causal information transfer between systems evolving in time, based on appropriately conditioned transition probabilities. Assuming that the system under study can be approximated by a stationary Markov process of order $k$, the transition probabilities describing the evolution of the system are $p(i_{n+1}|i_n,...,i_{n-k+1})$. If two processes $I$ and $J$ are independent, then the generalized Markov property

$$p(i_{n+1}|i_n,...,i_{n-k+1}) = p(i_{n+1}\,|\,i_n^{(k)},j_n^{(l)}), \tag{8.22}$$

holds, where $i_n^{(k)} = (i_n,...,i_{n-k+1})$ and $j_n^{(l)} = (j_n,...,j_{n-l+1})$ and $l$ is the number of conditioning state from process $J$. Schreiber proposed using the KLD (8.14) to measure the deviation of the transition probabilities from the generalized Markov property (8.22) and got the definition

$$T_{J\to I} = \sum p(i_{n+1},i_n^{(k)},j_n^{(l)}) \ln \frac{p(i_{n+1}|i_n^{(k)},j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})}, \tag{8.23}$$

denoted as *transfer entropy*. It can be understood as the excess amount of bits that must be used to encode the information of the state of the process by erroneously assuming that the actual transition probability distribution function is $p(i_{n+1}|i_n^{(k)})$, instead of $p(i_{n+1}|i^{(k)},j_n^{(l)})$. It was shown (for example in (42)) that the transfer entropy is in fact an equivalent expression for the conditional mutual information.

### 8.2.4 Comparison of coarse grained measures and two deterministic measures for causality detection in bivariate time series

A good causality detector in time series should have a low rate of false detections. Paluš and Vejmelka (63) experimentally analyzed causality detection for bivariate time series by coarse grained measures (CTIR, defined by (8.20)) and compared it to two common deterministic approaches. Numerous examples demonstrated what problems can appear in inference of causal relationship. This to the date unique comparative work on information-theoretic causality detectors to the deterministic ones definitely deserves our attention. Three approaches were compared. In all cases, the driving, autonomous system is denoted by $X$, and the driven, response system by $Y$. As Paluš et al. in (57) explain, the direction of coupling can be inferred from experimental data only when the underlying systems are coupled, but not yet synchronized. The CTIR defined by formula (8.20) was compared to the method from Le Van Quyen (72) and the method from Arnhold et al and Quian Quiroga (8) and (73) (belonging to the methods discussed in Section 8.3.4). The second method is based on cross-prediction using the idea of mutual neighbors. A neigborhood size $\delta$ is given. Considering a map from $X$ to $Y$, a prediction is made for the value of $y_{n+1}$ one step ahead using the formula

$$\widehat{y}_{n+1} = \frac{1}{|V_\delta(X_n)|} \sum_{j:X_j \in V_\delta(X_n)} y_{j+1}. \tag{8.24}$$

The volume $V_\delta(X_n) = \{X_{n'} : |X_{n'} - X_n| < \delta\}$ is $\delta$ neighborhood of $X_n$ and $|V_\delta(X_n)|$ denotes the number of points in the neighborhood. Using data rescaled to the zero mean and the unit variance, the authors define a crosspredictability index by subtracting the root-mean-square prediction error from one

$$P(X \to Y) = 1 - \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\widehat{y}_{n+1} - y_{n+1})^2}, \tag{8.25}$$

measuring how system $X$ influences the future of system $Y$.

The third method from Arnhold et al. and Quian Quiroga (8) and (73) uses mean square distances instead of the cross-predictions in order to quantify the closeness of points in both spaces. The time-delay embedding is first constructed in order to obtain state space vectors $X$ and $Y$ for both time series $\{x_i\}$ and $\{y_i\}$, respectively and then the mean squared distance to $k$ nearest neighbors is defined for each $X$ as

$$R_n^{(k)}(X) = \frac{1}{k}\sum_{j=1}^{k} |X_n - X_{r_{n,j}}|^2, \tag{8.26}$$

where $r_{n,j}$ the index of the $j$-th nearest neighbor of $X_n$. The $Y$-conditioned squared mean distance is defined by replacing the nearest neighbors of $X_n$ by the equal time partners of the nearest neighbors of $Y_n$ as

$$R_n^{(k)}(X|Y) = \frac{1}{k}\sum_{j=1}^{k} |X_n - X_{s_{n,j}}|^2, \tag{8.27}$$

where $s_{n,j}$ denotes the index of the $j-th$ nearest neighbor of $Y_n$. Then the asymmetric measure

$$S^{(k)}(X|Y) = \frac{1}{N}\sum_{j=1}^{N} \frac{R_n^{(k)}(X)}{R_n^{(k)}(X|Y)} \tag{8.28}$$

should reflect the interdependence in the sense that closeness of the points in $Y$ implies closeness of their equal time partners in $X$ and the values of $S^{(k)}(X|Y)$ approach to one, while, in the case of $X$ independent of $Y$ , $S^{(k)}(X|Y) \ll 1$. The quantity $S^{(k)}(Y|X)$ measuring the influence of $X$ on $Y$ is defined in full analogy.

