Causality detection based on information-theoretic approaches in time series analysis

Kateřina Hlaváčková-Schindler[1*], Milan Paluš[2], Martin Vejmelka[2], Joydeep Bhattacharya[1,3]

[1] *Commission for Scientific Visualization, Austrian Academy of Sciences*
Donau-City Str. 1, A-1220 Vienna, Austria
Phone: +43 (1) 515 81 6708, Fax: +43(1)20501 18900
katerina.schindler@assoc.oeaw.ac.at

[2] *Institute of Computer Science, Academy of Sciences of the Czech Republic*
Pod Vodárenskou Věží 2, 18207 Praha 8, Czech Republic

[3] *Department of Psychology, Goldsmiths College,*
University of London, New Cross SE14 6NW, London, UK

Abstract

Synchronization, a basic nonlinear phenomenon, is widely observed in diverse complex systems studied in physical, biological and other natural sciences, as well as in social sciences, economy and finance. While studying such complex systems, it is important not only to detect synchronized states, but also to identify causal relationships (i.e. who drives whom) between concerned (sub) systems. The knowledge of information-theoretic measures (i.e. mutual information, conditional entropy) is essential for the analysis of information flow between two systems or between constituent subsystems of a complex system. However, the estimation of these measures from a set of finite samples is not trivial. The current extensive literatures on entropy and mutual information estimation provides a wide variety of approaches, from approximation-statistical, studying rate of convergence or consistency of an estimator for a general distribution, over learning algorithms operating on partitioned data space to heuristical approaches. The aim of this paper is to provide a detailed overview of information theoretic approaches for measuring causal influence in multivariate time series and to focus on diverse approaches to the entropy and mutual information estimation.

PACS-1999: 05.10-a; 0.45-a; 07.05-t

*Corresponding author.

Keywords: causality, entropy, mutual information, estimation

Contents

1 Introduction

1.1 Causality

Detection and clarification of cause-effect relationships among variables, events or objects have been the fundamental questions of most natural and social sciences over the history of human knowledge. Despite some philosophers of mathematics like B. Russel [198](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. [171]), the language of all sciences, including mathematics and physics, has been using this term actively until now. Mathematical and physical relations are not limited only to equations. To advocate the Russell's view, any exact and sufficiently comprehensive formulation of what is causality is problematic. Causality can be understood in terms of a "flow" among processes and expressed in mathematical language and mathematically analysed. Current statistics understands causal inference as one of its most important problems.

The general philosophical definition of causality from the Wikipedia Encyclopedia [253] states: "The philosophical concept of causality or causation refers to the set of all particular "causal" or 'cause-and-effect' relations. A neutral definition is notoriously hard to provide, since every aspect of causation has received substantial debate. 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. Despite of these definitions, causal relationships are often investigated in situations which are influenced by uncertainty. 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 strengths of these relationships. Although there is an extensive literature on causality modeling, applying and combining mathematical logic, graph theory, Markov models, Bayesian probability, etc. (i.e. Pearl, [171]), the aim of our review is to focus only on the information-theoretic approaches which understand causality as a phenomenon which can be "measured" and quantified.

We want to provide a detailed overview of the information-theoretic approaches for measuring of a causal influence in multi-variate time series. Based on the definition of causality in the information-theoretic framework, we focus on approaches to the estimation of entropy and mutual information. The previous review papers on entropy estimation (Beirlant et al. [22] and Erdogmus [70]) focused on non-parametric entropy estimation. In this paper we not only update the state of art on non-parametric entropy estimation, but discuss also parametrical estimation.

1.2 Causal measures

As mentioned earlier, there has been no universally accepted definition of causality (see Granger, 1980 [91] for a lively discussion on this issue), so it would be futile to search for a unique causality measure. However, we mention here briefly the salient features of this debate for convenience. Most of the earlier research literature attempts to discuss unique causes in deterministic situations, and two conditions are important for deterministic causation: (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 (however, see [157] for a critique of Suppe's causality), there are some arbitrariness in practice in categorizing an event [91].

Till 1970, the causal modeling was mostly used in social sciences. This was primarily due to a pioneering work by Selltiz et al (1959) [210] 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 [252]: "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 [92] 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; and
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 [92]. 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 [88] is based on a linear regression model

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

where ξ_t are uncorrelated random variables with zero mean and variance σ^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. (1) to

$$Y_t = a_o + \sum_{k=1}^{L} b_{1k} Y_{t-k} + \tilde{\xi}_t. \tag{2}$$

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

$$GS = \frac{(R_2 - R_1)/L}{R_1/(N-2L)} \tag{3}$$

where R_1 is the residual sum of squares in (1) and R_2 is the residual sum of squares in (2). The GS test statistic has an F-distribution with L and $N - 2L$ degrees of freedom [2]. On the other hand, the Granger-Wald test is defined as

$$GW = N \frac{\hat{\sigma}^2_{\tilde{\xi}_t} - \hat{\sigma}^2_{\xi_t}}{\hat{\sigma}^2_{\xi_t}} \tag{4}$$

where $\hat{\sigma}^2_{\tilde{\xi}_t}$ is the estimate of the variance of $\tilde{\xi}_t$ from model (2) and $\hat{\sigma}^2_{\xi_t}$ is the estimate of the variance of ξ_t from model (1). The GW statistic follows the χ^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 [85] for a comprehensive survey of the literature), but also in diverse fields of natural sciences such as climatology (see [236] and references therein) or neurophysiology, where specific problems of multichannel electroencephalogram recordings were solved by generalizing the Granger causality concept to multivariate case [126, 36]. 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 economy (econometrics) on one side, and physics and other natural sciences on the other side. In the field of economy, Baek and Brock [14] and Hiemstra and Jones [110] 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 [94]. A non-parametric approach to non-linear causality testing, based on non-parametric regression, was proposed by Bell et al. [23]. Following Hiemstra and Jones [110], Aparicio and Escribano [10] 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 [177, 37]. Synchronization and related phenomena were observed not only in physical, but also in many biological systems. Examples include the cardio-respiratory interaction [199, 200, 166, 38, 220, 120] and the synchronization of neural signals [201, 187, 228, 160, 159]. 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. [201] and Quyen et al. [187] 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 [201, 187] 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 the 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. [12] 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 [188, 203], 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. [46] using local linear predictors. An important class of nonlinear predictors are based on so-called radial basis functions [42] which were used for nonlinear parametric extension of the Granger causality concept [7, 143]. Although they are not exactly based on information theory, they are connected to methods reviewed here and will be discussed more in detail in Section 6.

A non-parametric method for measuring causal information transfer between systems was proposed by Schreiber [206]. His *transfer entropy* is designed as a Kullback-Leibler distance (Eq. (15) in Sec. 2.1) of transition probabilities. This measure is in fact an information-theoretic functional of probability distribution functions.

Paluš et al. [160] 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 will be shown that, with proper conditioning, the Schreiber's transfer entropy [206] is equivalent to the conditional mutual information [160]. The latter, however, is a standard measure of information theory [50].

Turning our attention back to econometrics, we can follow further development due to Diks and DeGoede [62]. They again applied a nonparametric approach to nonlinear Granger causality using the concept of correlation integrals [94] and pointed out the connection between the correlation integrals and information theory. Diks and Panchenko [64] critically discussed the previous tests of Hiemstra and Jones [110]. As the most recent development in economics, Baghli [15] 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. The Schreiber's transfer entropy was used in climatology [149, 245], in physiology [245, 130], in neurophysiology [45] and also in analysis of financial data [144]. Paluš et al. [159, 160] 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. [111] and Pflieger and Greenblatt [176]. Causality or coupling directions in multimode laser dynamics is another diverse field where the

conditional mutual information was applied [156]. Paluš and Stefanovska [158] adapted the conditional mutual information approach [160] to analysis of instantaneous phases of interacting oscillators and demonstrated suitability of this approach for analyzing causality in cardio-respiratory interaction [160]. The later approach has also been applied in neurophysiology [39].

Having reviewed the relevant literature and also after extensive practical experience, 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. Therefore we focus in this review to the information theory and its applications in inference of causality from experimental time series, although we do not refrain mentioning other approaches.

The outline of the paper is the following. In Section 1 we explain the terms causality and its measures. Basic notions of information theory and their approaches to causality detection are discussed in Section 2. Section 3 classifies the methods which will be reviewed. The rest of the sections deals with the concrete methods: the nonparametric methods are treated in Section 4, the parametric methods in 5. Generalized Granger causality, although not being explicitly an information-theoretic approach, deserves our attention, too. It is discussed in Section 6. Section 7 is devoted to our conclusion.

2 Information theory as a tool for causality detection

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 [211], the founder of the information theory. 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

$$H(x) = -\int p(x) \log p(x) dx, \tag{5}$$

where log is natural logarithm. We assume that $H(x)$ is well-defined and finite. One can define the discrete version of differential entropy as follows. 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 the **entropy** $H(S)$:

$$H(S) = -\sum_{i=1}^{m} p_i \log p_i. \tag{6}$$

Entropy H(S) can be understood as the "quantity of surprise one should feel upon reading the result of a measurement" [78]. So entropy of S can be seen as the uncertainty of S.

More general term of entropy for which is Shannon differential entropy a special case, is Rényi entropy. **Rényi entropy** is for a continuous case defined as [191]

$$H_\alpha(x) = \frac{1}{1-\alpha} \int \log^\alpha p(x) dx, \tag{7}$$

and for the discrete case

$$H_\alpha(S) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_i^\alpha, \tag{8}$$

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's measure of entropy. Rényi's measure satisfies $H_\alpha(x) \leq H_{\alpha'}(x)$ for $\alpha > \alpha'$.

Besides Shannon and Rényi entropy, other entropy definitions (i.e. Tsallis, Havrda -Charvát, etc.) are studied in the mathematical literature, but Shannon entropy is the only one possessing all the desired properties of an information measure. Therefore its efficient and accurate estimate is of prime importance. Based on this, although Rényi entropy will be also discussed, we focus in our review mainly on Shannon entropy estimators and their application to mutual information. The definition of the latter follows, without lack of generality, only for discrete distributions.

The **joint entropy** $H(X, Y)$ of two discrete random variables X and Y is defined analogously

$$H(X, Y) = -\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} p(x_i, y_j) \log p(x_i, y_j) \tag{9}$$

where $p(x_i, y_j)$ denotes the joint probability that X is in state x_i and Y in state y_j (the number of possible states m_X and m_Y may differ). If the random variables X and Y are statistically independent, the joint entropy $H(X, Y)$ becomes $H(X, Y) = H(X) + H(Y)$. In general, the joint entropy may be expressed in terms of **conditional entropy** $H(X|Y)$ as follows

$$H(X, Y) = H(X|Y) + H(Y), \tag{10}$$

where

$$H(X|Y) = -\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} p(x_i, y_j) \log p(x_i|y_j) \tag{11}$$

and $p(x_i|y_j)$ denotes the conditional probability. The joint entropy expresses how much uncertainty remains in X when Y is known.

The **mutual information** $I(X, Y)$ between two random variables X and Y is then defined as [211]

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \tag{12}$$

Mutual information of two variables reflects the mutual reduction in uncertainty of one by knowing the other one. 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. It is invariant under one-to-one measurable transformations. Mutual information (MI) measures the strength of dependence in the sense that: 1) $I(X,Y) = 0$ iff X is independent of Y; 2) For the bivariate normal distributions, $I(X,Y) = \frac{1}{2}\log(1-\rho^2(X,Y))$, where ρ is the coefficient of correlation between X and Y.

The **conditional mutual information** [211] between random variables X and Y given Z is defined as

$$I(X,Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z). \tag{13}$$

For Z independent of X and Y we have

$$I(X,Y|Z) = I(X,Y). \tag{14}$$

Beside mutual information, there are other measures of relationships among variables. The most used measures like Pearson's correlation or Euclidean distance can reflect the degree of linear relationship between two variables. Mutual information is sensitive to other than linear functional relationships (i.e. non-linear) and therefore provides a more general criterion to investigate relationships between variables.

In the following we mention some other useful entropies and their relationship to mutual information.

The **Kullback-Leibler divergence** (KLD, also called relative entropy or cross-entropy), introduced by Kullback and Leibler [137], is an alternative approach to mutual information. The Kullback entropy $K(p,p^0)$ between two probability distributions $\{p\}$ and p^0 is

$$K(p,p^0) = \sum_{i=1}^{m} p_i \log(\frac{p_i}{p_i^0}). \tag{15}$$

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 (Cover and Thomas, [50]).

The **neg-entropy** is defined as

$$K(p,\phi_p) = \sum_{i} p_i \log(\frac{p_i}{\phi_p}), \tag{16}$$

(i.e. $p^0 = \phi_p$), where ϕ_p is multivariate Gaussian distribution having the same mean vector and covariance matrix as p. 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. [84]. So we can look at the results about Kullback-Leibler entropy as if they were applied to mutual information (relationship of KLD to entropy and conditional entropy can be also found in [84]).

2.2 Information, entropy and dynamical systems

A considerable amount of approaches to inferring causality from experimental time series have their roots in studies of synchronization of chaotic systems. It is therefore useful to make a few remarks about the connection between the theory of dynamical systems and information theory.

A. N. Kolmogorov, who introduced the theoretical concept of classification of dynamical system by information rates [134], was inspired by information theory and together with Y.G. Sinai generalized the notion of entropy of an information source [134, 216]. A possibility to use ideas and methods from the information theory in the field of nonlinear dynamics and related analyses of experimental data was demonstrated by Shaw [212, 213]. Fraser [79] analyzed information aspects of chaotic dynamics on strange attractors. Paluš [163] 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. Let us review several important details.

Consider n discrete random variables $X_1, \ldots, X_n$ with sets of values $\Xi_1, \ldots, \Xi_n$, respectively. The probability distribution for an individual X_i is $p(x_i) = \Pr\{X_i = x_i\}$, $x_i \in \Xi_i$. We denote the probability distribution function by $p(x_i)$, rather than $p_{X_i}(x_i)$, for convenience. Analogously, the joint distribution for the n variables $X_1, \ldots, X_n$ is

$p(x_1, \ldots, x_n) = \Pr\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\}$, $(x_1, \ldots, x_n) \in \Xi_1 \times \cdots \times \Xi_n$.

The **marginal redundancy** $\varrho(X_1, \ldots, X_{n-1}; X_n)$, which in the case of two variables reduces to the above defined mutual information $I(X_1; X_2)$, quantifies the average amount of information about the variable X_n contained in the $n-1$ variables $X_1, \ldots, X_{n-1}$, and is defined as:

$$\varrho(X_1, \ldots, X_{n-1}; X_n) = \sum_{x_1 \in \Xi_1} \cdots \sum_{x_n \in \Xi_n} p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{p(x_1, \ldots, x_{n-1}) p(x_n)}. \tag{17}$$

Now, let $\{X_i\}$ be a stochastic process, i.e. an indexed sequence of random variables, characterized by the joint probability distribution function $p(x_1, \ldots, x_n)$. The **entropy rate** of $\{X_i\}$ is defined as

$$h = \lim_{n \to \infty} \frac{1}{n} H(X_1, \ldots, X_n) \tag{18}$$

where $H(X_1, \ldots, X_n)$ is the joint entropy of the n variables X_1,..., X_n with the joint distribution $p(x_1, \ldots, x_n)$:

$$H(X_1, \ldots, X_n) = -\sum_{x_1 \in \Xi_1} \cdots \sum_{x_n \in \Xi_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n). \tag{19}$$

Consider further two processes $\{X_i\}$ and $\{Y_i\}$, their **mutual information rate** [118, 178] is

$$\imath(X; Y) = \lim_{n \to \infty} \frac{1}{n} I\big((X_1, \ldots, X_n); (Y_1, \ldots, Y_n)\big). \tag{20}$$

A way from the entropy rate of a stochastic process to the **Kolmogorov-Sinai entropy** (KSE) of a dynamical system can be straightforward due to the fact that any stationary stochastic process corresponds to a measure-preserving dynamical system, and vice versa [173]. Then for the definition of the KSE we can consider the equation (18), however, the variables X_i should be understood as m-dimensional variables, according to the dimensionality of a dynamical system. If the dynamical system is evolving in continuous (probability) measure space, then any entropy depends on a partition chosen to discretize the space and the KSE is defined as supremum over all finite partitions [48, 173, 217].

Possibilities to compute the entropy rates from data are limited to a few exceptional cases: for stochastic processes it is possible, e.g., for the finite-state Markov chains [50]. In the case of a dynamical system in a continuous measure space, the KSE can be in principle reliably estimated, if the system is low-dimensional and a large amount of (practically noise-free) data is available. In such a case, Fraser [79] proposed to estimate the KSE of a dynamical system from the asymptotic behavior of the marginal redundancy, computed from a time series generated by the dynamical system. In such an application one deals with a time series $\{y(t)\}$, considered as a realization of a stationary and ergodic stochastic process $\{Y(t)\}$. Then, due to ergodicity, the marginal redundancy (17) can be estimated using time averages instead of ensemble averages, and, the variables X_i are substituted as

$$X_i = y(t + (i-1)\tau). \tag{21}$$

Due to stationarity, the marginal redundancy

$$\varrho^n(\tau) \equiv \varrho(y(t), y(t+\tau), \ldots, y(t+(n-2)\tau); y(t+(n-1)\tau)) \tag{22}$$

is a function of n and τ, independent of t.

If the underlying dynamical system is m-dimensional and the marginal redundancy $\varrho^n(\tau)$ is estimated using a partition fine enough (to attain the so-called generating partition [79, 48, 217]) then the asymptotic behavior

$$\varrho^n(\tau) \approx H_1 - |\tau| h \tag{23}$$

is attained for $n = m+1, m+2, \ldots$, for some range of τ [79, 164, 163]. The constant H_1 is related to $\varrho^n(0)$ and h is the estimate of the Kolmogorov-Sinai entropy of the dynamical system underlying the analyzed time series $\{y(t)\}$.

2.3 Coarse-grained entropy and information rates

In order to obtain such estimates as those in equation (23), large amounts of data are necessary [163]. Unfortunately, such data requirement are not realistic in usual experiments. To avoid this, Paluš [162] 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 $\{X(t)\}$, $t = 1, 2, 3, \ldots$. In the following we denote

$x(t)$ as x and $x(t+\tau)$ as x_τ 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 τ_{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**

$$||I(x;x_\tau)|| = \frac{\Delta\tau}{\tau_{max} - \tau_{min} + \Delta\tau} \sum_{\tau=\tau_{min}}^{\tau_{max}} I(x;x_\tau) \tag{24}$$

with $\tau_{min} = \Delta\tau = 1$ sample as a usual choice. The CER h^1 is then defined as

$$h^1 = I(x, x_{\tau_0}) - ||I(x;x_\tau)||. \tag{25}$$

It was shown that the CER h^1 provides the same classification of states of chaotic systems as the exact KSE [162]. 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 (24). Paluš et al. [160] 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. [160] defined their **mutual coarse-grained information rate** (MCIR) by

$$i(X,Y) = \frac{1}{2\tau_{max}} \sum_{\tau=-\tau_{max}}^{\tau_{max};\tau\neq 0} I(x;y_\tau). \tag{26}$$

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. [160] proposed to evaluate the **conditional coarse-grained information rate** CCIR $i_0(X|Y)$ of $\{X\}$ given $\{Y\}$:

$$i_0(X|Y) = \frac{1}{\tau_{max}} \sum_{\tau=1}^{\tau_{max}} I(x;x_\tau|y), \tag{27}$$

considering the usual choice $\tau_{min} = \Delta\tau = 1$ sample. Recalling (14), 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. [160] define

$$i(X|Y) = i_0(X|Y) - i(X). \tag{28}$$

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 τ time units ahead (τ-future thereafter). This measure, however, could also contain an information about the τ-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 τ-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 τ as above

$$i_1(X, Y|X) = \frac{1}{\tau_{max}} \sum_{\tau=1}^{\tau_{max}} I(y; x_\tau|x); \tag{29}$$

In order to obtain the "net asymmetric" information measure, we subtract the symmetric MCIR (26):

$$i_2(X, Y|X) = i_1(X, Y|X) - i(X, Y). \tag{30}$$

Using a simple manipulation, we find that $i_2(X, Y|X)$ is equal to $i(X|Y)$ defined in Eq. (28). By using two different ways for definition of a directional information rate, Paluš et al. [160] 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. [160] 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. In transients from normal brain activity to epileptic seizures, asymmetries in information flow emerge or are amplified. The potential of the CTIR method for anticipating seizure onsets and for localization of epileptogenic foci was discussed in [159]. Paluš and Stefanovska [158] adapted the conditional mutual information approach [160] to the analysis of instantaneous phases of interacting oscillators and demonstrated suitability of this approach for analyzing causality in cardio-respiratory interaction [160].

2.4 Conditional mutual information and transfer entropy

The principal measure, used by Paluš et al. [160] 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 $\vec{X}(t)$ and $\vec{Y}(t)$, the time delay embedding vectors according to Takens [224] are used. Then, back in time-series representation, we have

$$I\big(\vec{Y}(t); \vec{X}(t+\tau)|\vec{X}(t)\big) = \tag{31}$$

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

where η and ρ are time lags used for the embedding of systems $\vec{X}(t)$ and $\vec{Y}(t)$, respectively. For simplicity, only information about one component $x(t+\tau)$ in the $\tau-$future of the system $\vec{X}(t)$ is used. The opposite CMI $I\big(\vec{X}(t); \vec{Y}(t+\tau)|\vec{Y}(t)\big)$ 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 [206] 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

$$p(i_{n+1}|i_n, ..., i_{n-k+1}) = p(i_{n+1} \,|\, i_n^{(k)}, j_n^{(l)}), \tag{32}$$

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 Kullback-Leibler divergence (15) to measure the deviation of the transition probabilities from the generalized Markov property (32). This results into the definition

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{p(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})}, \tag{33}$$

denoted as *transfer entropy*. The transfer entropy 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 $p(i_{n+1}|i^{(k)}, j_n^{(l)})$.

Considering the relation between the joint and conditional probabilities, from Eq. (33) we can obtain

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

and, after a few simple manipulations we have

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log p(i_{n+1}, j_n^{(l)}|i_n^{(k)}) \tag{34}$$

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

Now, considering Eq. (13), let us go back to the expression for conditional mutual information (31) and express it using conditional entropies as

$$I\big(\vec{Y}(t); \vec{X}(t+\tau)|\vec{X}(t)\big) = \tag{35}$$

$$H\Big(\big(y(t), y(t-\rho), \ldots, y(t-(m-1)\rho)\big)|\big(x(t), x(t-\eta), \ldots, x(t-(n-1)\eta)\big)\Big)$$

$$+H\Big(x(t+\tau)|\big(x(t), x(t-\eta), \ldots, x(t-(n-1)\eta)\big)\Big)$$

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

Now, we express the conditional entropies using the probability distributions. However, let us change our notations according to Schreiber by equating $I \equiv \{X(t)\}$, $m = k$, and $J \equiv \{Y(t)\}$, $n = l$, substitute t for n and set $\eta = \rho = \tau = 1$. We can see that we obtain the same expression as Eq. (34) for the transfer entropy. Thus the transfer entropy is in fact an equivalent expression for the conditional mutual information.

3 Basic classification of current methods for entropy and mutual information estimation

Calculations of mutual information occur mainly in the literature in four contexts in the analysis of observational data: learning theory questions, identification of nonlinear correlation (and consequently causality detection), determination of an optimal sampling interval and in the investigation of causal relationships concretely with directed mutual information.

The key problem for causality detection by means of conditional mutual information is to have an estimator of mutual information. Most entropy estimators in the literature, which are designed for multi-dimensional spaces, can be applied to mutual information estimation. Therefore this paper focuses mainly to entropy estimation in multidimensional spaces. In the following, we adopt the classification and mathematical criteria for evaluation of the differential entropy estimators from the overview of nonparametric methods from Beirlant et al. [22].

The basic properties of differential entropy are summarized e.g. in [50]. The differential entropy has some important extremal properties:

(i) If the density f is concentrated on the unit interval $[0, 1]$ then the differential entropy is maximal iff f is uniform on $[0, 1]$.

(ii) If the density is concentrated on the positive half line and has a fixed expectation then the differential entropy takes its maximum for the exponential distribution.

(iii) If the density has fixed variance then the differential entropy is maximized by the Gaussian density.

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

The following usual 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:
$(P1)$ $\int f(\log f)^2 < \infty$. (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 ([123]). When dealing with the convergence properties of the presented estimators, one needs the following definitions. By means of **Asymptotically consistent** estimator one understand that the series of the approximants converge in infinity to the function to be approximated (see i.e. [22]). **Asymptotically unbiased** estimator is that one which is unbiased in the limit.

3.2 Classification of methods for entropy estimation

There is an extensive literature dealing with entropy estimates, and in their classification we will roughly keep the schema given by Beirlant et al. in [22] and by Erdogmus in [70]. We extend them to newer methods and approaches and also to the non-parametric ones. At this point it is necessary to note that a great proportion of the literature dealing with entropy and MI estimation was originally motivated by other questions than detection of causality: by learning theory questions, i.e. blind separation, necessary for application of principal or independent component analysis (PCA and ICA) or by nonlinear dynamics applications.

Many of these methods, although accurate in one or two dimension, become inapplicable in higher dimensional spaces (because of their computational complexity). In this review paper we focus mainly on entropy estimation methods which are applicable in higher-dimensional spaces. The older methods (mostly adopted from [22]) will be presented briefly and the newer methods will be discussed more in detail.

4 Nonparametric estimators

4.1 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 [215]. In these estimates, a consistent probability density function estimator is substituted into the place of the pdf of a functional.

4.1.1 Integral estimates of entropy

These estimates have form given by

$$H_n = -\int_{A_n} f_n(x) \log f_n(x) dx \tag{36}$$

where, with the set A_n one typically excludes the small or tail values of f_n. These estimates evaluate approximately (or exactly) the integral. The first such estimator was introduced by Dmitriev and Tarasenko [66], who proposed to estimate H_n by (36) for $d = 1$, where $A_n = [-b_n, b_n]$ and f_n is the kernel density estimator (see section 4.8). The strong consistency of H_n defined by formula (36) was shown in Ref. [66] and in Ref. [181]. Mokkadem [150] calculated the expected L_r error of this estimate, also for the estimation of mutual information.

To evaluate the (infinite) integral form of entropy (the exact or approximate one), numerical integration must be performed, which is not easy to be computed if f_n is a kernel density estimator.

Joe [121] estimated entropy $H(f)$ by the sequence of integral estimators H_n given by formula (36) when f is a multivariate pdf, but he pointed out that the calculation of H_n, when f_n is a kernel estimator, gets more difficult for

$d \geq 2$. He therefore excluded the integral estimate from further study (curse of dimensionality). The integral estimator can however be easily calculated if, for example, f_n is a histogram [99].

4.1.2 Resubstitution estimates

The resubstitution estimation is of the form

$$H_n = -\frac{1}{n}\sum_{i=1}^{n} \log f_n(X_i). \tag{37}$$

This approach includes the approximation of the expectation operator (i.e. the expected value of an argument) in the entropy definition with the sample mean or by polynomial expansions. Polynomial expansions of pdfs in order to estimate entropy were lately applied by Van Hulle [241] who used Edgeworth expansion (see Section 5 Parametric methods) and by Viola [251]. Ahmad and Lin in [3] proposed estimating $H(f)$ by (37), where f_n is a kernel density estimate. They showed the mean square consistency of (37) under some mild conditions.

Joe [121] considered the estimation of $H(f)$ for multivariate pdf's by (37), also based on a kernel-based estimate. Joe obtained asymptotic bias and variance terms, and showed that non-unimodal kernels satisfying certain conditions can reduce the mean square error. He concluded that in order to obtain accurate estimates especially in multivariate situations, the number of samples required increased rapidly with the dimensionality d of the multivariate density (curse of dimensionality). These results strongly rely on conditions T2 and P2.
Hall and Morton [105] investigated both the case when f_n is a histogram density estimator and when it is a kernel estimator in (37). Root-n consistency of form of asymptotic normality was proven for histogram under certain tail and smoothness conditions with $\sigma^2 = Var(\log f(X))$. The histogram-based estimator can only be root-n consistent when $d = 1$ or 2. However, the estimator has in case of $d = 2$ significant bias. They suggest an empirical rule for the bandwidth, using a penalty term. The effects of tail behavior, distribution smoothness and dimensionality on convergence properties were studied with the conclusion that root-n consistency of entropy estimation requires appropriate assumptions about each of these three features. These results are valid for a wide class of densities f having unbounded support.

4.1.3 Splitting data estimate

The approach of these methods is similar to the approach of 4.1.2 except that the sample set is divided into two subsets, $X_1, \ldots, X_l$ and $X_1^*, \ldots, X_m^*$, $n = l + m$; One is used for density estimation, while the other one for sample mean. Based on $X_1, \ldots, X_l$, one constructs a density estimate f_l and then, using this density estimate and then the second sample, estimates $H(f)$ by

$$H_n = -\frac{1}{m}\sum_{i=1}^{m} I_{[X_i^* \in A_l]} \log f_l(X_i^*). \tag{38}$$

This approach was used for f_l being the histogram density estimate in Ref. [98], for f_l being the kernel density estimate in Ref [99] and for f_l being any L_1-consistent density estimate such that $[X_i^* \in A_l] = [f_l(X_i^* \geq a_l], 0 < a_l \to 0$ in Ref [100]. Under some mild tail and smoothness conditions of f, the strong consistency was shown for general dimension d.

4.1.4 Cross-validation estimate

This class of estimators uses a leave-one-out principle in the resubstitution estimate. The entropy estimate is obtained by averaging the leave-one-out resubstitution estimates of the data set. If $f_{n,i}$ denotes a density estimate based on $X_1, \ldots, X_n$ leaving X_i out, then the corresponding density estimate is of the form

$$H_n = -\frac{1}{m} \sum_{i=1}^{n} I_{[X_i \in A_n]} \log f_{n,i}(X_i). \tag{39}$$

Ivanov and Rozhkova [119] proposed such an estimator for Shannon entropy when $f_{n,i}$ is a kernel-based pdf estimator. They showed strong consistency, and also made a statement regarding the rate of convergence of the moments $E|H_n - H(f)|^r$, $r \geq 1$.

Hall and Morton [105] also studied entropy estimates of the type (39) based on kernel estimator. For $d = 1$, properties of H_n were studied in the context of Kullback-Leibler distance by Hall [103]. Under some conditions the analysis by Hall and Morton [105] yields a root-n consistent estimate of the entropy when $1 \leq d \leq 3$.

4.1.5 Convergence properties of discrete Plug-in estimates

Convergence properties of discrete plug-in estimators were studied by Antos and Kontoyiannis [9] in a more general scope. They investigated a class of additive functionals where a discrete random variable is given by its distribution $\{p(i); i \in \mathcal{H}\}$. The plug-in estimate for F is defined by

$$\hat{F_n} = g((\sum_{i \in \mathcal{H}} f(i, p(i))),$$

where

$$p_n(i) = \frac{1}{n} \sum_{j=1}^{n} I_{\{X_j = i\}}$$

is the empirical distribution induced by the samples $(X_1, \ldots, X_n)$ on $\mathcal{H}$. In other words, $\hat{F_n} = F(p_n)$. It is assumed that f and g are arbitrary real-valued functions with the only restriction that f is always nonnegative. For additive functionals, including the cases of the mean, entropy, Rényi entropy and mutual information, satisfying some mild conditions, the plug-in estimates of F were shown to be universally consistent and consistent in L_2. The L_2-error of the plug-in estimate is of order $O(\frac{1}{n})$. In other words, in the case of discrete

estimators, the convergence results obtained by Antos and Kontoyiannis [9] are in agreement with the convergence results of the all above mentioned plug-in methods but they were done in general for additive functionals among which all plug-in methods belong (under some mild conditions).

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.

It was shown in [9] that there is no universal rate at which the error goes to zero, no matter what estimator we select, even when our sample space is discrete (albeit infinite). Given any such assumed rate a_N, we can always find some distribution P for which the true rate of convergence is infinitely slower than a_N. Antos and Kontoyiannis [9] proved identical theorems for the mutual information, as well as a few other functionals of P.

4.2 Estimates of entropy based on partitioning of the observation space

This popular class of estimators divides the observation space into a set of partitions. The methods belonging to this class can be classified according to the number of features. 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. In the following, algorithms employing fixed schemes as well as algorithms using adaptive schemes are presented.

4.2.1 Fixed partitioning of the observation space

Consider a pair of random variables x and y with values in the measurable spaces X and Y, respectively. Recalling definitions (6), (9) and (12), their mutual information is

$$I(X,Y) = \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \tag{40}$$

Classical Histogram methods
One of the most straightforward and widely used nonparametric approach to estimate (40) is approximation of the probability distributions $p(x_i, y_j)$, $p(x_i)$ and $p(y_j)$ by a histogram estimation [43]. The range of a variable, say x, is partitioned into m_X discrete bins $a_i \in A$, each with width h_X. Let k_i denotes the number of measurements which lie in the bin a_i. The probability $p(x_i)$ is approximated by relative frequencies of occurrence $p_X(a_i) = \frac{k_i}{N}$, where N is the size (the number of points) of the data set. Analogously, we estimate the probability $p(y_i)$ using elements b_j of the width h_y belonging to the partition B as $p_Y(b_j) = \frac{k_j}{N}$.