These three measures were tested on the examples of the Rössler system driving the Lorenz system and for the unidirectionally coupled Henon system and then on unidirectionally coupled Rössler systems. Neither the cross-predictability, nor the mutual nearest neighbours statistics gave consistent results when using three different examples of unidirectionally coupled systems. Only the coarse-grained transinformation rate correctly identified the direction of the causal influence in the above three examples as well as in many other systems of different origins (tested in other works from the authors). In the above mentioned examples of unidirectionally coupled systems, the used measures were generally non-zero in both directions even before the systems became synchronized and comparison of the values of such measures did not always reflect the true causality given by the unidirectional coupling of the systems. The intuitively understandable implication that the lower prediction error (better predictability) implies the stronger dependence cannot be in general applied to nonlinear systems. When the coupling of the systems is weaker than what is necessary for the emergence of synchronization, as used in the above examples, any smooth deterministic function between the states of the systems does not have to exist yet. However, there is already some statistical relation valid on the coarse-grained description level. Although the deterministic quantities are based on the existence of a smooth functional relation, when estimated with finite precision they usually give nonzero values influenced not only by the existing statistical dependence but also by the properties of the systems other then the coupling. Therefore it is necessary to use quantities proposed for measuring statistical dependence such as information-theoretic measures which have solid mathematical background. Conditional mutual information vanishes in the uncoupled direction in the case of unidirectional coupling so that the causal direction can be identified by its statistically significant digression from zero, while in the uncoupled direction it does not cross the borders of a statistical zero. From this respect has the CTIR based causality detector a special position in the comparison to the deterministic ones. Factors and influences, which can lead to either decreased test sensitivity or to false causality detections, were identified and concrete remedies proposed in (63) in order to perform tests with high sensitivity and low rate of false positive results.

### *8.2.5 Classification and criteria for methods for entropy estimation*

The key problem for causality detection by means of conditional mutual information is to have a ”good” estimator of mutual information. Most entropy estimators in the literature, which are designed for multidimensional spaces, can be applied to mutual information estimation. In the following, we adopt mathematical criteria for evaluation of the entropy estimators from Beirlant et al. (11).

### *8.2.6 Conditions and criteria*

If for the identically independent distributed (i.i.d.) sample $X_1, \ldots, X_n$, $H_n$ is an estimate of $H(f)$, then the following types of consistencies can be considered:
**Weak consistency**: $\lim_{n\to\infty} H_n = H(f)$ in probability; **Mean square consistency**: $\lim_{n\to\infty} E(H_n - H(f))^2 = 0$; **Strong (universal) consistency**: $\lim_{n\to\infty} H_n = H(f)$ a.s. (almost sure); **Slow-rate convergence**: $\limsup_{n\to\infty} \frac{E|H_n - H|}{a_n} = \infty$ for any sequence of positive numbers $\{a_n\}$ converging to zero; **Root-$n$ consistency** re-

sults are either of form of **asymptotic normality**, i.e. $\lim_{n\to\infty} n^{1/2}(H_n - H(f)) = N(0,\sigma^2)$ convergence in distribution, of $L_2$ **rate of convergence**: $\lim_{n\to\infty} nE(H_n - H(f))^2 = \sigma^2$ or for the **consistency in** $L_2$, $\lim_{n\to\infty} E(H_n - H(f))^2 = 0$.

The conditions on the underlying density $f$ are: **Smoothness conditions**:$(S1)$ $f$ is continuous. $(S2)$ $f$ is $k$ times differentiable. **Tail conditions**: $(T1)$ $H([X]) < \infty$, where $[X]$ is the integer part of $X$. $(T2)$ $\inf_{f(x)>0} f(x) > 0$. **Peak conditions**: $(P1)$ $\int f(\ln f)^2 < \infty$. (This is also a mild tail condition.) $(P2)$ $f$ is bounded.

Many probability distributions in statistics can be characterized as having maximum entropy and can be generally characterized by Kagan-Linnik-Rao theorem ((44)). When dealing with the convergence properties of estimates, one needs the following definitions. **Asymptotically consistent** estimator means that the series of the approximants converge in infinity to the function to be approximated (see i.e. (11)). **Asymptotically unbiased** estimator is that one which is unbiased in the limit.

In (42) we reviewed current methods for entropy estimation. Most of the methods were originally motivated by other questions than detection of causality: by learning theory questions, or by nonlinear dynamics applications. Many of them, although accurate in one or two dimension, become inapplicable in higher dimensional spaces (because of their computational complexity). Here we discuss these methods from their consistency point of view.

## 8.3 Non-parametric entropy estimators

### *8.3.1 Plug-in estimates*

Plug-in estimates are based on a consistent density estimate $f_n$ of $f$ such that $f_n$ depends on $X_1,\ldots,X_n$. Their name "plug-in" was introduced by Silverman (83). A consistent probability density function estimator is substituted into the place of the pdf of a functional. The most used plug-in estimators are integral estimators, resubstitution estimates, splitting date estimates and cross-validation estimates. Strong consistency of integral estimators was proven by Dmitriev and Tarasenko in Ref. (24) and by Prasaka Rao in Ref. (70). The resubstitution estimates have the mean square consistency which was proven by Ahmad and Lin (2). Splitting data estimate have under some mild tail and smoothness condition on $f$ strong consistency for general dimension $d$ (37). Ivanov and Rozkova showed strong consistency of cross-validation estimates (43). Convergence properties of discrete plug-in estimators were studied by Antos and Kontoyiannis (5) in a more general scope. They proved that for additive functionals, including the cases of the mean, entropy, Rényi entropy and mutual information, satisfying some mild conditions, the plug-in estimates are universally consistent and consistent in $L_2$ and he $L_2$-error of the plug-in estimate is of order $O(\frac{1}{n})$. For discrete estimators, the convergence results obtained by Antos and Kontoyiannis (5) are in agreement with the convergence results of the all above mentioned plug-in methods. On the other hand, for a wide class of other functionals, including entropy, it was shown that the universal convergence rates cannot be obtained for any sequence of estimators. Therefore, for positive rate-of-convergence results, additional conditions need to be placed on the class of considered distributions.

## *8.3.2 Entropy estimates based on the observation space partitioning*

### 8.3.2.1 Fixed partitioning

These estimators divide the observation space into a set of partitions. The partition is generated either directly or recursively (iteratively). The algorithms employ a fixed scheme independent of the data distribution or an adaptive scheme which takes the actual distribution of the data into account. The most widely used methods with fixed partitioning are classical histogram methods, where the approximation of the probability distributions $p(x_i, y_j)$, $p(x_i)$ and $p(y_j)$ is by a histogram estimation (17). These methods work well only up to three scalars. An insufficient amount of data, occurring especially in higher dimensions, leads to a limited occupancy of many histogram bins giving incorrect estimations of the probability distributions and consequently leads to heavily biased, usually overestimated values of mutual information. Consistency of histogram methods was analysed by Lugosi and Nobel in (52), who presented general sufficient conditions for the almost sure $L_1$-consistency of multivariate histogram density estimates based on data-dependent partitions. Analogous conditions guarantee the almost-sure risk consistency of histogram classification schemes based on data-dependent partitions.

### 8.3.2.2 Adaptive partitioning

**Marginal equiquantization**
Any method for computation of mutual information based on partitioning of data space is always connected with the problem of quantization, i.e. a definition of finite-size boxes covering the state (data) space. The probability distribution is then estimated as relative frequencies of the occurrence of data samples in particular boxes (the histogram approach). A naive approach to estimate the mutual information of continuous variables would be to use the finest possible quantization, e.g., given by a computer memory or measurement precision. One must however keep in mind that a finite number $N$ of data samples is available. Hence, using a quantization that is too fine, the estimation of entropies and mutual information can be heavily biased - we say that the data are overquantized.

As a simple data adaptive partitioning method, Paluš (60; 58) used a simple box-counting method with marginal equiquantization. The marginal boxes are not defined equidistantly but so that there is approximately the same number of data points in each marginal bin. The choice of the number of bins is, however, crucial. In Ref. (58) Paluš proposed that computing the mutual information $I^n$ of $n$ variables, the number of marginal bins should not exceed the $n+1$-st root of the number of the data samples, i.e. $q \leq \sqrt[n+1]{N}$. The equiquantization method effectively transforms each variable (in one dimension) into a uniform distribution, i.e. the individual (marginal) entropies are maximized and the MI is fully determined by the value of the joint entropy of the studied variable. This type of MI estimate, even in its coarse-grained version, is invariant against any monotonous (and nonlinear) transformation of the data (61). Due to this property, MI, estimated using the marginal equiquantization method, is useful for quantifying dependence structures in data as well as for statistical tests for nonlinearity which are robust against static nonlinear transformations of the data (58).

Darbellay and Vajda (20) demonstrated that MI can be approximated arbitrarily closely in probability and proved the weak consistency. Their method was experimentally compared to maximum-likelihood estimators (Sec. 8.3.6). The partitioning scheme used by Darbellay and Vajda (20) was originally proposed by Fraser and Swinney (28) and in physics literature is referred to as the Fraser-Swinney algorithm, while in the information-theoretic literature as the Darbellay-Vajda algorithm.

### 8.3.3 Ranking

Pompe (69) proposed an estimator of dependencies of a time series based on second order Rényi entropy. Pompe noticed that if the time series is uniformly distributed, some of the desirable properties of Shannon entropy can be preserved for the second order Rényi entropy. Moreover, the second order Rényi entropy can be effectively estimated using the Grassberger-Procaccia-Takens Algorithm (GPTA) (35). The idea of Pompe's entropy is in finding a transformation of an arbitrarily distributed time series to a uniform distribution and is accomplished by sorting the samples using some common fast sorting algorithm. There are no consistency results (even in their weakest form) known.

### 8.3.4 Estimates of entropy and mutual information based on nearest neighbor search

Estimators of Shannon entropy based on k-nearest neighbor search in one dimensional spaces were studied in statistics already almost 50 years ago by Dobrushin (25) but they cannot be directly generalized to higher dimensional spaces. For general multivariate densities, the nearest neighbor entropy estimate is defined as the sample average of the algorithms of the normalized nearest neighbor distances plus the Euler constant. Under the condition (P1) introduced in Section 8.2.6, Kozachenko and Leonenko (47) proved the mean square consistency for general $d \geq 1$. Tsybakov and van der Meulen (86) showed root-n rate of convergence for a truncated version of $H_n$ in one dimension for a class of densities with unbounded support and exponential decreasing tails, such as the Gaussian density.

Leonenko et al. (51) studied a class of $k$-nearest-neighbor-based Rényi estimators for multidimensional densities. It was shown that Rényi entropy of any order can be estimated consistently with minimal assumptions on the probability density. For Shannon entropy, and for any $k > 0$ integer, the expected value of the $k$-nearest neighbor estimator (including the KSG algorithm described below) converges with the increasing size of data set $N$ to infinity to the entropy of $f$ if $f$ is a bounded function (asymptotical unbiasedness). For any $k > 0$ integer, the $k$-nearest neighbor estimator converges for the Euclidean metric ($L_2$ rate of convergence), with the increasing size of data set $N$ to infinity, to the entropy of $f$ if $f$ is a bounded function (consistency).