The joint probability $p(x_i, y_j)$ is then approximated using the product partition $A \times B$: $p_{X,Y}(a_i \times b_j) = \frac{k_{i,j}}{N}$. Then the estimator of the mutual information is

$$I(X,Y) = \log N + (1/N) \sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} k_{i,j} \log \frac{k_{i,j}}{k_i k_j}, \tag{41}$$

where $k_{i,j}$ is the number of measurements for which x lies in a_i and y in b_j. This method is also referred to as equidistant binning, as all the bins of the histogram have the same size (Fig. 1).

`Insert Fig. 1 here`

It can be demonstrated [221] that the estimate of mutual information given by (41) fluctuates around the true value or gets systematically overestimated. Moreover, these methods fail in higher dimensions and work well only for two or 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.

The accuracy of the histogram entropy estimator is closely related to the histogram problem: given a scalar data set X, how many elements should be used to construct a histogram of X? The histogram problem has a long history and has been examined by several investigators. A systematic theoretic development of the question is given by Rissanen et al. [194], who use a minimum description length argument to conclude that the optimal value of the number of elements to use in a histogram is the value that gives a minimum value of the stochastic complexity. As a representative application of histogram methods to mutual information estimation we mention here Moddemeijer [148] (using a simple histogram-based method in a procedure to estimate time-delays between recordings of electroencephalogram (EEG) signals originating from epileptic animals or patients) or the work from Knuth et al. [133] who introduced so called optimal binning techniques, developed for piecewise-constant, histogram-style models of the underlying density functions.

Generalized binning with B-Splines

Daub et al. [56] developed a method for estimating multidimensional entropies using B-splines. In classical histogram approaches to mutual information estimation, data points close to bin boundaries can cross over to a neighboring bin due to noise or fluctuations, and in this way they introduce additional variance into the computed estimate. Even for sets of moderate size, this variance is not negligible. To overcome this problem, Daub et al. [56] proposed a generalized histogram method, which uses B-spline functions to assign data points to bins. The sample space is divided into equally sized bins as in equidistant binning. The major difference between the classical histogram methods and the generalized binning is that a data point is assigned to multiple bins simultaneously with

weights given by B-spline functions which are implicitly normalized. The shape of the B-spline functions is determined by their order k, which is a parameter of the method. As an example, B-Splines of order 3 are shown for 4 bins in Fig. 2.

`Insert Fig. 2 here`

When B-splines of order 1 are selected, each point is assigned to one bin only and the method is equivalent to simple equidistant binning. The proposed method is therefore a fixed binning scheme extended by a preprocessing step to reduce the variance. This strategy also somewhat alleviates the choice-of-origin problem of classical histogram methods by smoothing the effect of transition of data points between bins due to shifts in origin.

The probability $p(a_i)$ of each bin is estimated by

$$\hat{p}(a_i) = \frac{1}{N} \sum_{j=1}^{N} B_{i,k}(\tilde{x}_j), \tag{42}$$

where $B_{i,k}$ is a B-spline function of order k evaluated at bin i; $\tilde{x}_j$ is an appropriately scaled data sample mapping the values of x into the domain of the B-spline functions [56, 58]. In two dimensions the joint pdf is computed as

$$\hat{p}(a_i, b_j) = \frac{1}{N} \sum_{l=1}^{N} B_{i,k}(\tilde{x}_l)) \times B_{j,k}(\tilde{y}_l)). \tag{43}$$

The mutual information $I_{M,k}(X;Y)$ can then be estimated from

$$I_{M,k}(X;Y) = H_{M,k}(X) + H_{M,k}(Y) - H_{M,k}(X,Y) \tag{44}$$

and each of the terms may be computed using the standard formulas applied to the probabilities (42), (43). The notation $I_{M,k}(X;Y)$ and $H_{M,k}(X,Y)$ indicates that the method has two parameters: M, the number of bins and k, the order of the B-spline. The procedure can be theoretically extended to a higher number of dimensions, but the performance of this estimator in the multidimensional case has not been systematically studied.

Consistency or other properties in the framework of the section 3.1 are not known. Daub et al. [56] gave numerical estimates of bias and variance with data set size of N for the estimator $I_{M,k}(X;Y)$ for statistically independent data sets and for $k = 3$. The $I_{M,3}$ estimator was found to have bias scaling as $\sim 1/N$ but the slope was significantly lower than for the classical histogram method (equivalent to $I_{M,1}$). The same is true for the standard deviation which also scaled as $1/N$, but with a significantly lower slope than $I_{M,1}$.

Daub et al. compared their estimator to the estimator $\hat{H}_{BUB}$ also using binning, introduced by Paninski [167] (more details in Section 4.6) for independent data sets. The scaling behavior of the bias was found to be similar. The standard deviation of their algorithm is however lower than that of $\hat{H}_{BUB}$.

The estimator from Daub et al. [56] was also compared to the kernel density estimator KDE, i.e. [151], [215], [221].

4.2.2 Adaptive partitioning of the observation space

Marginal equiquantization

Any method for computation of mutual information based on partitioning of data space is always connected with the problem of quantization. By the quantization we understand 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 described above). 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: Estimating the joint entropy of n variables using q marginal bins one obtains q^n boxes covering the state space. If the value q^n approaches the number N of data samples, or even $q^n > N$, the estimate of $H(X_1, \ldots, X_n)$ can be equal to $\log N$, or, in any case, it can be determined more by the number of data samples and/or by a number of distinct data values than by a structure in the data, i.e., by properties of the system under study. In such a case we say that the data are overquantized. Even a "natural" quantization of experimental data given by an A/D (analog to digital) converter can be too fine for reliable estimation of the mutual information from limited number of samples.

Emergence of overquantization is given by the number of boxes covering the state space, i.e., the higher the space dimension (the number of variables), the lower the number of marginal quantization levels that can cause the overquantization. Recalling the definition of mutual information by formula (12), one can see that while the estimate of the joint entropy can be overquantized, i.e., saturated on a value given by the number of the data samples and/or by the number of distinct data values, the estimates of the individual (marginal) entropies are not and they increase with fining the quantization. Thus the overquantization causes an overestimation of the mutual information and in the case of the lagged mutual information, it obscures its dependence on the lag τ [163, 161].

As a simple data adaptive partitioning method, Paluš [164, 161, 163] used a simple box-counting method with marginal equiquantization. It means that 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. An example of an equiquantized observation space is in Fig. 3.

`Insert Fig. 3 here`

In Ref. [161] 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 mutual information estimate, even in its coarse-grained version, is invariant against any monotonous (and nonlinear) transformation of the data [165]. Due to this property, the mutual information, 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 [161].

Equiprobable binning was recently used also by Celluci et al. [44], however, the number of bins is determined using the minimum description length criterion. It is proposed in their work that calculation of mutual information should be statistically validated by application of a χ^2 test of the null hypothesis of statistical independence. Additionally, the partition of the XY plane, which is used to calculate the joint probability distribution P_{XY}, should satisfy the Cochran criterion on the expectancies E_{XY} [44]. A procedure for a non-uniform XY partition is proposed which reduced sensitivity to outlying values of X and Y and provides an approximation of the highest partition resolution consistent with the expectation criterion.

Celluci et al. compare this simple algorithm, adaptive in one dimension, with the locally data adaptive approach of Fraser and Swinney [78] which is technically also equivalent to the Darbellay-Vajda algorithm [54, 55] (more details in the following section). The latter approach is probably the method with the smallest bias provided the unlimited amount of data. Using the limited number of samples (less then 10^4 samples), this algorithm introduced false structures and the simple marginal equiquantization method gives better results, not to speak about the CPU time used (see the comparison done by Celluci et al. [44]).

It should be noted that while Fraser and Swinney algorithm uses a χ^2 criterion to control subdivisions of the XY plane locally, it does not, in contrast to the algorithm proposed by [44], provide a global statistical assessment of an $I(X,Y)$ calculation that includes the probability of the null hypothesis of statistical independence.

Adaptive partitioning in two (and more) dimensions

Darbellay and Vajda [54, 55] demonstrated that mutual information can be approximated arbitrarily closely in probability (i.e. the weak consistency was proven) by calculating relative frequencies on appropriate partitions and achieving conditional independence on the rectangles of which the partitions are made. This method was experimentally compared to maximum-likelihood estimators (see Sec. 4.6). The partitioning scheme used by Darbellay and Vajda [54, 55] (described below) was originally proposed by Fraser and Swinney [78, 79] and in physical literature is referred to as the Fraser-Swinney algorithm. Darbellay and Vajda [54, 55] proved the weak consistency of this estimate and tested the method on a number of probability distributions. In the mathematical and

information-theoretic literature the method has recently been referred to as the Darbellay-Vajda algorithm.

The consistency proof of Darbellay and Vajda [54, 55] starts from the following definition of mutual information due to Dobrushin [67]:

$$I(\mathbf{X_a}, \mathbf{X_b}) \equiv \sup_{\{A_i\}\{B_i\}} \sum_{i,j} P_X(A_i \times B_j) \log \frac{P_X(A_i \times B_j)}{P_{X_a}(A_i) P_{X_b}(B_j)} \tag{45}$$

where X_a, X_b are random vectors with values in R^{d_a}, R^{d_b} respectively.

$P_{X_a}(A_i)P_{X_b}(B_j) = (P_{X_a} \times P_{X_b})(A \times B)$ is the product measure defined as the probability measure for A, B elements of the respective σ-algebras of R^{d_a} and R^{d_b}.

The supremum in (45) is taken over all finite partitions $\Gamma_a = \{A_i | i \in 1, ..., m\}$ of R^{d_a} and all finite partitions $\Gamma_b = \{B_j | j \in 1, ..., n\}$ of R^{d_b}.

A finite partition of a set X is any finite system of sets $\Gamma = \{C_k | k \in 1, ..., q\}$ which satisfies $C_i \cap C_j = \emptyset$ for $i \neq j$ and $\bigcup_{k=1}^{q} C_k = X$. Each set C_k is called a *cell* of the partition Γ. A partition $\Lambda = \{D_l | l = 1, ..., r\}$ is a *refinement* of the partition Γ if for each $D_l \in \Lambda$ there exists $C_k \in \Gamma$ such that $D_l \subset C_k$.

Darbellay [55] notes that the sequence of numbers

$$D_\Gamma \equiv \sum_k P_X(C_k) \log \frac{P_X(C_k)}{P_{X_a \times X_n}(C_k)} \tag{46}$$

never decreases as the partition Γ is made successively finer and finer. An important fact with respect to the developed algorithm is that if Λ is a refinement of the partition Γ such that $C_k \in \Gamma = \bigcup_l D_{k,l} \in \Lambda$ for some set of indices l depending on k then

$$D_\Gamma = D_\Lambda \iff \frac{P_X(D_{k,l})}{P_{X_a \times X_b}(D_{k,l})} = \frac{P_X(C_k)}{P_{X_a \times X_b}(C_k)} \forall k, l. \tag{47}$$

This means that the random vectors $\mathbf{X_a}, \mathbf{X_b}$ must be conditionally independent if they attain values in the cell C_k. If this is true for all cells C_k, then the mutual information $I(\mathbf{X_a}, \mathbf{X_b})$ can be estimated as D_Γ.

The algorithm works with d-dimensional hyperrectangles. To split a given cell, each of its d edges is split into $\alpha \geq 2$ equiprobable intervals (marginal equiquantization). At every partitioning step, a cell is split into α^d subcells. Initially, the entire space $\mathcal{R}^d$ is one cell. The algorithm follows by first checking the condition on the right side of (47) for each cell C_k and if the cell does not satisfy the condition, then it is split by marginal equiquantization. The parameter α is usually set to 2 since the recursive nature of the algorithm allows further splitting of regions where conditional independence is not achieved. Fig. 4 illustrates how such a partition might look like.

`Insert Fig. 4 here`

It is advantageous to combine all the conditional independence tests (47)

into one statistic. Here the $\chi^2(\{D_{k,l}\})$ statistic is used. To increase robustness of the test, when testing for conditional independence, the splitting can be done at multiple levels of refinement $t = 1, 2, ...\beta$. This means that the cell C_k is broken into α^{td} cells for each t. A higher value of β prevents the algorithm from stopping the partitioning process too early, while a value of β too high might force the splitting process to continue until there is a very small number of points in each partition. The choice of β also depends on the number of points available and the depth of the cell. When testing the algorithms numerically, β was set to 2 for partitioning depth up to 3 and then to 1 for deeper cells if the problem was more than 2-dimensional. For 2-dimensional problems the authors used $\beta = 2$. This setup provides some guidelines for selecting the values of α and β.

The estimator was tested on correlated Gaussians, where another estimate of the mutual information is available via a maximum likelihood estimator. The recursive space partitioning estimator appears to be asymptotically unbiased and numerical tests also suggest that it is $\sqrt{N}$-consistent for a large class of distributions. The estimator is not universally consistent since the examples of distributions, where the estimator does not converge, are known. These distributions are however rather 'exotic' (e.g. exhibiting some symmetries, which prevent the conditional independence test to succeed). Asymptotically, the estimates of mutual information $\hat{I}$ have a normal distribution, except for very low values of mutual information.

4.3 Ranking

Pompe [179] proposed an estimator of dependencies of a time series based on second order Rényi entropy (more details on Rényi entropy in section 4.5.1). In general, Rényi entropy does not possess the attractive properties of Shannon entropy, such as non-negativity and it is not possible to infer independence with vanishing Rényi entropy. However, 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) [94, 225].

Consider a stationary discrete time series $\{X_t\}$ such that X_t attains one of k different values $x(n), n \in 1, 2, ..., k$.

The statistical dependency between a d-dimensional vector of 'past' values $\vec{X}_t = (X_{t-\Theta_{d-1}}, ..., X_{t-\Theta_0})$ and one 'future' value $X_{t+\tau}$ is examined, d is $1, 2, 3, ...$ and $\Theta_{d-1} > \Theta_{d-2} > ... > \Theta_0 = 0$ and also $\tau \geq 0$. The joint probabilities are denoted as

$$p_{m_{d-1},...,m_0,n}(\tau) = p\{X_{t-\Theta_{d-1}} = x(m_0), ..., X_{t-\Theta_0} = x(m_0), X_{t+\tau} = x(n)\} \tag{48}$$

where each of the indices $m_i, i \in \{d-1, ..., 0\}$ and n are in $1, 2, ..., k$. The vector of values $(m_{d-1}, ..., m_0)$ is hereafter denoted as $\vec{m}$. Using the above notation, Pompe defines the contingency

$$\varphi_d^2(\tau) \equiv \sum_{\vec{m},n=1}^{k} \frac{[p_{\vec{m},n}(\tau) - p_{\vec{m}}p_n]^2}{p_{m_{d-1}} \cdots p_{m_0} p_n}. \tag{49}$$

The key point is the assumption that X_t is *uniformly distributed*. All the probabilities thus satisfy $p_{m_{d-1}} = ... = p_{m_0} = p_n = 1/k = \epsilon$ and it is possible to rewrite the equation for contingency

$$\varphi_d^2(\tau) \equiv \epsilon^{-(d-1)} \sum_{\vec{m},n=1}^{k} p_{\vec{m},n}^2(\tau) - \epsilon^{-d} \sum_{\vec{m},n}^{k} p_{\vec{m}}^2. \tag{50}$$

The contingency $\phi_d^2(\tau)$ can be related to the *generalized mutual information*

$$I_d^{(2)}(\tau) \equiv H_1^{(2)} + H_d^{(2)} - H_{d+1}^{(2)}(\tau) \tag{51}$$

with the formula

$$I_d^{(2)}(\tau) = \log\left(\frac{\varphi_D^2(\tau)}{\epsilon^{-d} \sum_{\vec{m}=1} p_{\vec{m}}^2} + 1\right). \tag{52}$$

From the equation (52) it can be seen that $I_d^{(2)}(\tau) \geq 0$ because $\varphi_d^2(\tau) \geq 0$ and the argument of the logarithm is always ≥ 1. It was also noted in [179] that $I_d^{(2)}(\tau) = 0$ iff $\vec{X}_t$ and $\vec{X}_{t+\tau}$ are independent, i.e. it is true that $p_{\vec{m},n}(\tau) = p_{\vec{m}}p_n$ for all combinations $\vec{m}, n$. It is further proven that

$$0 \leq I_d^{(2)}(\tau) \leq H_1^{(2)} = \log k = -\log \epsilon, \tag{53}$$

always assuming the uniform distribution of data.