An improvement of KL algorithm for using in higher dimensions was proposed by Kraskov, Stögbauer and Grassberger (KSG) (48). The estimator differs from the KL that it uses different distance scales in the joint and marginal spaces. Any consistency results are not known. A nearest neighbor approach to estimate Kullback-Leibler divergence was studied in (51) and by Wang et al. in (88)and its asymptotical consistency was proven.

### 8.3.5 Estimates based on learning theory methods

#### 8.3.5.1 Motivated by signal processing problems

Entropy and mutual information are often used as a criterion in learning theory. Entropy as a measure of dispersion is applied in many other areas, in control, search, or in the area of neural networks and supervised learning, i.e. Refs. (67), (27), (79). Many of the developed methods belong as well to non-parametric plug-in estimators. Learning theory is interested in computationally simpler entropy estimators which are continuous

and differentiable in terms of the samples, since the main objective is not to estimate the entropy itself but to use this estimate in optimizing the parameters of an adaptive (learning) system. The consistency properties of an estimator are not questioned strictly in this field since for relatively small data sets it is not critical to have a consistent or an inconsistent estimate of the entropy as long as the global optimum lies at the desired solution. These methods work in general also in higher dimensional spaces and therefore can be applicable to mutual information. From the variety of learning theory applications, we mention here the nonparametric estimator of Rényi entropy from Erdogmus (27), based on Parzen (window) estimate (65) and some neural network-based approaches. The former estimator is consistent if the Parzen windowing and the sample mean are consistent for the actual pdf of the iid samples.

#### 8.3.5.2 Estimates by neural network approaches

In the probabilistic networks, the nodes and connections are interpreted as the defining parameters of a stochastic process. The net input to a node determines its probability of being active rather than its level of activation. The distribution of states in a stochastic network of these nodes can be calculated with models from statistical mechanics by treating the net inputs as energy levels. A well-known example of this type of network is Boltzmann Machine (i.e. (41)). The entropy of the stochastic process can then be calculated from its parameters, and hence optimized. The nonparametric technique by Parzen or kernel density estimation leads to an entropy optimization algorithm in which the network adapts in response to the distance between pairs of data samples. Such entropy estimate is differentiable and can therefore be optimized in a neural network, allowing to avoid the limitations encountered with parametric methods and probabilistic networks. The consistency of such method depends on the optimization algorithm.

### *8.3.6 Entropy estimates based on maximum likelihood*

When maximizing the likelihood, we may equivalently maximize the log of the likelihood and the number of calculations may be reduced. The log-likelihood is closely related to entropy and Fisher information. Popular methods for maximum likelihood are the Expectation-Maximization (EM) (i.e. Demster et al. (21)) and Improved Iterative Scaling (IIS) algorithms (Berger (12)). These methods are often used in classification tasks, especially in speech recognition. Paninski (64) used an exact local expansion of the entropy function and proved almost sure consistency (strong consistency) for three of the most commonly used discretized information estimators, namely the maximum likelihood (MLE) estimator the MLE with the so-called Miller-Madow bias correction (53), and for the jackknifed version of MLE from Efron and Stein (26).

### *8.3.7 Correction methods and bias analysis in undersampled regime*

These entropy estimates are mostly analytical and their bias can be computed. Most of them use Bayesian analysis and are asymptotically consistent (Miller (53), Paninski (64), Nemenman et al. (55)) but there is also another approach from Grassberger (36) applying Poisson distribution.

### 8.3.8 Kernel methods

*Kernel density estimation methods (KDE)*

Mutual information was first estimated by this approach by Moon et al. (54). The KDE methods have more advantages than the classical histogram methods: they have a better mean square error rate of convergence of the estimate to the underlying density, are insensitive to the choice of origin and the window shapes are not limited to the rectangular window. Kernel density estimator introduced by Silverman (83) in one dimensional space is defined

$$f(x) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h}), \tag{8.29}$$

where $h$ is the kernel width parameter and $K(x)$ the kernel function. It was shown by Kulkarni et al. in (49) that these estimators are consistent for any dimension.
Prichard and Theiler (71) introduced a method to compute information theoretic functionals based on mutual information using correlation integrals. Correlation integrals were introduced by Grassberger and Procaccia in ((35)). Consistency of this method was proven by Borovkova et al in (15). Schreiber (80) proposed to compute the transfer entropy also using the correlation integrals. (35).

## 8.4 Parametric estimators

Some assumption about either the functional form of the density or about its smoothness can be appropriate in some cases. The most common is to assume that the density has a parametric form. This approach is preferred when there is confidence that the pdf underlying the samples belongs to a known parametric family of pdf's. It is effective when the assumed parametric family is accurate but it is not appropriate in adaptation scenarios where the constantly changing pdf of the data under consideration may not lie in a simple parametric family. Parametric entropy estimation is a two step process. First, the most probable density function is selected from the space of possible density functions. This often requires a search through parameter space (for example maximum likelihood methods). Second, the entropy of the most likely density is evaluated.

Verdugo Lazo and Rathie (87) computed a table of explicit Shannon entropy expressions for many commonly used univariate continuous pdfs. Ahmed and Gokhale (3) extended this table and results to the entropy of several families of multivariate distributions, including multivariate normal, normal, log-normal, logistic and Pareto distributions. Consistent estimators for the parametric entropy of all the above listed multivariate distributions can be formed by replacing the parameters with their consistent estimators (computed by Arnold (7)).

### 8.4.1 Entropy estimators by higher-order asymptotic expansions

This class includes Fourier Expansion, Edgeworth Expansion and Gram-Charlier Expansion and other expansions (38). These methods are recommended especially for distributions which are "close to the Gaussian one" (42). The Edgeworth expansion, similarly as the Charlier-Gram expansion approximates a probability

distribution in terms of its cumulants. All the three expansion types are consistent, i.e. in infinity converge to the function which they expand, Cramer (19).

## 8.5 Generalized Granger causality

The classical approach of Granger causality as mentioned in Sec. 1.2 is intuitively based on the temporal properties, i.e. the past and present may cause the future but the future cannot cause the past (31). Accordingly, the causality is expressed in terms of predictability: if the time series $Y$ causally influences the time series $X$, then the knowledge of the past values of $X$ and $Y$ would improve a prediction of the present value of $X$ compared to the knowledge of the past values of $X$ alone. The causal influence in the opposite direction can likewise be checked by reversing the role of the two time series. Although this principle was originally formulated for wide classes of systems, both linear and nonlinear systems, the autoregressive modeling framework (Eq. (1)) proposed by Granger was basically a linear model, and such a choice was made primarily due to practical reasons (32). Therefore, its direct application to nonlinear systems may or may not be appropriate.

### *8.5.1 Nonlinear Granger causality*

Ancona et al. (4) extended Granger's causality definition to nonlinear bivariate time series. To define linear Granger causality (31), the vector autoregressive model was modeled by radial basis neural networks (16). A directionality index was introduced measuring the unidirectional, bidirectional influence or uncorrelation which was computed again by means of conditional mutual information applying generalized correlation integral (35).

### *8.5.2 Nonparametric Granger causality*

Despite the computational benefit of model-based (linear and/or nonlinear) Granger causality approaches, it should be noted that the selected model must be appropriately matched to the underlying dynamics, otherwise model mis-specification would arise, leading to spurious causality values. A suitable alternative would be to adopt nonparametric approaches which are free from model mismatch problems. We discuss here those nonparametric approaches which can be expressed in the information theoretic terms. Let us first reformulate the Granger causality in information theoretic terms (23; 22): For a pair of stationary, weakly dependent, bivariate time series $\{X_t, Y_t\}$, $Y$ is a Granger cause of $X$ if the distribution of $X_t$ given past observations of $X$ and $Y$ differs from the distribution of $X_t$ given past observations of $X$ only. Thus $\{Y_t\}$ is a Granger cause of $\{X_t\}$ if

$$F_{X_{t+1}}(x|F_X(t), F_Y(t)) \neq F_{X_{t+1}}(x|F_X(t)), \tag{8.30}$$

where $F_{X_{t+1}}$ represents the cumulative distribution function of $X_{t+1}$ given $F$, and $F_X(t)$ and $F_Y(t)$ represents the information contained in past observations of $X$ and $Y$ up to and including time $t$. The idea of the Granger causality is to quantify the additional amount of information on $X_{t+1}$ contained in $\mathbf{Y}_t$, given $\mathbf{X}_t$.

Now, the average amount of information which a random variable $X$ contains about another random variable $Y$ can be expressed in terms of generalized correlation integrals (see the equivalent Eq. (9)) as $I_q(X,Y) = \ln C_q(X,Y) - \ln C_q(X) - \ln C_q(Y)$ where the generalized correlation integral (35), $C_q$ can be estimated by

$$C_q(\mathbf{X},\varepsilon) = \frac{1}{N(N-1)^{q-1}} \sum_{j=1}^{N} \Big[ \sum_{i \neq j} \Theta(||X_j - X_i|| - \varepsilon) \Big]^{q-1}; \tag{8.31}$$

$\Theta$ is the Heaviside function, $\|.\|$ a norm and the last term is related to kernel density estimation. The extra amount of information that $\mathbf{Y}_t$ contains about $X_{t+1}$ in addition to the information already contained in $\mathbf{X}_t$ will be measured by the information theoretic measure of Granger causality: $I^{GC}_{Y\to X} = I(\mathbf{X}_t, \mathbf{Y}_t; X_{t+1}) - I(\mathbf{X}_t; X_{t+1}) = \ln C(\mathbf{X}_t, \mathbf{Y}_t, X_{t+1}) - \ln C(\mathbf{X}_t, X_{t+1}) - \ln C(\mathbf{X}_t, \mathbf{Y}_t) + \ln C(\mathbf{X}_t)$.

In order to obtain statistical significance, bootstrapping procedure is recommended to check if the statistic is significantly larger than zero (22).

Here the causality measure is based on conditional entropy, and unlike mutual or time-lagged information measures, can distinguish actually transported information from that produced as a response to a common driver or past history (80). Interestingly, these entropies can be expressed in terms of generalized correlation integrals whose nonparametric estimation is well known. Correlation integral based nonparametric Granger causality test was originally proposed by Baek and Brock (9) and then later modified by Hiemstra and Jones (39) in the field of econometrics. More details to this method can be found in (42).