The transformation of an arbitrarily distributed time series to a uniform distribution is accomplished by sorting the samples using some common fast sorting algorithm such as *quicksort* or *heapsort* and replacing each sample by its rank in the sorted sequence. If the generating process Y_t is continuous then the above transformation function would be equivalent to the distribution function of Y_t. A side-effect is that the estimation of $I_d^{(2)}$ in this way is insensitive to non-linear invertible distortions of the original signal. After the transformation, Pompe recommends estimating the *generalized mutual information* using the GPTA algorithm as

$$I_d^{(2)}(\tau) \simeq \log \frac{C_{d+1,\epsilon/2}(\tau)}{C_{d,\epsilon/2} C_{1,\epsilon/2}} \text{ for } \epsilon \to 0, \tag{54}$$

where $C_{X,\delta}$ represents the correlation integral ([94], [108], [106]) with the appropriate time-delay embedding $\vec{X}_t$ for $C_{d,\epsilon/2}$, $(\vec{X}_t, X_{t+\tau})$ for $C_{d+1,\epsilon/2}$, $X_{t+\tau}$ for $C_{1,\epsilon/2}$ and the neighborhood size δ.

The above considerations are however not entirely applicable to the quantized and sampled time-series, as the possibility of two different samples coinciding is non-zero. In fact this commonly occurs in practice. Under these conditions

it is not possible to obtain a unique ranking of the original sequence as equal samples can be arbitrarily interchanged. Pompe suggests using a neighborhood size equal to at least $\rho_{max}\epsilon_q$, where ϵ_q represents the relative quantization error and ρ_{max} is the maximum value of the one-dimensional distribution density of the original data. In practice ρ_{max} would be estimated as $l_{eq,max}/T$ where $l_{eq,max}$ is the maximal number of equal data samples in the series. By adhering to this policy, problems stemming from the non-uniqueness of the samples are circumvented.

4.4 Estimates of entropy and mutual information based on computing distances

4.4.1 Based on sample spacings

These methods are defined only for $d = 1$ and their generalization to multivariate cases is not trivial. Let $X_1, \ldots, X_2$ be a sample of i.i.d. real valued random values and let $X_{n,1} \leq X_{n,2} \leq \ldots, X_{n,n}$ be the corresponding order statistics. Then $X_{m,i+m} - X_{n,i}$ is called m-spacing ($1 \leq i \leq i+m \leq$n). Based on spacings, it is possible to construct a density estimate:

$$f_n(x) = \frac{m}{n} \frac{1}{X_{n,im} - X_{n,(i-1)m}} \tag{55}$$

if $x \in [X_{n,(i-1)m}, X_{n,im})$. This density estimate is consistent if for $n \to \infty$ hold

$$m_n \to \infty, m_n/n \to 0. \tag{56}$$

The estimate of entropy based on sample spacings can be derived as a plug-in integral estimate or resubstitution estimate using a spacing density estimate. Surprisingly, although the m-spacing density estimates might not be consistent, their corresponding m-spacing entropy estimates might turn out to be (weakly) consistent [101].

(i) *m-spacing estimate for fixed m has the form*

$$H_{m,n} = \frac{1}{n} \sum_{i=1}^{n-m} \log(\frac{n}{m}(X_{n,i+m} - X_{n,i})) - \psi(m) + \log m \tag{57}$$

where $\psi(x) = -(\log \Gamma(x))'$ is the digamma function. Then the corresponding density estimate is not consistent. This implies that in (57) there is an additional term correcting the asymptotic bias. For uniform f the consistency of (57) was proven by Tarasenko [227] and by Beirlant and van Zuijlen [20]. Hall proved the weak consistency of (57) for densities satisfying T2 and P2 [101]. The asymptotic normality of $H_{m,n}$ was studied under the conditions T2 and P2 in Refs. [53], [68], [101] and [21], who all proved the asymptotic normality under T2 and P2 with

$$\sigma^2 = (2m^2 - 2m + 1)\psi'(m) - 2m + 1 + Var(\log f(X)), \tag{58}$$

which for $m = 1$ gives

$$\sigma^2 = \frac{\pi^2}{6} - 1 + Var(lnf(X)).$$

(ii) *m_n-spacing estimate with $m_n \to \infty$*

This case is considered in the papers of Vašíček [243], Dudewitz and van der Meulen [68], Beirlant and van Zuijlen [20], and of Beirlant [21], Hall [106], van Es [239]. In these papers, the weak and strong consistencies are proven under condition (56). Consistencies for densities with unbounded support is proved only in Tarasenko [227] and in Beirlant [20]. Hall [106], van Es [239] proved asymptotic normality with $\sigma_2 = Var(\log f(X))$ if f is not uniform but satisfies T2 and P2. Hall in [106] showed this result also for the non-consistent choice of $M_n/n \to \rho$ if ρ is irrational. This asymptotic variance is the smallest one for an entropy estimator if f is not uniform. If f is uniform on $[0, 1]$ then Dudewitz and van Meulen [68] and van Es [239] showed, respectively for $m_n = o(n^{1/3-\delta}), \delta > 0$, and for $m_n = o(n^{1/3})$ that

$$\lim_{n\to\infty} (mn)^{1/2}(\hat{H}_n - H(f)) = N(0, 1/3), \tag{59}$$

for slight modifications $\hat{H}_n$ of the m_m-spacing estimate H_n. (Since sample spacings are defined only in one dimension, for our application are not these methods suitable. The generalization of these estimates is in higher dimension non-trivial.)

4.4.2 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 [67] and by Vašíček [243], but they cannot be directly generalized to higher dimensional spaces (and therefore not applied to mutual information).

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. More precisely, let $\rho_{n,i}$ be the nearest neighbor distance of X_i and the other $X_j : \rho_{n,i} = min_{j\neq i, j\leq n}||X_i - X_j||$. Then the *nearest neighbor entropy estimate* is defined as

$$H_n = \frac{1}{n}\sum_{i=1}^{n} \log(n\rho_{n,i}) + \log 2 + C_E, \tag{60}$$

where C_E is the Euler constant $C_E = -\int_0^\infty e^{-t} \log t dt$. Under the condition (P1) introduced in Section 3.1, Kozachenko and Leonenko [135] proved the mean square consistency for general $d \geq 1$. Tsybakov and van der Meulen [237] showed root-n rate of convergence for a truncated version of H_n when $d = 1$ for a class of densities with unbounded support and exponential decreasing tails,

such as the Gaussian density. Bickel and Breiman [32] considered estimating a general functional of density. Under general conditions on f they proved asymptotic normality. Their study unfortunately excludes the entropy.

We will describe here more in detail two nearest neighbor entropy estimators: KL introduced by Kozachenko and Leonenko [135] and its theoretical analysis and then its improved modification from Kraskov et al. [136].

Victor [248] applied the KL estimator and claimed that the algorithm was dramatically more efficient than standard bin-based approaches, such as the direct method from Strong at al. [222] for amounts of data typically available from laboratory experiments.

The KL estimator

For simplicity reasons, we describe the estimators in R^2. The idea is to rank, for each point $z_i = (x_i, y_i) \in R^2$ its neighbors by distance $d_{i,j} = \|z_i - z_j\| : d_{i,j_1} \leq d_{i,j_2} \leq \ldots$ (supposing $\|.\|$ be a metrics) and than to estimate $H(X)$ from the average distance to the $k-$nearest neighbor, averaged over all x_i.

Shannon entropy $H(X) = -\int dx \mu(x) \log \mu(x)$ can be understood as an average of $\log \mu(x)$. Having an unbiased estimator $\widehat{\log \mu(x)}$ of $\log \mu(x)$, one would get an unbiased estimator $\widehat{H(X)} = -\frac{1}{N} \sum_{i=1}^{N} \widehat{\log \mu(x_i)}$. In order to estimate $\widehat{\log \mu(x_i)}$, the probability distribution $P_k(\epsilon)$ is considered for the distance between x_i and its $k-th$ nearest neighbor. The probability $P_k(\epsilon)d\epsilon$ can be derived from the trinomial formula

$$P_k(\epsilon) = k \binom{N-1}{k} dp_i(\epsilon)/d\epsilon p_i^{k-1} (1-p_i)^{N-k-1}. \tag{61}$$

The expectation value of $\log p_i(\epsilon)$ can be from $P_k(\epsilon)$ derived as $E(\log p_i) = \psi(k) - \psi(N)$, where $\psi(x)$ is the digamma function (i.e. logarithmic derivative of the gamma function, see [2]). The expectation is taken over the positions of all other $N-1$ points, x_i is kept fixed. An estimator for $\log \mu(x)$ is then obtained by assuming that $\mu(x)$ is constant in the whole $\epsilon > 0$ ball. This gives $p_i(\epsilon) \approx c_d \epsilon^d \mu(x_i)$, where d is the dimension of x and c_d is the volume of the d-dimensional unit ball. (For the maximum norm $c_d = 1$, for the Euclidean $c_d = \pi^{d/2}/\Gamma(1 + d/2)/2^d$). Then $\log \mu(x_i) \approx \psi(k) - \psi(N) - dE(\log \epsilon) - \log c_d$, which leads to

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log(c_d) + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon(i) \tag{62}$$

where $\epsilon(i)$ is twice the distance from x_i to its k-th nearest neighbor. In most investigated cases (including Gaussian and uniform densities in bounded domains with a sharp cutoff) the approximation error is approximately k/N or $k/N \log(N/k)$ [136].

Mutual information can be obtained by estimating $H(X)$, $H(Y)$, and $H(X,Y)$ separately and by applying formula (12). But in this way, the errors made in the individual estimates would not have to cancel (see also the discussion below).

Leonenko et al. [139] studied a class of k-nearest-neighbor-based Rényi estimators for multidimensional densities (as we have already mentioned above, Shannon entropy is Rényi entropy for $q = 1$). They investigated theoretically a class of estimators of the Rényi and Tsallis (also called Havrda-Charvát, see i.e. [51]) entropies of an unknown multidimensional distribution based on the k-nearest distances in a sample of independent identically distributed vectors. 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 both versions of KSG algorithm of the MI estimators $I^{(1,2)}$ 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). These statements in a more general form for Rényi and Tsallis entropies were proven in [139]. Kullback-Leibler distance (KLD) of two functions in d-dimensional Euclidean space was also examined and for its nearest neighbor estimator similarly proven to be asymptotically unbiased and consistent. A central limit theorem for functions h of the nearest neighbor method was proven by Bickel and Breiman (for $k = 1$ in [32]) and Penrose (for $k > 1$ in [172]) but only under the condition that the nearest neighbor estimator computing entropy is bounded (nearest neighbor estimators of Rényi entropies, including Shannon entropy are not in general bounded). At present, neither exact nor asymptotic results on the distribution of the k-nearest neighbor entropy estimator are known. Goria et al. [87] presented a simulation study showing that for many distribution families used in statistics, the hypothesis of asymptotic normal distribution of the nearest neighbor estimator seems to be acceptable (for Beta, Cauchy, Gamma, Laplace and Student t distributions). It was shown that by increasing of parameter k, one can influence the approximation precision in higher dimensional spaces (the estimator with bigger k was much more accurate as for $k = 1$). Similarly, by setting $k = k_N$, the precision can be influenced for increasing N [139]. For the Kullback-Leibler divergence (KLD), three various nearest neighbor estimators were tested on one-dimensional function f given by 10000 data points generated by a Student distribution with 5 degrees of freedom (t_5). All the three tested estimators converged to t_5. Although the asymptotical unbiasedness and consistency of the estimators was proven for multidimensional spaces, up to our best knowledge, any functional relationship of k with respect to the approximation error of the estimate is not known.

The KSG estimators

The following algorithm is from Kraskov, Stögbauer and Grassberger (KSG) [136]. Assume $Z = (X, Y)$ to be the joint random variable with maximum norm. The estimator differs from formula (62) that vector x is replaced by z,

dimension d is replaced by $d_Z = d_X + d_Y$ and volume c_d by $c_{d_X} c_{d_Y}$. We get

$$\hat{H}(X,Y) = -\psi(k) + \psi(N) + \log c_{d_X} c_{d_Y} - ((d_X + d_Y)/N) \sum_{i=1}^{N} \log \epsilon(i), \quad (63)$$

where $\epsilon(i)$ is twice the distance from x_i to its k-th neighbor. Now the formula (12) for mutual information can be applied for the same k. In this way, the different distance scales would be effectively used in the joint and marginal spaces. However, the biases of formula (62) resulting from the nonuniformity of the density would be different for the estimates $H(X), H(Y)$ and $H(X,Y)$ and would not cancel. To avoid this, Kraskov et al. recommend not to use fixed k for marginal entropy estimation. Two estimators are proposed. Assume that the k-th neighbor of x_i is on one of the vertical sides of the square of size $\epsilon(i)$ (in two dimensions). Then if there are altogether $n_x(i)$ points within the vertical lines $(x_i - \epsilon(i)/2, x_i + \epsilon(i)/2)$, then $\epsilon(i)/2$ is the distance to the $(n_x(i)+1)$-th neighbor of and x_i and

$$\hat{H}(X) = -\frac{1}{N} \sum \psi[n_x(i) + 1] + \psi(N) + \log c_{d_X} + \frac{d_X}{N} \sum_{i=1}^{N} \log \epsilon(i). \quad (64)$$

For the coordinate Y, this is not exactly true, i.e. $\epsilon(i)$ is not exactly equal twice the distance to the $(n_y(i)+1)$-th neighbor if $n_y(i)$ is analogously defined as the number of points in $(y_i - \epsilon(i)/2, y_i + \epsilon(i)/2)$. The first estimator uses hyper-cubes in the joint space and is given in dimension d by

$$I^{(1)}(X_1, \ldots, X_d) = \psi(k) - (d-1)\psi(N) - < \psi(n_{x_1}) + \cdots + \psi(n_{x_d}) > \quad (65)$$

where $< \cdots >= (1/N) \sum_{i=1}^{N} E[\ldots(i)]$ and n_{x_i} is the number of points x_j so that $\|x_j - x_j\| < \epsilon(i)/2$. The second estimate uses hyper-rectangles and is given in dimension d by

$$I^{(2)}(X_1, \ldots, X_d) = \psi(k) - \frac{d-1}{k} + (d-1)\psi(N) - < \psi(n_{x_1}) \cdots + \psi(n_{x_d}) > . \quad (66)$$

More details can be found in [136]. Both estimators (for $k = 1$) for correlated Gaussian distributions give approximately the same results, only in very high dimensions gives $I^{(2)}$ better results because $\epsilon(i)$ tends to be much larger than the marginal $\epsilon_{x_j}(i)$. Some hints for selection of parameter k, influencing the precision of approximation, can be found in [136]. Both estimators appeared in the experiments adaptive (i.e. the resolution is higher where data are more numerous) and had minimal bias.

Fig. 5 (a) shows how $\epsilon(i), n_x(i)$, and $n_y(i)$ are determined in the first algorithm ($I^{(1)}$), for $k = 1$ and some fixed i. In this case, $n_x(i) = 5$ and $n_y(i) = 3$. The two bottom images of Fig. 5 show how to find $\epsilon_x(i), \epsilon_y(i), n_x(i), n_y(i)$ in the second algorithm ($I^{(2)}$) for $k = 2$. The left image (b) indicates the case where

the above are determined by a single point and the right image (c) depicts a situation where two different points influence the values $\epsilon_x(i), \epsilon_y(i), n_x(i), n_y(i)$.

Insert Fig. 5 here

KSG has the following important property. Numerically, both estimators become exact for independent distributions, i.e. the estimator error vanishes (up to statistical fluctuations) if $\mu(x, y) = \mu(x)\mu(y)$ as confirmed by experiments. The results for correlated Gaussians are shown in Fig. 6. This holds for all tested marginal distributions and for all dimensions of x and y (see below). Many points in a large set may have identical coordinates. In that case, the numbers $n_x(i)$ and $n_y(i)$ need no longer be unique (the assumption of continuously distributed points is violated). The nearest neighbor counting can lead to wrong results. [136] solved this by adding very low-amplitude noise to the data. The k-nearest neighbor search in KSG is done by so called Box-assisted algorithm from Grassberger [93]. This algorithm is recommended to be used with KSG in lower dimensions (up to 3), while k-d trie (a data representation structure similar to k-d trees) showed up to be considerably more advantageous in higher dimensional spaces (Vejmelka and Hlaváčková -Schindler [244]). The estimators were applied to assess the actual independence of components obtained from independent component analysis (ICA), to improve ICA and to estimate blind source separation. Rossi et al. [195] applied the KSG estimator of MI in higher dimensional spaces to selection of relevant variables in spectrometric nonlinear modeling, another application is from Sorjamaa et al. [219].

Insert Fig. 6 here

The KSG method was experimentally compared to the adaptive partitioning method from Darbellay and Vajda [54] and was slower. On the other hand, mutual information estimated by KSG estimates $I^{(1,2)}$, worked for more non-Gaussian general distributions, where the adaptive partitioning method failed. KSG and Edgeworth expansion method (for more details about Edgeworth expansion, see Section 5.2.2) for entropy and MI of Gaussian distributions were experimentally compared in Ref. [241]. One can see that the Edgeworth expansion has an advantage for Gaussian distributions or distributions "close" to Gaussian, since the error is caused only by the cumulants. On the other hand, the parameter k gives flexibility both to the KL and KSG estimator to widen its approximation ability to more general distributions.

4.5 Estimates based on learning theory methods

4.5.1 Motivated by signal processing problems

Entropy and mutual information are often used as a criterion in learning theory. For example, classification and pattern recognition literature uses it them

for feature extraction, i.e. Torkkola, [233] or for principal or independent component analysis, i.e. Xu et al. [254], [184], [185]. 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. [174], [186], [70], [71], [205] and [251]. 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. Since these methods work in general also in higher dimensional spaces (and therefore can be applicable to mutual information), they definitely deserve our attention. From this variety of learning theory applications, we mention here the nonparametric estimator of Rényi entropy from Erdogmus [70] and some neural network-based approaches. Concerning the former estimator, we will first explain the Parzen estimation and then the properties of Parzen estimation of Rényi entropy, including Rényi divergence, mutual information and their estimators.

Quadratic Rényi entropy estimator

The nonparametric estimator for Rényi quadratic entropy introduced by Principe and Erdogmus [70] uses Parzen windowing with Gaussian kernels in the following manner. Let the (continuous) quadratic entropy be given by

$$H_2(X) = -\log \int_{-\infty}^{\infty} f_X^2(x)dx. \tag{67}$$

Let $x_i, \ldots, x_N$ are identically distributed samples of the random variable X. The *Parzen (Window) estimate* [170] of the pdf using an arbitrary kernel function $\kappa_\sigma(.)$ is given by

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} \kappa_\sigma(x - x_i), \tag{68}$$

where the kernel function κ_σ is a valid pdf in general and is continuous and differentiable. If Gaussian kernels $G_\sigma(.)$ with standard deviation σ

$$\kappa_\sigma(y) = G(y, \sigma^2 I) = \frac{1}{2/\pi^{M/2}\sigma^M} \exp(-\frac{y^T y}{2\sigma^2}), \tag{69}$$

is substituted into the quadratic entropy expression (67), the following quadratic Rényi entropy estimator is derived by Erdogmus in [70]:

$$\hat{H}_2^{old}(X) = -\log \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma\sqrt{2}}(x_j - x_i). \tag{70}$$

Consider the discrete version of the Rényi entropy written with the expectation operator

$$H_\alpha(X) = \frac{1}{1-\alpha} \log E_X[f_X^{\alpha-1}(X)]. \tag{71}$$

By approximating the expectation operator with the sample mean we get

$$H_\alpha(X) \approx \frac{1}{1-\alpha} \log \frac{1}{N} \sum_{j=1}^{N} f_X^{\alpha-1}(x_j). \tag{72}$$

By substituting the Parzen window estimator into the previous equation, we get

$$H_\alpha^{new}(X) = \frac{1}{1-\alpha} \log \frac{1}{N^\alpha} \sum_{j=1}^{N} (\sum_{i=1}^{N} \kappa_\sigma(x_j - x_i))^{\alpha-1}. \tag{73}$$

For $\alpha = 2$ and Gaussian kernels with standard deviation $\sigma\sqrt{2}$, the old and new estimator become identical. The new estimator can be used for entropy evaluation or when it is desired to adapt the weights of a learning system based on entropic performance index [257]. The new estimator is consistent if the Parzen windowing and the sample mean are consistent for the actual pdf of the iid samples. In case of estimating the joint entropy of an n-dimensional random vector X from its samples $\{x^1, \ldots, x^N\}$, using a multidimensional kernel that is the product of single-dimensional kernels, the estimate of the joint entropy and the estimate of the marginal entropies are consistent [70].

Similarly as for Shannon entropy, the Kullback-Leibler divergence (KLD) is defined also for Rényi entropy. Erdogmus [70] derived analogously a kernel-based resubstitution estimate for Rényi order α divergence. The computational complexity in both cases is $O(N^2)$.

In the Shannon's case is the mutual information between the components of an n-dimensional random vector X equal to the KLD of the joint distribution of X from the product of the marginal distributions of the components X. Rényi order-α mutual information is defined as the Rényi divergence between the same quantities.

Letting $f_X(.)$ be the joint distribution and $f_{X^b}(.)$ be the marginal density of the b^{th} component, *Rényi mutual information* becomes (Rényi, 1976, [192])

$$I^\alpha(X) = \frac{1}{\alpha-1} \log \sum_{i=1}^{N} \cdots \sum_{i=1}^{N} \frac{f_X^\alpha(x_1^i, \ldots, x_n^i)}{\prod_{b=1}^{n} f_{X^b}^{\alpha-1}(x_b^i)}. \tag{74}$$

It is again possible to write kernel-based resubstitution estimator for Rényi mutual information by approximating the joint expectation with the sample mean and then by replacing the pdfs with their Parzen estimators that use consistent kernels between the marginal and joint pdf estimates. The nonparametric mutual information estimator is then

$$\hat{I}_\alpha(X) = \frac{1}{\alpha-1} \log \frac{1}{N} \sum_{j=1}^{N} \left(\frac{(\frac{1}{N}\sum_{i=1}^{N} \prod_{b=1}^{n} \kappa_{\sigma_b}(x_b^j - x_b^i))}{\prod_{b=1}^{n} (\frac{1}{N}\sum_{i=1}^{N} \kappa_{\sigma_b}(x_b^j - x_b^i))}\right)^{\alpha-1}. \tag{75}$$

This estimator can be used in problems where it is necessary to evaluate the mutual information between sets of samples and in adaptation scenarios, where optimizing according to the mutual information between certain variables is the primary objective.

In order to improve the performance by smoothing its learning curve, Erdogmus [70] designed two recursive nonparametric quadratic entropy estimators. One is an exact recursion that provides the exact estimate given by the batch estimator, and the other one a forgetting recursion that incorporates the advantages of a forgetting factor for successful entropy tracking in non-stationary environments. The gradient of the latter estimator directly yields a recursive entropy gradient, called *recursive information gradient (RIG)*. (The stochastic information gradient is shown to be a special case of this corresponding to zero memory, as expected).

Other entropy applications in signal processing are from Bercher and Vignat [27] and Viola [250], [251]; they use spectral-estimation based or polynomial expansion type pdf estimates substituted for the actual pdf in Shannon entropy definition. Viola derived a differential learning rule called EMMA that optimizes entropy by kernel (Parzen) density estimation. Entropy and its derivative can then be calculated by sampling from this density estimate. EMMA was applied for the alignment of three-dimensional models to complex natural images and for detection and correction of corruption in magnetic resonance images. These applications outperform the results done by parametrical methods.

Bercher and Vignat [27] presented an entropy estimator of continuous signals. The estimator relies on a simple analogy between the problems of pdf estimation and power spectrum estimation. The unknown probability density of data is modeled in the form of an autoregressive (AR) spectrum density and regularized long-AR models are applied to identify the AR parameters. The corresponding estimator does not require the explicit estimation of the pdf but only of some samples of a correlation sequence. It was evaluated and compared with other estimators based on histograms, kernel density models, and order statistics. An adaptive version of this entropy estimator was applied for detection of law changes, blind deconvolution, and source separation.

4.5.2 Estimates by neural network approaches

The computation of entropy from a data set by a neural network unfortunately requires explicit knowledge of the local data density. This information is usually not available in the learning from samples case. Schraudolph [204] analyzed three following methods for making density estimation accessible to a neural network: parametric modeling, probabilistic networks and nonparametric estimation (by Parzen window estimator). By imposing their own structure to the data, parametric density models implement impoverished but tractable forms of entropy such as the log-variance (see Section 5).

In the probabilistic networks, neural network node activities are interpreted as the defining parameters of a stochastic process. The net input to such 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. Since the distribution is analytically available, entropy can be optimized directly, but it is the entropy of the network rather than that of the data itself. The result of entropy optimization in such a system therefore depends on the nature of the stochastic model. A well-known example of this type of network is Boltzmann Machine (i.e. [112]). The entropy of the 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 nonparametric estimate of the empirical entropy of Y by Parzen method was derived in the form [204]:

$$\hat{H}(Y) = -\frac{1}{|S|}\sum_{y_i \in S} \log \hat{p}(y_i) = -\frac{1}{|S|}\sum_{y_i \in S} \log \sum_{y_j \in T} \kappa_\sigma(y_i - y_j) + \log |T| \qquad (76)$$

where κ is defined by formula (69). Note that Schraudolph does not use Renyi entropy as is used in the estimators in formulas (70) and (73) but the Shannon one. The Parzen density was used to estimate and optimize the entropy at the output of a parametrized mapping such as a neural network. This resulted in a simple yet efficient batch learning rule that operates on pairs of input samples.

Taleb and Jutten [226] proposed optimization of entropy by neural networks. To optimize the output entropy, one needs to estimate the output pdf (more precisely its derivatives). They suggest to apply a multilayer perceptron (MLP) in unsupervised learning with the weight vector $\mathbf{w}$. Let $\mathbf{x}$ be the input vector, and $\mathbf{y}$ the output of this MLP. The definition of Shannon entropy (5) can be expressed as

$$-E[\log p_Y(y)]. \qquad (77)$$

The weight vector $\mathbf{w}$ is trained (under some constraints) to optimize (77) and the stochastic gradient learning algorithm is

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu_t \nabla \mathbf{w} \mathbf{y}^T \nabla \mathbf{y} \log p_Y(\mathbf{y}), \qquad (78)$$

where μ_t is the learning rate and the sign of the rate depends on if we want to maximize or minimize output entropy. Taleb and Jutten applied a method for the estimation of $\nabla \mathbf{y} \log p(\mathbf{y})$ called score functions. Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_n}) \in R^n$ be a random variable, with differentiable pdf $p_X(x)$. Score function in the multivariate case is defined as:

$$\psi_{\mathbf{X}}(\mathbf{x}) = (\frac{\delta \log p_X(\mathbf{x})}{\delta x_1}, \ldots, \frac{\delta \log p_X(\mathbf{x})}{\delta x_n})^T.$$

Suppose that $\psi_{\mathbf{X}}(\mathbf{x})$ is known. Then, using function approximation ability of neural networks, one can use a simple MLP with one input and one output unit

to provide an estimation $h(\mathbf{w}, x)$ of $\psi_{\mathbf{X}}(\mathbf{x})$. The parameter vector $\mathbf{w}$ is trained to minimize the mean squared error:

$$\varepsilon = \frac{1}{2} E[(h(\mathbf{w}, x) - \psi_{\mathbf{X}}(x))^2]. \tag{79}$$

A gradient descent algorithm on (79) leads to the following weights update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \mu_t \nabla_{\mathbf{w}} \varepsilon,$$

where

$$\nabla \mathbf{w} \varepsilon = E[h(\mathbf{w}, x) \nabla_{\mathbf{w}} h(\mathbf{w}, x) + \nabla_{\mathbf{w}} \frac{\delta h(\mathbf{w}, x)}{\delta x}]. \tag{80}$$

Since $\psi_{\mathbf{X}}(x)$ in (80) disappears, the supervised learning algorithm changes into the unsupervised one. The method can be easily extended into the multivariate case by using a multilayer perceptron with n inputs. To improve the speed of the learning algorithm based on a simple gradient descent, one can use second order minimization techniques. This algorithm was applied to the blind source separation.

To mention another application of entropy estimation, Rigoll [193] used entropy as a learning criterion for perceptron-like networks using self-organizing training. On the other hand, based on the entropy values, the complexity of a neural network approximating a function can be determined [113].

4.6 Entropy estimates based on maximum likelihood

Maximum likelihood estimation (MLE) is a popular statistical method used to make inferences about parameters of the underlying probability distribution of a given data set. The method was pioneered by geneticist and statistician Sir R. A. Fisher already in 1912 [75, 76]. We could classify this approach as well as a parametrical one (Section 5).

When maximizing the likelihood, we may equivalently maximize the log of the likelihood, since log is a continuous (monotonously) increasing function over the range of the likelihood (and the number of calculations may be reduced). The log-likelihood is closely related to entropy and Fisher information (the latter is the negative of expectation of the second derivative of the log of f with respect to θ, where f is the probability function and θ is a parameter). Popular methods for maximum likelihood are the Expectation-Maximization (EM) (i.e. Demster et al. [60] and Berger, Neal and Hinton [152]) and Improved Iterative Scaling (IIS) algorithms (Berger [26]). These methods are often used in classification tasks, especially in speech recognition.

Maximum likelihood estimator

The joint pdf $f(X, \theta)$, given in a parametric form, is computed, where $X = \{x_1, \ldots, x_N\}$ is the set of randomly drawn samples from this pdf. By statistical independence, $f(X, \theta) = \prod_{i=1}^{N} f(x_i, \theta)$, which is known as the likelihood

function of θ. The maximum likelihood (ML) method estimates θ such that the likelihood function takes its maximum value [232], that

$$\hat{\theta}_{ML} = arg \max_{\theta} \prod_{i=1}^{N} f(x_i, \theta). \tag{81}$$

The maximum a posteriori probability (MAP) estimate $\hat{\theta}_{MAP}$ is defined as the point where $f(\theta|X) = f(\theta)f(X|\theta)$ becomes maximum. A method applying log maximum likelihood approach to kernel entropy estimation and using Expectation-Maximization algorithm will be discussed in Section 4.8.

Paninski [167] used an exact local expansion of the entropy function and proved almost sure consistency (strong consistency) and central limit theorems for three of the most commonly used discretized information estimators, namely the maximum likelihood (MLE) estimator $\hat{H}_{MLE}(p_N) = \sum_{i=1}^{N} p_{N,i} \log p_{N,i}$, in our terminology plug-in (see above), the MLE with the so-called Miller-Madow bias correction [146], [145]

$$\hat{H}_{MM}(p_N) = \hat{H}_{MLE}(p_N) + \frac{\hat{m} - 1}{2N} \tag{82}$$

where $\hat{m}$ is some estimate of the number of bins with nonzero $p-$probability ([167] considers $\hat{m}$ to be the number of bins with non-zero p_N probability), and the jackknifed version of MLE from Efron and Stein [69]

$$\hat{H}_{JK} = N\hat{H}_{MLE} - \frac{N-1}{N} \sum_{j=1}^{N} \hat{H}_{MLE-j} \tag{83}$$

where $\hat{H}_{MLE-j}$ is the MLE based on all but the $j-th$ sample.

This framework leads to the estimator $\hat{H}_{BUB}$ (*Best Upper Bounds estimator*) equipped with the bounds on the maximum error over all possible underlying probability distributions; this maximum error is very small. This estimator was applied both on real and simulated data.

Mixture models

Mixture models provide more flexibility into the density estimation. Here, the unknown density is modeled as a mixture of M densities

$$f(x) = \sum_{m=1}^{M} f(x|m) P_m, \tag{84}$$

where $\sum_{m=1} P_m = 1$. Thus, this modeling assumes that each point x may be drawn from any of the M model distributions with probability $P_m, m = 1, \ldots, M$. The density components have a parametric form, $f(x|m, \theta)$ and then the unknown parameters θ and P_m, $m = 1, \ldots, M$ must be computed from the samples. Since the contribution of mixture density is not known, the maximum

likelihood principle cannot be easily employed, and one can apply the EM algorithm to solve the problem. Because of the additional flexibility the mixture models add to the parametric models, this method may be regarded as semi-parametric. There are basically two drawbacks associated with the parametric density estimation schemes discussed above. In the case that an information theoretic measure is used as a cost function for training adaptive systems, the parametric methods require solving an optimization problem within an optimization problem, where the 'external' optimization is the adaption process (using for example gradient-based learning algorithms). The second drawback is the insufficiency of parametric models for general-purpose modeling tasks. The selected parametric family may be too limiting to be able to accurately model the data distributions in question; it may be as well difficult to select the right parametric class.

4.7 Correction methods and bias analysis in undersampled regime

Basharin [18] and Herzel [109] pointed out that to the second order, the bias for an entropy estimation is independent of actual distribution. One can use Bayesian approaches or use very strong assumptions about the estimator to get a small bias, but estimators with very small bias, i.e. [202], [180] have unfortunately large statistical errors. In this subsection we discuss entropy estimates, which are mostly analytical and their bias can be computed.