## 8.6 Conclusion

The main objective of this paper was to show that information theory and information theoretical measures, in particular conditional mutual information, can detect and measure causal link and information flow between observed variables. However, it opens a more difficult question: How to reliably estimate these measures from a finite data set? Research literature abounds with various estimators with a diverse range of assumptions and statistical properties. Theoretically, for a good entropy estimator, the condition of consistency seems to be important. However, it should be noted that the conditions for desired consistency might be too restrictive for an experimental environment. Accordingly, we also critically reviewed those methods which have surprisingly good overall performance (i.e. small systematic and statistical error for a wide class of pdfs) though their consistency properties are not yet known. Last but not least, let us mention some informal comments on the detection of causality which are relevant to any causality measure applied. One needs to be extra careful before claiming a causal relationship between observed variables. From the viewpoint of establishing new models, inferences and control strategies, establishing a causal relationship is always tempting. However, one has to first carefully scrutinize the statistical properties of the observed data sequences and the completeness of the model or the assumptions necessary for the estimation of the information theoretic measures. Otherwise, spurious results could often be obtained (i.e. as discussed in Section 8.2.4). Despite these precautionary remarks, we would like to stress again that there are enough good reasons, contrary to B. Russel's arguments (75), to investigate causality, offering numerous applications in natural and physical sciences.

**Acknowledgements** The author thanks her colleagues M. Paluš, M. Vejmelka and J. Bhattacharya for their valuable discussions and for the support of the Czech Grant Agency by project MSMT CR 2C06001 (Bayes).

## References


[1] Abarbanel, H.D.I.: Introduction to Nonlinear Dynamics for Physicists. In: Lecture Notes in Physics. World Scientific, Singapore (1993)

[2] Ahmad, I.A., Lin, P.E.: A nonparametric estimation of the entropy for absolutely continuous distributions. IEEE Transaction on Information Theory **22**, 372-375 (1976)

[3] Ahmed, N.A., Gokhale, D.V.: Entropy expressions and their estimators for multivariate distributions. IEEE Transaction on Information Theory **35**, 688-692 (1989)

[4] Ancona, N., Marinazzo D., Stramaglia, S.: Radial basis function approach to nonlinear Granger causality of time series. Physical Review E **70**, 056221 (2004)

[5] Antos A., Kontoyiannis, I.: Convergence properties of functional estimates for discrete distributions. Random Structures and Algorithms, Special issue: Average-Case Analysis of Algorithms 19, 163-193 (2002)

[6] Aparicio, F.M, Escribano, A.:Information-theoretic analysis of serial dependence and cointegration. Studies in Nonlinear Dynamics and Econometrics 3, 119-140 (1998)

[7] Arnold, B.C.: Pareto Distributions. International Co-Operative Publishing House, Burtonsvile, MD, 1985

[8] Arnhold, J., Grassberger, P., Lehnertz, K., Elger, C.E.: A robust method for detecting interdependences: Application to intracranially recorded EEG. Physica D **134**, 419–430 (1999)

[9] Baek, E.G., Brock, W.A.: A general test for nonlinear Granger causality: Bivariate model, Working paper, Iowa State University and University of Wisconsin, Madison (1992)

[10] Baghli, M.: A model-free characterization of causality. Economics Letters **91**, 380–388 (2006)

[11] Beirlant, J., Dudewitz, E.J., Györfi, L., van der Meulen, E.C.: Nonparametric entropy estimation: An overview, Int. J. Math. And Statistical Sciences, **6**, 17-39 (1997)

[12] Berger, A.: The improved iterative scaling algorithm: A gentle introduction (http://www.cs.cmu.edu/afs/cs/user/aberger/www/ps/scaling.ps (1997)

[13] Blinowska, K.J., Kuś, R., Kamiński, M.: Granger causality and information flow in multivariate processes. Phys.Rev. E **70**, 050902(R) (2004)

[14] Boccaletti, S., Kurths, J., Osipov, G., Valladares, D.L., Zhou, C.S.: The synchronization of chaotic systems. Physics Reports **366**, 1-101 (2002)

[15] Borovkova, S., Burton, R., Dehling, H.: Consistency of the Takens estimator for the correlation Dimension. Annals of Applied Probability **9** 2, 376-390 (1999)

[16] Broomhead, D.S., Lowe, D.: Multivariate functional interpolation and adaptive networks. Complex Systems **2**, 321-355 (1988)

[17] Butte, A.J., Kohane, I.S.: Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements, Pac. Symp. Biocomput. 418-29 (2000)

[18] Chen, Y., Rangarajan, G., Feng, J., Ding, M: Analyzing mulitiple nonlinear time series with extended Granger causality. Phys. Lett. A **324**, 26-35 (2004)

[19] Cramer, H. On the composition of elementary errors. Skand. Aktuarietidskr. **11**, 13-4 and 141-80 (1928)

[20] Darbellay G., Vajda, I.: Estimation of the information by an adaptive partitioning of the observation space. IEEE Transaction on Information Theory **45**, 1315-1321 (1999)

[21] Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B **39**, 138 (1977)

[22] Diks, C., DeGoede, J.: A general nonparametric bootstrap test for Granger causality. In: Global Analysis of Dynamical Systems, Chapter 16, Broer, Krauskopf and Vegter (eds.), 391-403 (2001)

[23] Diks, C., Panchenko, V.: A note on the Hiemstra-Jones test for Granger non-causality. Studies in Nonlinear Dynamics and Econometrics **9** 4, 1-7 (2005)
[24] Dmitriev, Y.G., Tarasenko, F.P.: On the estimation functions of the probability density and its derivatives. Theory Probab. Appl. **18**, 628-633 (1973)
[25] Dobrushin, R.L.: A simplified method of experimentally evaluating the entropy of a stationary sequence. Teoriya Veroyatnostei i ee Primeneniya **3**, 462-464 (1958)
[26] Efron, B., Stein, C.: The jackknife estimate of variance, Annals of Statistics **9**,586-596, (1981)
[27] Erdogmus, D.: Information theoretic learning: Renyi's Entropy and its Application to Adaptive System Training, PhD thesis, University of Florida, 2002
[28] Fraser, A., Swinney, H.: Independent coordinates for strange attractors from mutual information. Phys. Rev. A **33** 11341140 (1986)
[29] German, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis. Chapman and Hall/A CRC Press Company, Texts in Statistical Science Series, 2004
[30] Geweke, J.: Inference and causality in economic time series models. In: Handbook of Econometrics, Griliches, Z., Intriligator, M.D. (eds.), North-Holland, vol. **2**, 1101–1144 (1984)
[31] Granger, C.W.J.: Investigating causal relations by econometric and cross-spectral methods, Econometrica **37** 424-438 (1969)
[32] Granger, C.W.J., Newbold, P.: Forecasting Economic Time Series. Academic Press, New York,(1977)
[33] Granger, C.W.J.: Testing for causality: A personal viewpoint, Journal of Economic Dynamics and Control **2**, 329-352 (1980)
[34] Granger, C.W.J.: Time series analysis, cointegration, and applications. Nobel Lecture, December 8, 2003. In: Les Prix Nobel. The Nobel Prizes 2003, Frängsmyr, T. (ed.), (Nobel Foundation, Stockholm, 2004) pp. 360–366. http://nobelprize.org/nobel_prizes/economics/laureates/2003/granger-lecture.pdf
[35] Grassberger, P., Procaccia, I.: Measuring of strangeness of strange attractors, Physica D **9**, 189-208 (1983)
[36] Grassberger, P.: Finite sample corrections to entropy and dimension estimates, Phys. Lett. A **128**, 369-373 (1988)
[37] Györfi, L., Van der Meulen, E.C.: On nonparametric estimation of entropy functionals. In: Nonparametric Functional Estimation and Related Topics, G. Roussas (ed.), pp. 81-95. Kluwer Academic Publisher, Amsterdam (1990)
[38] Haykin, S.:Neural Networks: A Comprehensive Foundation. Second Edition. Prentice Hall, Englewood Cliffs, NJ (1998)
[39] Hiemstra, C., Jones, J.D.: Testing for linear and nonlinear Granger causality in the stock pricevolume relation. Journal of Finance **49**, 1639-1664 (1994)
[40] Hinrichs, H., Heinze, H.J., Schoenfeld, M.A.: Causal visual interactions as revealed by an information theoretic measure and fMRI, NeuroImage **31**, 1051–1060 (2006)
[41] Hinton, G., Sejnowski, T.: Learning and relearning in Boltzmann machines In: Parallel Distributed procesing, Rumelhart, D., and J. McClelland J. (eds.) MIT Press, Cambridge, 1986, Vol **1**, Chapter 7, pp. 282-317.
[42] Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., Bhattacharya, J.: Causality detection based on information-theoretic approaches in time series analysis, Physics Reports **441** (1), 1-46 (2007) – doi:10.1016/j.physrep.2006.12.004
[43] Ivanov, A.V., Rozhkova, A.: Properties of the statistical estimate of the entropy of a random vector with a probability density. Problems of Information Transmission **10**, 171-178 (1981)
[44] Kagan, A.M., Linnik, Y.V., Rao, C.R.: Characterization Problems in Mathematical Statistics. Wiley, New York (1973)

[45] Katura, T., Tanaka, N., Obata, A., Sato, H., Maki, A.: Quantitative evaluation of interrelations between spontaneous low-frequency oscillations in cerebral hemodynamics and systemic cardiovascular dynamics. Neuroimage **31**, 1592-1600 (2006)
[46] Kolmogorov, A.N.: Entropy per unit time as a metric invariant of automorphism. Dokl. Akad. Nauk SSSR **124**, 754–755 (1959)
[47] Kozachenko, L.F., Leonenko, N.N.: Sample estimate of the entropy of a random vector. Problems of Information Transmission **23**, 95-100 (1987)
[48] Kraskov, A., Stögbauer H., Grassberger, P.: Estimation mutual information. Physical Review E **69**, 066138 (2004)
[49] Kulkarni, S.R., Posner, S.E. , Sandilya, S.: Data-dependent $k-NN$ and kernel estimators consistent for arbitrary processes. IEEE Transactions on Information Theory Vol. **48** , No. 10 (2002)
[50] Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22**, 79-86 (1951)
[51] Leonenko, N., Pronzato, L., Savani, V.: A class of Rényi information estimators for multidimensional densities. Laboratoire I3S, CNRS–Universit de Nice-Sophia Antipolis, Technical report I3S/RR-2005-14-FR (2005)
[52] Lugosi,G., Nobel, A.: Consistency of data-driven histogram methods for density estimation and classification. Ann. Statist. Vol. **24**, No. 2, 687-706 (1996)
[53] Miller, G.: Note on the bias of information estimates. In: Quastler, H. (ed.) Information theory in psychology II-B, (Glencoe, IL, pp.95-100, Free Press (1955)
[54] Moon, Y., Rajagopalan, B., Lall, U.: Estimation of mutual information using kernel density estimators. Physical Review E **52**, 2318-2321 (1995)
[55] Nemenman, I., Bialek, W., de Ruyter van Stevenick, R.: Entropy and information in neural spike trains: progress on sampling problem. Physical Review E **69**, 056111 (2004)
[56] Otsuka, K., Miyasaka, Y., Kubota, T.: Formation of an information network in a self-pulsating multimode laser. Phys. Rev. E **69**, 046201 (2004)
[57] Paluš, M., Komárek, V., Hrnčíř, Z., Štěrbová, K.: Synchronization as adjustment of information rates: Detection from bivariate time series. Phys. Rev. E **63**, 046211 (2001)
[58] Paluš, M.: Testing for nonlinearity using redudancies: Quantitative and qualitative aspects. Physica D **80**, 186-205 (1995)
[59] Paluš, M.: Coarse-grained entropy rates for characterization of complex time series. Physica D **93**, 64-77 (1996)
[60] Paluš, M.: Identifying and quantifying chaos by using information-theoretic functionals. In: Time series prediction: Forecasting the future and understanding the past, Weigend, A.S., Gershenfeld, N.A.(eds.) Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol XV., Addison-Wesley, Reading, MA, 1993 pp. 387-413
[61] Paluš, M.: Detecting nonlinearity in multivariate time series. Phys. Lett. A **213**, 138-147 (1996)
[62] Paluš, M., Hoyer, D.: Detecting nonlinearity and phase synchronization with surrogate data. IEEE Engineering in Medicine and Biology **17**, 40-45 (1998)
[63] Paluš, M., Vejmelka, M.: Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. Phys. Rev. E **75** (2007) 056211 – doi:10.1103/PhysRevE.75.056211
[64] Paninski, L.: Estimation of entropy and mutual information. Neural Computation **15**, 1191-1253 (2003)
[65] Parzen, E.: On estimation of a probability density function and mode. In: Time Series Analysis Papers (Holden-Day, Inc., San Diego, California (1967)
[66] Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, New York (2000)