Let us first consider the simplest and the most straightforward one, the naive ("likelihood") estimator, where one replaces the discrete probabilities $p_i, i = 1, \ldots, K$ (N is the number of observations, and n_i the frequency of realization i among all observations) in the Shannon entropy formula by $\hat{p}_i = \frac{n_i}{N}$. We get

$$\hat{H}_{naive} = -\sum_{i=1}^{K} \hat{p}_i \log \hat{p}_i. \tag{85}$$

$\hat{H}_{naive}$ is also a maximum likelihood estimator S_{ML}, since the maximum likelihood estimate of the probabilities is given by the frequencies. This estimator leads to a systematic underestimation of entropy H (i.e. the difference of the real entropy and its estimator is positive).

Among the first corrections of the estimation error belongs the work of Miller [146], applying a Taylor expansion around p_i to the log function into the naive estimator with the correction term of $O(1/N)$. Paninski [167] applied Bernstein polynomials (i.e. a linear combination of binomial polynomials, from which he derived the estimator $\hat{H}_{BUB}$ discussed in Section 4.6 above, more details [167]) and achieved that the maximum (over all p_i) systematic deviations are of $O(1/N^2)$. Unfortunately, the variance of the corresponding estimator turns out to be very large [167]. Thus a good estimator should have the bounds on bias and variance minimized simultaneously. The result is a regularized least-squares problem. There is however no guarantee that the solution of the regularized

problem implies a good polynomial approximation of the entropy function; this also depends on the priority, what is more important whether reducing bias or variance, or vice versa.

In a more recent work by the same author [168], the entropy estimation is investigated in the undersampled regime (i.e. on m bins given fewer than m samples). It has been long known [146] that the crucial quantity in this estimation problem is the ratio N/m: if the number of samples is much greater than the number of bins, the estimation problem is easy, and vice versa. Paninski concentrated on this part of the problem: how can one estimate the entropy when is N/m_N bounded? He showed that a consistent estimator $H(p_N)$ exists in this regime (by proving his main conjecture from [167]). The most surprising implication of this result is that it is possible to accurately estimate the entropy on m bins, given N samples, even when N/m_N is small (provided that both N and m are sufficiently large).

Nemenman et al. [153, 154] studied properties of near-uniform (Dirichlet) priors for learning undersampled probability distributions on discrete non-metric spaces and entropy and information in neural spike trains. The authors argue that for the estimates of entropy using knowledge of priors, fixing one parameter (beta in the Dirichlet priors) specifies the entropy almost uniquely.

A Bayesian entropy estimator S_{ML} was introduced by Nemenman and Bialek [155] as a maximum likelihood estimator and was applied to synthetic data inspired by experiments and to real experimental spike trains. The estimator S_{ML} was inspired by the Ma's entropy estimation by counting coincidences for uniform distributions. Ma's idea was generalized to an arbitrary distribution. It is well known that one needs $N \sim K$ (N is the size of the data set and K the number of all possible values of a distribution) to estimate entropy universally with small additive or multiplicative errors [167]. Thus the main question is: does a particular method work well only for abstract model problems, or does work also on natural data? The goal of [154] was to show that the method introduced in [153] can generate reliable estimates well into a classically undersampled regime for an experimentally relevant case of neurophysiological recordings.
In an experiment, in N examples each possibility i occurred n_i times. If $N \gg K$, one can use the naive estimator $\hat{H}_{naive}$ given by formula (85). It is known that S_{ML} underestimates entropy ([167]). With good sampling ($N \gg K$), classical arguments due to Miller [146] show that the S_{ML} estimate should be corrected by a universal term $\frac{K-1}{2N}$ (compare to the negative results of Paninski for $K \sim N$ in the mentioned discussion above from [167], where $K = m_N$). There are other correcting approaches, but however they work only when the sampling errors are in some sense a small perturbation. To make progress outside of the asymptotically large N regime, one needs an estimator that does not have a perturbative expansion in $1/N$ with S_{ML} as the zero order term. The estimator S_{ML} from [153] has this property. S_{ML} is the limiting case of Bayesian estimation with Dirichlet priors.

Maximum likelihood estimation is Bayesian estimation with this prior in the limit $\beta \to 0$, while the natural "uniform" prior is $\beta = 1$. The key observation of [153] is that while these priors are quite smooth on the state space of $\mathbf{p}$,

the distributions drawn at random from $\mathcal{P}_\beta$ all have very similar entropies, with a variance that vanishes as K becomes large. This is the origin of the sample size dependent bias in entropy estimation. The goal is to construct a prior on the space of probability distributions which generates a nearly uniform distribution of entropies. It is probable that such a uniform distribution prior would largely remove the sample size dependent bias in entropy estimation, but it is crucial to test it experimentally. In particular, there are infinitely many priors which are approximately (and even exactly) uniform in entropy, and it is not clear which of them will allow successful estimation in real world problems. An estimator S^{NSB} was computed in Ref. [154] and about it proven that the NSB prior almost completely removed the bias in a model problem and that the S^{NSB} is a consistent Bayesian estimator (the derivation of this and of the entropy estimator S^{NSB} can be found in [153] and [154]). Since the analysis is Bayesian, one obtains not only S^{NSB} but also the a posteriori standard deviation, an error bar on our estimate. Secondly, for real data in a regime where undersampling can be beaten down by data, the bias is removed to yield agreement with the extrapolated ML estimator even at a very small sizes. Finally and most crucially, applied to natural and nature-inspired synthetic signals, the NSB estimation performs smoothly and stably over a wide range of $K \gg N$. This opens new possibilities for the information theoretic analysis of experiments.

In the following we focus on other correction methods, which do not use Bayesian analysis. Grassberger [95] derived an estimator (in more general form for Rényi entropy) which is at least asymptotically unbiased for large N, and is also a 'good' approximation in the case of small samples. The corresponding estimator of the Shannon entropy (assumes that the observation space is divided into $M \gg 1$ boxes, each with probability p_i, $\sum_i p_i = 1$ so that each n_i is a random variable with a Poisson distribution) has the form

$$\hat{H}_\psi = \sum_{i=1}^{M} \frac{n_i}{N} (\log N - \psi(n_i) - \frac{(-1)^{n_i}}{n_i(n_i+1)}) \tag{86}$$

where $\psi(n) = \log \Gamma(n)/dn$ is the digamma function. In the case of small probabilities $p_i \ll 1$, is this estimate less biased both than the above mentioned the naive estimator and the Miller's correction estimate.

Grassberger [96] modified the previous estimator into the form

$$\hat{H}_G = \sum_{i=1}^{M} \frac{n_i}{N} (\psi(N) - \psi(n_i) - (-1)^{n_i} \int_0^1 \frac{t^{n_i-1}}{t+1} dt). \tag{87}$$

In the high sampling regime (i.e. $\gg 1$ points in each box), both estimators have exponentially small biases. In the low sampling regime, the errors increase but are smaller than for most other estimators (i.e. Miller's correction [146]). The correction term of the estimator (86) is recovered by a series expansion of the integrand in (87) up to the second order. The higher order terms of the integrand lead to successive bias reductions compared to (86).

Schürmann [209] proposed a class of parametrical entropy estimators and determined their systematical error analytically. The estimator of Shannon entropy is of the form

$$\hat{H}_S(\xi) = \psi(N) - \frac{1}{N}\sum_{i=1}^{M} n_i S_{n_i}(\xi) \tag{88}$$

where $\hat{H}_S(\xi) = \sum_{i=1}^{M} \hat{h}(\xi, n_i)$ and $S_n(\xi) = \psi(n) + (-1)^n \int_0^{1/\xi - 1} \frac{t^{n-1}}{1+t} dt$ and $\hat{h}$ is an estimator of $h(p) = -p \log p$ satisfying $\hat{h}$. For bias of $\hat{h}$ holds

$$b(\xi, p) = -p \int_0^{1-p/\xi} \frac{t^{N-1}}{1-t} dt \text{ and } E[\hat{H}(\xi, n)] = -p \log p + b(\xi, p).$$

This estimator is unbiased for $\xi = p$ and there is a turning point for $\xi = pN$. The estimator is asymptotically unbiased, i.e. $b(\xi, p) \to 0$ for $N \to \infty$ if $\xi \geq p/2$. The mean square error (i.e. statistical error) is $\sigma^2(\xi, p) = E[(\hat{h}(\xi, n) - h(p))^2]$ (where $h(p) = -p \log p$). For $\xi = 1$ the estimator $\hat{h}$ in the asymptotic regime $n \gg 1$ it leads to the Miller's correction. For $\xi = e^{-1/2}$ is the Grassberger estimator $\hat{H}_\psi$ a special case. For $\xi = 1/2$ is the estimator identical to the estimator $\hat{H}_G$ from Grassberger. This estimator is less biased than Miller's correction and the estimator $\hat{h}(e^{-1/2}, n)$, but the statistical error is bigger. The experiments in [205] indicate that it is not possible to decide which estimator should be generally preferred. A good choice of the parameter is always application dependent.

4.8 Kernel methods

Kernel density estimation methods (KDE)

Mutual information was first estimated by this approach by Moon et al. [151]. According to Steuer et al. [221], the KDE methods were found to be superior to the classical histogram methods (see Section 4.2.1) from the following reasons: 1. they have a better mean square error rate of convergence of the estimate to the underlying density; 2. they are insensitive to the choice of origin; 3. the window shapes are not limited to the rectangular window. Kernel density estimator introduced by Silverman [215] in one dimensional space is defined

$$f(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x - x_i}{h}) \tag{89}$$

where h is the kernel width parameter. Kernel function $K(x)$ is required to be a (normalized) probability density function. It follows that also f itself is a probability density. The selection of h is crucial but the methods for selection of h are usually computationally intensive. Silverman suggests, as the optimal width h to use the one which minimizes the mean integrated square error, assuming the underlying distribution is Gaussian:

$$h_{opt} = (\frac{4}{3N}^{1/5})\sigma \approx (1.06\sigma N^{1/5}) \tag{90}$$

where σ denotes the standard deviation of the data. For two dimensional spaces, we use two-dimensional Gaussian kernel estimate

$$F_g(x) = \frac{1}{2\pi N h^2} \sum_{i=1}^{N} \exp^{\frac{d_i(x,y)^2}{2h^2}} \tag{91}$$

where $d_i(x, y)$ is the Euclidean distance of (x, y) from (x_i, y_i). According to Silverman [215], under the assumption that the density F_g is Gaussian, an approximately optimal value is given by

$$h_{opt} \approx \sigma(\frac{4}{d+2})^{\frac{1}{d+4}} N^{\frac{-1}{d+4}} \tag{92}$$

where d is the dimension of the data set and s the average marginal standard deviation. Steuer et al. [221] made objections against a straightforward introduction of a kernel density estimator into the logarithmic formula of mutual information. The reason is that kernel estimation can be used for a continuous form of mutual information while we are interested in the MI of discrete states. The discretization of the (x, y)-plane into infinitesimal bins corresponds to the continuous form of MI

$$I(X, Y) = \int_X \int_Y f(x, y) \log(\frac{f(x, y)}{f(x)f(y)}) dxdy. \tag{93}$$

But such a correspondence does not hold for the individual entropies used in the formula $I(X, Y) = H(X) + H(Y) - H(X, Y)$. The discretization introduced by numerical integration for computing the above integral does not correspond to the partition of data. It is shown in Ref. [221] that the estimated mutual information is much less sensitive than the probability density itself.

Generalized Cross-redundancies

Prichard and Theiler [182] introduced a method to compute information theoretic functionals based on mutual information using correlation integrals. Correlation integrals were introduced by Grassberger and Procaccia ([94]) as

$$C_q(x, \epsilon) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \Theta(\epsilon - ||x_i - x_j||) \tag{94}$$

where q is the order of the integral, N is the number of samples, ϵ is the radius of the neighborhood and $\Theta(x)$ is the Heaviside function. A similar work on correlation integrals is from [108] and [106]. Prichard and Theiler [182] introduced the generalized redundancy

$$I_q(x_1; x_2, l, \epsilon) = H_q(x_1(t), \epsilon) + H_q(x_2(t - l), \epsilon) - H_q(x_1(t), x_2(t - l), \epsilon) \tag{95}$$

which is a time-lagged mutual information functional between $x_1(t)$ and $x_2(t-l)$ parametrized by $\epsilon > 0$ and q which is the order of Rényi entropy. They were inspired by the work of Green and Savit [97] on statistics quantifying dependencies between variables. Setting $q = 1$, mutual information based on Shannon entropy is obtained. This cross-redundancy [182] can be expressed using correlation integrals as

$$I_q(x_1; x_2, l, \epsilon) = -\log_2 \frac{C_q(x_1(t), \epsilon) C_q(x_2(t-l), \epsilon)}{C_q((x_1(t), x_2(t-l)), \epsilon)}. \tag{96}$$

Entropy from maximum likelihood kernel density estimation

Although both maximum likelihood methods and Parzen estimator were discussed already in the previous chapters, we will allow ourselves to present here these methods once more in the framework of multidimensional kernel methods. Schraudolph [205] proposed an estimate of entropy based on kernel density estimation (Parzen window estimation). The underlying assumption is that the probability density $p(\mathbf{y})$ of the generating process is a smoothed version of the empirical pdf of the sample. The estimate based on Parzen windows based on a sample of data T can be written as

$$\hat{p}(\mathbf{y}) = \frac{1}{|T|} \sum_{\mathbf{y}_j \in T} K(\mathbf{y} - \mathbf{y}_j) \tag{97}$$

where K is the kernel. This is an unbiased density estimate of the true density. The kernel used in the work [205] is

$$K(\mathbf{y}) = N(0, \mathbf{\Sigma}) = \frac{\exp(-\frac{1}{2}\mathbf{y}^T \mathbf{\Sigma}^{-1} \mathbf{y})}{(2\pi)^{\frac{n}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}}, \tag{98}$$

with dimensionality n and the covariance matrix $\mathbf{\Sigma}$. The obvious problem is the choice of the covariance matrix $\mathbf{\Sigma}$: in one extreme the estimated pdf will converge to the form of the kernel regardless of the sample distribution and in the other extreme the estimated pdf is too dependent on the particular set of samples in T (thus inducing large variance in the estimate). A suitable kernel between these extremes can be found by the maximum likelihood method. An empirical estimate of the maximum likelihood kernel is the kernel which makes a second sample S drawn independently from the pdf $p(\mathbf{y})$. Usually, the logarithm of the maximum likelihood is maximized for numerical reasons

$$\hat{L} = \log \prod_{\mathbf{y}_i \in S} \hat{p}(\mathbf{y}_i) = \sum_{\mathbf{y}_i \in S} \log \sum_{\mathbf{y}_j \in T} K(\mathbf{y}_i - \mathbf{y}_j) - |S| \log |T|. \tag{99}$$

The estimated log-likelihood from formula (99) (which equals to formula (76) multiplied by $-|S|$) assumes two independent sample sets S and T. In practice, not enough data might be available to create separate sample sets S and T. The data is also likely to be quantized by some measurement process. Both of

the above effects distort the shape of the log-likelihood function and thus can shift the position of the maximum. Schraudolph uses a technique called leave-one-out to ensure $S \cap T = \emptyset$. When estimating the pdf at the sample $\mathbf{y}_i$ the set $T_i = S - \{\mathbf{y}_i\}$ is used. This method ensures optimal use of the sample T while respecting the maximum likelihood requirement $S \cap T = \emptyset$. The quantization effect is mitigated by reintroducing the quantization noise into the kernel

$$K(\mathbf{y}) = \frac{[\exp(-\frac{1}{2}(\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} + \kappa\mathbf{b}^T\boldsymbol{\Sigma}^{-1}\mathbf{b})]}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \tag{100}$$

where $\mathbf{b}$ is the vector of the quantization bin widths in each dimension and $\kappa = \frac{1}{12}$ if the condition $\mathbf{y} \notin T$ holds.

Schraudolph [205] showed how the maximum likelihood $\hat{L}$ can be maximized using gradient ascent for a diagonal $\boldsymbol{\Sigma}$ matrix. Because the performance of standard gradient ascent is not satisfactory due to the shape of the $\hat{L}$ function, it is recommended to use exponentiated gradient ascent with step-size adaptation [205]. Using this approach, the convergence of the method has significantly improved.

Schraudolph also presents an Expectation Maximization (EM) algorithm variant (for the original EM algorithm see i.e. [107]) , which has $\boldsymbol{\Sigma}$ as the only optimized variable, with fixed centers (i.e. data points $\mathbf{y}_i$). This is possible since the kernel used the problem can be understood as the estimation of a mixture of Gaussians. In the E-step, using a given $\boldsymbol{\Sigma}$, so called proximity factors π_{ij} are computed. The proximity factors indicate how is each data point $\mathbf{y}_i$ responsible for the mixture j; they are estimated by

$$\pi_{ij} = \frac{K(\mathbf{y}_i - \mathbf{y}_j)}{\sum_{\mathbf{y}_k \in T} K(\mathbf{y}_i - \mathbf{y}_k)}. \tag{101}$$

In the maximization step of the EM algorithm, the new covariance matrix is computed as the covariance of the proximity weighted data

$$\boldsymbol{\Sigma} = \frac{1}{|S|}\sum_{\mathbf{y}_i \in S}\sum_{\mathbf{y}_j \in T} \pi_{ij}(\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)^T. \tag{102}$$

The convergence of the algorithm is further improved by *overrelaxation*, where the covariance matrix is modified as

$$\boldsymbol{\Sigma}(t) = \boldsymbol{\Sigma}(t)\boldsymbol{\Sigma}^{-1}(t-1)\boldsymbol{\Sigma}(t). \tag{103}$$

It is recommended to use the covariance matrix of the entire sample (uniformly weighted) to initialize the EM algorithm.

An estimate of the Shannon entropy can be computed using the kernel density estimate defined above by formula (76).

4.8.1 Transfer entropy

The causality detection approach based on so-called transfer entropy [206] was introduced and discussed in Sec. 2.4, showing that the expression for the transfer entropy (33) is equivalent to the conditional mutual information defined in the same set-up (dimensions, time lags). Here we remind again the paper of Schreiber [206], now from the point of view of the estimation method.

Schreiber [206] proposed to compute the transfer entropy using the correlation integrals [94]

$$\hat{p}_r(x_{n+1}, x_n, y_n) = \frac{1}{N} \sum_{n'} \Theta \left(r - \left| \begin{pmatrix} x_{n+1} - x_{n'+1} \\ x_n - x_{n'} \\ y_n - y_{n'} \end{pmatrix} \right| \right) , \tag{104}$$

where Θ is a suitable kernel and $|\cdot|$ is a norm. The generalized correlation integral based on time series x_n, y_n approximates the probability measure $p(i_{n+1}|i_n^{(k)}, j_n^{(l)})$. The step kernel $\Theta(x > 0) = 1; \Theta(x \leq 0) = 0$ and the maximum norm are used. It is recommended to exclude dynamically correlated pairs (e.g. using a Theiler window [231]).

The transfer entropy was tested on a lattice of 100 unidirectionally coupled Ulam maps (for the definition see e.g. Ref [238]). The direction of information transfer was correctly shown. Moreover, the bifurcation points, where the behavior of the lattice changed, was identified. The analysis was done using 10^5 data points in each series [206].

Verdes [245] proposed a modification of Schreiber's method which generalizes the above mentioned correlation integral to non-cubic neighborhoods. Instead of using the standard neighborhood, which uses the same ϵ value in each dimension, Verdes uses a neighborhood

$$p^*(x_i, y_i, z_i) = \frac{1}{N_{\text{pairs}}} n(\Delta x_{ij} < \epsilon, \Delta y_{ij} < \Delta_Y, z_{ij} < \delta_Z). \tag{105}$$

Verdes conjectures that due to the limited amount of available data and noise in the data bounds, reasonable δ_Y values from below and for large δ_Y the conditioning has no effect; a possible choice of δ_Y is

$$\delta_Y = \arg\max \frac{n(\Delta x_{ij} < \epsilon, \Delta y_{ij} < \Delta_Y)}{n(\Delta y_{ij} < \delta_Y)}. \tag{106}$$

The value for δ_Z is selected analogically.

5 Parametric estimators

No finite sample can determine density or the entropy directly. Therefore some assumption about either the functional form of the density or about its smoothness can be appropriate in some cases. The most common approach 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. Then, it becomes necessary to estimate the entropy non-parametrically. 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.

When the parametric assumption is violated, the resulting algorithms are incorrect. The most common assumption, that the data follow the Gaussian density, is especially restrictive. An entropy maximization technique that assumes that data are Gaussian, but operates on data drawn from a non-Gaussian density, may in fact end up minimizing entropy.

The popularity of the Gaussian function is based on three considerations: (1) finding the Gaussian that fits the data best is very easy, (2) the entropy of the Gaussian can be directly calculated from its variance, and (3) an affine transformation of a Gaussian random variable remains Gaussian. Entropy of a Gaussian density is $h(X) = -E_X[\log g_\psi(x-\mu)] = \frac{1}{2}\log 2\exp\pi\psi$, where $g_\psi(x-\mu)$ is the Gaussian density with variance ψ and mean μ and E_X is the expectation over the random variable X. It is well known that given a sample set A, the most likely Gaussian density has its mean the mean of A and as its variance the variance of A. As a result, if we assume that a random variable is Gaussian, its empirical entropy is proportional to the log of the sample variance. More simply, when the data is assumed Gaussian, maximizing entropy is equivalent to maximizing variance.

Schraudolph in [204] argues that one does not have to assume a particular shape for the density in order to set up a parametric model for entropy optimization: Let X be a continuous random variable with density $P(x) = Prob[X = x]$, and let Y be the linear function $Y = \sigma X + \mu$. Since the density of Y is given by $P(x) = Prob[X = x] = p((y-\mu)/\sigma)/\sigma$, its entropy is

$$H(Y) = -E[\log(p(y-\mu)/\sigma/\sigma)] = -E[\log p(x) - \log\sigma] = H(X) + \log\sigma.$$

That is, regardless of the shape of pdf $p(x)$, the entropy of a linear function of X scales with the log of the variance. Matching $p(x)$ to empirical data with a linear transformation thus changes the model's entropy in proportion to the log-variance $\log\sigma$ of the data.

5.1 Entropy expressions for multivariate distributions

Verdugo Lazo and Rathie [246] computed a table of explicit Shannon entropy expressions for many commonly used univariate continuous pdfs. Ahmed and Gokhale [4] 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 [11]). Besides an explicit functional form or a smoothness condition for density estimation, one can assume that the pdf could be estimated i.e. by a neural network of a given type. In this way we can define parametric neural network estimators for pdf and then consequently entropy estimation. Neural network entropy estimation was discussed already in section 4.5.2 (since neural network approaches can be classified both as learning theory methods and parametric methods).

5.2 Entropy estimators by higher-order asymptotic expansions

This class includes Fourier Expansion, Edgeworth Expansion and Gram-Charlier Expansion and other expansions [107]. We will discuss here only the last two. An earlier work applying the Gram-Charlier polynomial expansion to MI estimation for a blind separation algorithm is from Hua Yang and Amari [116]. They applied the Gram-Charlier expansion and the Edgeworth expansion (both to the forth order cumulants) to approximate the pdf of the outputs. Their computer simulations showed that Gram-Charlier expansion is superior to the Edgeworth expansion for blind separation.

5.2.1 Mutual information estimation by Gram-Charlier polynomial expansion

Trappenberg et al. [234] introduced a variable selection scheme to a statistical dependency test based on mutual information. They compared several methods for mutual information estimation, namely a standard (equally binned) histogram method (HG), an adaptive partitioning histogram method (AP) from Darbellay and Vajda [54] and the MI estimation based on the Gram-Charlier polynomial expansion (GC) [35].

The CG method of MI estimation is based on the Gram-Charlier polynomial expansion of a probability density function derived by Blinnikov and Moessner [35] in the form

$$f(x) \approx \sum_{n=0}^{\infty} c_n \frac{d^n Z(x)}{dx^n}, \tag{107}$$

where $Z(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$ is a Gaussian function and c_n are factors that determine the weights of different order derivations of $Z(x)$. Using the truncated polynomial expansion for marginal pdf's, Amari et al. [6] derived an approximation of the marginal entropy

$$\hat{H}(x) = \frac{2e\pi}{2} - \frac{(k_3^x)^2}{2.3!} - \frac{(k_4^x)^2}{2.4!} + \frac{(5.k_3^x)^2 k_4^x}{8} + \frac{(k_4^x)^3}{16} \tag{108}$$

where k_3^x and k_4^x are third and fourth order cumulants. Using the fourth order Gram-Charlier expansion for two-dimensional joint pdf, Akaho et al. [5] derived

the joint entropy

$$H(x,y) = H(r,s) + \frac{1}{2}\log(1-\rho^2) \tag{109}$$

where $\rho = E[xy]$, r and s are a linear combination of x and y, $\begin{bmatrix} r \\ s \end{bmatrix} = \begin{pmatrix} c^+ & c^- \\ c^- & c^+ \end{pmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ $c^+ = [(1+\rho)^{-1/2} + (1-\rho)^{-1/2}]/2$, $c^- = [(1+\rho)^{-1/2} - (1-\rho)^{-1/2}]/2$ and $\hat{H}(r,s) = 1 + \log 2\pi - \frac{1}{2.3!}[(\beta_{3,0})^2 + 3(\beta_{2,1})^2 + 3(\beta_{1,2})^2 + (\beta_{0,3})^2] - \frac{1}{2.4!}[(\beta_{4,0})^2 + 4(\beta_{3,1})^2 + 6(\beta_{2,2})^2 + 4(\beta_{1,3})^2 + (\beta_{0,4})^2]$ where $\beta_{k,l} = E\{r^k s^l\} - \beta_0^{k,l}$,

$$\beta_0^{k,l} = \begin{cases} 3 & k = 4 \text{ or } l = 4 \\ 1 & k = l = 2 \\ 0 & \text{otherwise.} \end{cases} \tag{110}$$

Mutual information can then be calculated from these estimates using formula (12), which corresponds to a polynomial of high order cumulants.

The comparison of these three MI estimators can be summarized as follows: The advantage of MI estimation with Gram-Charlier expansion is that it only calculates the expectation value of different powers of the samples. Thus, it is fast and easy to calculate. The disadvantage of the GC method is that the estimate might suffer from the truncation of the expansion in the case of non-Gaussian signals and result into the underestimation of mutual information. The histogram based methods are in this sense more general than polynomial expansion-based methods because they are less sensitive to the bin partitioning. A rough partitioning might result in bias towards high MI, while fine-grained partitions might result in underestimating MI. A good choice of bin width is particularly important for MI estimation as the regions with low data densities carry large information content (such as the tails of a distribution).

5.2.2 Edgeworth approximation of entropy and mutual information

The Edgeworth expansion, similarly as the Charlier-Gram expansion approximates a probability distribution in terms of its cumulants. According to Hall [104], it provides in general accurate approximations to the finite-sample distribution and can be used in deriving the higher-order accuracy of the bootstrap methods. The advantage of the Edgeworth series with respect to the Gram-Charlier series is that the error is controlled, so it is a true asymptotic expansion. Edgeworth expansion is consistent, i.e. in infinity converges to the function which it expands, Cramer [52].

The Edgeworth expansion of a function is estimated in terms of a known distribution f with the same pdf as the function to be approximated, and cumulants κ_i. The density f is generally chosen to be that of normal distribution. Here we mention the definition of the Edgeworth expansion for multivariate density $p(v)$, $v \in R^d$ and up to the fifth order about its best normal estimate $\phi_p(v)$ (i.e. with the same mean and covariance matrix as p) and corresponding

multivariate entropy $H(p)$ is it was used in [242]:

$$p(v) \approx \phi_p(v)(1 + (1/3!)\sum_{i,j,k} \kappa^{i,j,k} h_{ijk}(v) + \kappa) \tag{111}$$

with the $ijk-th$ Hermite polynomial and $\kappa^{i,j,k} = \kappa^{ijk}/(s_i^2 s_j^2 s_k^2)^{-1/2}$ where κ^{ijk} is the sample third cumulant over input dimensions i, j, k, and s_i the sample second central moment over input dimension i. The term κ collects the terms in h_{ijkl} and h_{ijklpq} that are dropped out in the first -order entropy derivation below. (For the precise calculation of the terms of expansion we refer the reader to i.e. Blinnikov and Moessner [35]). Since the differential entropy H(p) = $H(\phi_p) - J(p)$, with J neg-entropy, the $H(p)$ can be approximated as [242]:

$$H(p) \approx H(\phi_p) - \frac{1}{12}[\sum_{i=1}^{d}(\kappa^{i,i,i})^2 + 3\sum_{i,j=1,i\neq j}^{d}(\kappa^{i,i,j})^2 + \frac{1}{6}(\sum_{i,j,k=1,i<j<k}^{d}(\kappa^{i,j,k})^2] \tag{112}$$

which converges on the order of $O(N^{-2})$ with N number of data points. The term $H(\phi_p)$ is the expression for the d-dimensional entropy:
$H(\phi_p) = 0.5 \log |\sum| + \frac{d}{2}\log 2\pi + \frac{d}{2}$, where $|.|$ denotes determinant.

To our best knowledge, the first application of Edgeworth expansions for neg-entropy in the univariate case in the literature was proposed in [122] and was generalized for multivariate case and for entropy and Kullback-Leibler distance in [141].

The Edgeworth approximations for the KLD and neg-entropy in the experiments of Lin et al. ([141]) required in the simulations that the distributions p and p^0 are not far from the Gaussian distribution [103] (to avoid a big approximation error). In the case of differential entropy, one can get it from KLD by using the formula $H(X) = H(p) = H(\phi_p) - K(p, \phi_p)$, where $H(X)$ is given by (6) and p is distribution of the random event. So by using KLD, one can also obtain differential entropy.

To summarize, to approximate both KLD and differential entropy, and consequently mutual information by Edgeworth expansion makes sense only for "close"-to-Gaussian distributions. On the other hand, differential entropy by the Edgeworth expansion avoids the density estimation problems. Furthermore, the order of Edgeworth approximation of differential entropy is $O(N^{-3/2})$ and for KLD $O(N^{-1/2})$, while the density method approximation is of order $O(N^{-1/2})$ where N is size of processed sample. The density estimation cannot be used for differential entropy and KLD estimation for dimension $d > 2$ (because of its speed), while the Edgeworth expansion of neg-entropy produces very good approximations also for more-dimensional Gaussian distributions [141]. Furthermore, the error rate of the histogram estimator depends not only on sample size N but also on the choice of the bandwidth value h; the total error is $O(h^2)+O(N^{-1/2})$ [103]. In the case of histogram, density and kernel estimation, the error is $O(N^{-1/2})$ for dimension $d < 3$. The kernel estimator is much less sensitive to choices of the bandwidth h compared to the histogram estimator.

Finally, both the differential entropy and KLD estimator by Edgeworth expansion can be evaluated for distributions of arbitrary dimension, while the other three mentioned methods can be practically applied only to low-dimensional distributions ($d \leq 3$).
It appears from the above that the most important advantage of the Edgeworth expansion is its applicability in multidimensional spaces. On the other hand, it has the following drawbacks: 1. its behavior on probability distributions that differ significantly from Gaussian distribution and 2. the EE approximation can give approximation of pdf having negative values [190]. Gaztanga et al. [83] addressed the negativity problem by exploring expansions around such pdfs, which would yield positive densities when the variance is large enough and proposed to use Gamma probability function (the Gamma pdf, also called negative binomial or Pearson Type 3 (PT3) arises from the χ^2 distribution with N degrees of freedom, where $1/\rho^2 = N/2$ is taken to be a continuous parameter the Gamma pdf). Gamma expansion has, by construction, the exponential tails and a better general behavior than the Edgeworth expansion, both with respect to the positivity of approximating pdf and the approximation error. The experiments in [83] confirmed that the Gaussian EE has tails dropping quickly to zero as the underlying Gaussian pdf, while the Gamma EE has exponential-type tails. Differences in both expansions might be slight, especially around the peak of pdf, as to the first order both expansions are formally equivalent. Therefore, which one best fits the data set is a matter of careful data analysis. The Gamma EE is a real competitor to the Gaussian EE as it can be generalized to multivariate case.

A multivariate case of the Gaussian EE expansion estimate of differential entropy and mutual information was experimentally compared to the following methods in Ref. [241]: 1-spacings from [101], Parzen window plug-in estimate from [3] and KL and KSG method [135] and [136]. The Edgeworth expansion up to the order three was used and comparisons were performed on the normal and exponential distributions. Each distribution was considered along each dimension and size of the data sample. The best performance results were achieved for the KSG method, the EE method was since it was biased for the exponential distribution.

6 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 [88]. 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 [90]. Therefore, its direct application to nonlinear systems may or may not be appropriate. In the following subsection, we discuss some recent methods to extend the Granger's concept to nonlinear cases.