[67] Peters M.A., Iglesias, P.A.: Minimum entropy control for discrete-time varying systems. Automatica **33**, 591-605 (1997)
[68] Pikovsky, A., Rosenblum, M., Kurths, J.: Synchronization. A Universal Concept in Nonlinear Sciences. Cambridge University Press, Cambridge (2001)
[69] Pompe, B.: Measuring statistical dependencies in a time series. J. Stat. Phys. **73**, 587-610 (1993)
[70] Prasaka Rao, B.L.S.: Nonparametric Functional Estimation. Academic Press, New York (1983)
[71] Prichard, D., Theiler, J.: Generalized redundancies for time series analysis. Physica D **84**, 476–493 (1995)
[72] Le Van Quyen, M., Martinerie, J., Adam, C., Varela, F.J.: Nonlinear analyses of interictal EEG map the brain interdependences in human focal epilepsy. Physica D **127**, 250-266 (1999)
[73] Quian Quiroga, R., J. Arnhold, J., Grassberger,P.: Learning driver-response relationships from synchronization patterns, Phys. Rev. E **61**(5), 5142-5148 (2000)
[74] Rényi, A.: On measures of entropy and information, In: Proc. Fourth Berkeley Symp. Math. Stat. and Probability, Vol. 1 Berkeley, CA, University of California Press, (1961) pp. 547-561
[75] Russel,B.: On the notion of cause. In: Proceedings of the Aristotelian Society, New Series **13**, 1-26 (1913)
[76] Schäfer, C., Rosenblum, M.G., Kurths, J., Abel, H.H.: Heartbeat synchronized with ventilation. Nature **392**, 239-240 (1998)
[77] Schiff, S.J., So, P., Chang, T., Burke, R.E., Sauer, T.: Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. Phys. Rev. E **54**, 6708-6724 (1996)
[78] Schmitz, A.: Measuring statistical dependence and coupling of subsystems. Phys. Rev. E **62**, 7508-7511 (2000)
[79] Schraudolph, N.: Gradient-based manipulation of non-parametric entropy estimates. IEEE Trans. On Neural Networks **14**, 828-837 (2004)
[80] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2000) 461-464.
[81] Selltiz, C., Wrightsman, L.S., Cook, S.W.: Research Methods in Social Relations. Holt, Rinehart and Winston, New York (1959)
[82] Shannon, C.E.: A mathematical theory of communication, Bell System Tech. J. **27**, 379-423 (1948)
[83] Silverman, B.W.: Density Estimation. Chapman and Hall, London (1986)
[84] Sinai, Y.G.: On the concept of entropy for a dynamic system. Dokl. Akad. Nauk SSSR **124**, 768-771 (1959)
[85] Takens, F: In: Rand, D.A., Young, D.S. (eds.) Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics **898** (Springer, Berlin, 1981), p.365
[86] Tsybakov A.B., van Meulen,E.C.: Root-n consistent estimators of entropy for densities with unbounded support. Scand. J. Statist. **23**, 75-83 (1994)
[87] Verdugo Lazo, A.C.G., Rathie, P.N.: On the entropy of continuous probability distributions. IEEE Transactions on Information Theory **24**, 120-122 (1978)
[88] Wang, Q., Kulkarni, S.R., Verdú, S.: A nearest-neighbor approach to estimating divergence between continuous random vectors. ISIT 2006, Seattle, USA, July 9-14 (2006)
[89] Wiener, N.: The theory of prediction, in: Modern Mathematics for Engineers, Beckenbach, E.F. (ed.), McGraw-Hill, New York (1956)
[90] http://en.wikipedia.org/wiki/Causality