6.1 Nonlinear Granger causality

Ancona et al. [7] extended Granger causality definition to nonlinear bivariate time series. To define linear Granger causality [88], the vector autoregressive model (VAR) (for two series x and y) is used, which considers the time series x as a vector-weighted sum of both series $\mathbf{X}$ and $\mathbf{Y}$ (similarly for x) and autoregressive predictions (AR) ($x = \mathbf{V}_1.\mathbf{X}$ and $y = \mathbf{V}_2.\mathbf{Y}$, $\mathbf{V}_1$ and $\mathbf{V}_2$ to be estimated by least square fit). A directionality index is introduced measuring the unidirectional, bidirectional influence or uncorrelation. The index

$$D = \frac{c_2 - c_1}{c_1 + c_2} \tag{113}$$

(where $c_1 = \epsilon_x - \epsilon_{xy}$ and $c_2 = \epsilon_y - \epsilon_{yx}$) varies from 1 in the case of unidirectional influence ($x \to y$) to -1 in the opposite case ($y \to x$), with the intermediate values corresponding to bidirectional influence. According to this definition of causality, the following property holds for sufficiently large $M (M = N - m$, N is the length of the time series, m the order of the model).

If the first time series $\mathbf{Y}$ is uncorrelated with $\mathbf{X}$ and x then $\epsilon_x = \epsilon_{xy}$ (where ϵ_x is the estimate of the variance of $x - \mathbf{V}_1.\mathbf{X}$, $\mathbf{V}_1.\mathbf{X}$ is prediction of x, similarly ϵ_y is the estimate of the variance of $y - \mathbf{V}_2.\mathbf{Y}$, $\mathbf{V}_2.\mathbf{Y}$ is prediction of y; ϵ_{xy} and ϵ_{yx} are the prediction errors of the VAR model, defined as the estimated variance of $x - \mathbf{W}_{11}.\mathbf{X} - \mathbf{W}_{12}.\mathbf{Y}$ and $y - \mathbf{W}_{21}.\mathbf{X} - \mathbf{W}_{22}.\mathbf{Y}$, respectively). This means that in this case VAR and AR modelings of the x_i time series coincide. Analogously, if $\mathbf{X}$ is uncorrelated with $\mathbf{Y}$ and y then $\epsilon_y = \epsilon_{yx}$. These properties are fundamental and make the linear prediction approach suitable to evaluate causality. On the other hand, for nonlinear systems higher order correlations may be relevant.

Ancona et al. proposed that any prediction scheme providing a nonlinear extension of Granger causality should satisfy the following property: (P1) *if* $\mathbf{Y}$ *is statistically independent of* $\mathbf{X}$ and x, then $\epsilon_x = \epsilon_{xy}$; *if* $\mathbf{X}$ *is statistically independent of* $\mathbf{Y}$ and y, then $\epsilon_y = \epsilon_{yx}$. The approach applying locally linear models suggested by Hua-Yang and Amari in Ref. [114] for evaluation of nonlinear causality needs very long time series to satisfy $P1$. To construct a method working effectively on moderately long time series, the problem of extending Granger causality can be formulated as finding classes of nonlinear models satisfying property $P1$. Radial basis function method (RBF) [42] is suggested to be applied to the family of models given by

$$\begin{aligned} x &= \mathbf{w}_{11}.\mathbf{\Phi}(\mathbf{X}) + \mathbf{w}_{12}.\mathbf{\Psi}(\mathbf{Y}) \\ y &= \mathbf{w}_{21}.\mathbf{\Phi}(\mathbf{X}) + \mathbf{w}_{22}.\mathbf{\Psi}(\mathbf{Y}) \end{aligned} \tag{114}$$

where $\{\mathbf{w}\}$ are four n-dimensional real vectors, $\mathbf{\Phi} = (\phi_1, \ldots, \phi_n)$ are n given nonlinear real functions of m variables, and $\mathbf{\Psi} = (\psi_1, \ldots, \psi_n)$ are n other real functions of m variables. The prediction errors are given by empirical risks

$$\epsilon_{xy} = \frac{1}{M} \sum_{k=1} M[x^k - \mathbf{w}_{11}\mathbf{\Phi}(\mathbf{X^k}) - \mathbf{w_{12}}\mathbf{\Psi}(\mathbf{Y^k})]^\mathbf{2} \tag{115}$$

$$\epsilon_{yx} = \frac{1}{M} \sum_{k=1} M[y^k - \mathbf{w}_{21}\mathbf{\Phi}(\mathbf{X^k}) - \mathbf{w_{22}}\mathbf{\Psi}(\mathbf{Y^k})]^\mathbf{2}.$$

Fixed $M \gg n$ where n is the number of centers $\{\hat{\mathbf{X}}^\rho\}_{\rho=1}^n$ in the space of $\mathbf{X}$ vectors, are determined by a clustering (or other) procedure applied to data $\{\mathbf{X}\}_{\mathbf{k=1}}^{\mathbf{M}}$. Analogously, n centers $\{\hat{\mathbf{Y}}^\rho\}_{\rho=1}^n$ in the space of $\mathbf{Y}$ vectors, are determined by a clustering applied to data $\{\mathbf{Y}\}_{\mathbf{k=1}}^{\mathbf{M}}$. Ancona et al. suggest to choose

$$\phi_\rho(\mathbf{X}) = \exp(-\|\mathbf{X} - \hat{\mathbf{X}}^\rho\|^\mathbf{2}/\mathbf{2}\sigma^\mathbf{2}), \rho = \mathbf{1}, \ldots, \mathbf{n} \tag{116}$$

$$\psi_\rho(\mathbf{Y}) = \exp(-\|\mathbf{Y} - \hat{\mathbf{Y}}^\rho\|^\mathbf{2}/\mathbf{2}\sigma^\mathbf{2}), \rho = \mathbf{1}, \ldots, \mathbf{n} \tag{117}$$

where ρ is a fixed parameter, whose order of magnitude is the average spacing between the centers. The centers $\hat{\mathbf{X}}^\rho$ are prototypes of the $\mathbf{X}$ variable. Functions ϕ measure the similarity to these typical patterns, analogously, ψ measure the similarity to typical patterns of $\mathbf{Y}$.

The method was tested on chaotic maps and on time series of heart rate and breath rate of a sleeping human suffering from sleep apnea. There is a growing evidence that suggests a causal link between sleep apnea and cardiovascular disease. This data set has been already analyzed by Schreiber in [206], measuring the rate of information flow (transfer entropy), and a stronger flow of information from the heart rate to the breath rate was found. In this example, the rate of information flow entropy and Granger nonlinear causality give consistent results. Both these quantities, in the end, measure the departure from the generalized Markov property [206]. The results in [7] showed that the value of the directionality index D may in some cases be very sensitive to statistical fluctuations, especially when the interdependence is weak.

According to Ancona et al., the standard RBF model of bivariate time series in comparison to formula (115) (as described i.e. by Bishop [34]) does not satisfy in general property $P1$ and therefore is not suited to evaluate causality. Ancona and Stramaglia [8] argued that not all nonlinear prediction schemes are suitable to evaluate causality between two time series, since they should be invariant if statistically independent variables are added to the set of input variables. This property guarantees that, at least asymptotically, one would be able to recognize variables without causality relationship. Marinazzo et al. [143] used the theoretical results from [8] to find the largest class of RBF models suitable to evaluate causality, and in this sense they extended the results of [8]. Moreover, they showed the application of causality to the analysis of cardio-circulatory interaction and studied the mutual influences in inhibitory and excitatory model neurons.

6.2 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 misspecification would arise, leading to spurious causality values. A suitable alternative would be to adopt nonparametric approaches which are free from model mismatch problems. Since the topic of this paper is causality based on information theory, we discuss primarily those nonparametric approaches which can be expressed in the information theoretic terms.

Let us first reformulate the Granger causality in information theoretic terms [64, 62]: 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

$$F_{X_{t+1}}(x|F_X(t), F_Y(t)) \neq F_{X_{t+1}}(x|F_X(t)) \tag{118}$$

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.

Given two time series, the delay vectors are first constructed as follows: $\mathbf{X}_t = (X_t, X_{t-\tau_x}, \ldots, X_{t-\tau_x(d_x-1)})$, and $\mathbf{Y}_t = (Y_t, Y_{t-\tau_y}, \ldots, Y_{t-\tau_y(d_y-1)})$ where time delays are τ_x and τ_y, and embedding dimensions are d_x and d_y, respectively. 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

$$I_q(X, Y) = \log C_q(X, Y) - \log C_q(X) - \log C_q(Y) \tag{119}$$

where the generalized correlation integral [94], C_q can be estimated by

$$C_q(\mathbf{X}, \epsilon) = \frac{1}{N(N-1)^{q-1}} \sum_{j=1}^{N} \Big[\sum_{i \neq j} \Theta(||X_j - X_i|| - \epsilon) \Big]^{q-1}; \tag{120}$$

Θ is the Heaviside function, $\|.\|$ a norm and the last term is related to kernel density estimation. $C_2(X, \epsilon)$ is simply the probability that a distance between two independent realizations of X is smaller than or equal to ϵ. For computational ease, $q = 2$ is preferred [62], though $q = 1$ is also used elsewhere [45]. We refer the interested readers to Ref. [45, 182] for computational and statistical properties of correlation integral with different choices of order (q) and the length scale (ϵ). For visual clarity, both of these indices are omitted from the following equations.

Now, the amount of information about X_{t+1} contained in both $\mathbf{X}_t$ and $\mathbf{Y}_t$ will be:

$$I(\mathbf{X}_t, \mathbf{Y}_t; X_{t+1}) = \log C(\mathbf{X}_t, \mathbf{Y}_t, X_{t+1}) - \log C(\mathbf{X}_t, \mathbf{Y}_t) - \log C(X_{t+1}) \tag{121}$$

whereas the amount of information of X_{t+1} contained in $\mathbf{X}_t$ is

$$I(\mathbf{X}_t; X_{t+1}) = \log C(\mathbf{X}_t, X_{t+1}) - \log C(\mathbf{X}_t) - \log C(X_{t+1}). \tag{122}$$

Given the past values of X at any specific time instant t, if past values of Y does not contain any information about the future values of X, then $I(\mathbf{X}_t, \mathbf{Y}_t; X_{t+1}) = I(\mathbf{X}_t; X_{t+1})$, otherwise when the past values of Y do contain information about the future, the following inequality $I(\mathbf{X}_t, \mathbf{Y}_t; X_{t+1}) > I(\mathbf{X}_t; X_{t+1})$ is expected.

Accordingly, 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 Eq. (121) − Eq. (122) which provides the information theoretic measure of Granger causality:

$$I^{GC}_{Y\to X} = I(\mathbf{X}_t, \mathbf{Y}_t; X_{t+1}) - I(\mathbf{X}_t; X_{t+1}) \tag{123}$$

$$= \log C(\mathbf{X}_t, \mathbf{Y}_t, X_{t+1}) - \log C(\mathbf{X}_t, X_{t+1}) - \log C(\mathbf{X}_t, \mathbf{Y}_t) + \log C(\mathbf{X}_t).$$

In order to obtain the statistical significance, bootstrapping procedure is recommended to check if the statistic is significantly larger than zero [62, 115].

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 [86, 206]. Interestingly, these entropies can be expressed in terms of generalized correlation integrals whose nonparametric estimation is well known. Correlation integrals are routinely employed in nonlinear time series analysis [1]. Additionally, correlation integral based entropies require minimal assumptions about the underlying dynamics of the systems and the nature of their coupling, thus the applications of these entropies are no longer restricted to deterministic systems but are suitable for any arbitrary, stationary and weakly mixing systems [182].

Correlation integral based nonparametric Granger causality test was originally proposed by Baek and Brock [14] and then later modified by Hiemstra and Jones [110] in the field of econometrics. They proposed the test statistic as

$$T^{GC}_{Y\to X} = \frac{C(\mathbf{X}_t, \mathbf{Y}_t; X_{t+1})}{C(\mathbf{X}_t, \mathbf{Y}_t)} - \frac{C(\mathbf{X}_t, X_{t+1})}{C(\mathbf{X}_t)} \tag{124}$$

The null hypothesis - Y is not Granger causing X - could be rejected if T^{GC} is too large because higher values of T^{GC} are expected when past observations of Y contain information about future observations of X. T^{GC} has some initial bias since it depends on the length N of the series [14] but it was shown [110] that asymptotically the statistic under the null hypothesis is normally distributed. However, one has to be careful in accepting the null hypothesis by using this statistic for real data applications [65].

Both statistic, I^{GC} and T^{GC}, are closely related (compare Eq. (124) to Eq. (123)), however, there is no one-to-one mapping between the outcomes of two statistic. Additionally, the statistical significance of the two statistic are measured in different ways, one by using Monte Carlo bootstrapping and the other by asymptotical distribution theory.

It is worth mentioning that both statistics neither provide any specific information about the nature (linear or nonlinear) nor the sign (positive or negative influence) of the causality. For I^{GC} statistic, one could calculate the linearized version on the basis of redundancies as proposed by Paluš [161] (see also [45]) which reflects only the dependence contained in the linear correlation matrix (of the variables which are assumed to be Gaussian). If the general statistic I^{GC} passes the test of significance, then its linearized counterpart is subject to test; if the linear version indicates a significant causality, then the causal influence is most probably due to a linear Gaussian process, otherwise a nonlinear type of causal influence can be inferred. For T^{GC} statistic, one could use the residuals from the linear autoregressive model fit to the two time series and any remaining incremental predictive power of one residual series for another can be considered nonlinear [14].

The T^{GC} statistic was applied primarily in the field of econometrics and finance, for example, an unidirectional information flow from relative money supply to exchange rate in European Monetary System [142], bidirectional causality between daily stock returns and trading volume in Korean market [214] where as volume Granger causes returns for Standard and Poor's index [65] (but see also their result of weakening this influence on the basis of a modified test statistic), bi-directional causality between volume and volatility in the New York Stock Exchange [41], to name a few.

On the other hand, I_{GC} statistic is relatively new but has been applied to wide classes of systems, from climatological [63], cardiological [115], to neurophysiological ones [45].

Other alternative nonparametric tests, such as non-causality test based on additive models proposed by Bell et al. [23], or test for conditional independence based on Hellinger distances [223], also exist, however, their applications have been quite limited, so far.

7 Conclusion

Natural phenomena usually emerge from behaviour or evolution of so-called complex systems which consist of many parts or subsystems. These subsystems interact in a complex, non-linear way and the behaviour of the whole system cannot be explained by a linear summation of dynamics of the system parts. In real-life situations, we do not have a direct access to the equations governing the underlying dynamics; instead, we are faced with a data set representing the temporal dynamics of possibly interacting variables recorded from possibly interacting subsystems. How can we tell from these observed sequences whether there exists any causal relationship between two ore more variables?

The basic rationale of this paper was that information theory provides a crucial key to this answer, 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. The overall objective of this paper was to present the current state of the art of these estimators of information theoretical measures which could be applied to detect causality. To the best of our knowledge, there is no other review paper available in the literature which deals with causality and its estimation from this point of view. We classified and discussed two types of estimators: parametric and non-parametric estimators.

Theoretically, for a good entropy estimator, the condition of consistency seems to be important. We specifically highlighted those estimators whose consistency results are known or could be derived. 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. Despite these precautionary remarks, we would like to stress again that there are enough good reasons, contrary to B. Russel's arguments [198], to investigate causality, offering numerous applications in natural and physical sciences.

Acknowledgements

K. H.-S. was supported by grant of Austrian Research Fonds FWF-H-226 (2005) under Charlotte Bühler Program and partly by ASCR 1ET 100 750 401, Project Baddyr. M.P. and M.V. were supported by the EC FP6 project BRACCIA (Contract No 517133 NEST), and partly by the Institutional Research Plan AV0Z10300504. J. B. was supported by JST.ERATO Shimojo project.

References

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

[2] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing (Dover, New Yoprk, 1972) pp. 946-949.

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

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

[5] S. Akaho, Y. Kiuchi and S. Umeyama, MICA: Multimodal independent component analysis, Proceedings of IJCNN (1999) 927–932.

[6] S. Amari, A. Cichocki and H.H. Yang, A new learning algorithm for blind signal separation, Advances in Neural Information Processing Systems, vol. 8, eds D.S. Touretzky, M.C. Mozer and M.E. Hasselmo (MIT Press, 1996) vol 8, 757–763.

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

[8] N. Ancona and S. Stramaglia, An invariance property of predictors in kernel-induced hypothesis spaces, Neural Computation 18 (2006) 749-759.

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

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

[11] B.C. Arnold, Pareto Distributions (International Co-Operative Publishing House, Burtonsvile, MD, 1985).

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

[13] J.H. Badsberg, An Environment for Graphical Models, Part I, PhD Thesis, (http://www.math.aau.dk/ jhb/Thesis, 1995).

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

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

[16] O.E. Barndorr-Nielsen and D.R. Cox, Inference and Asymprotics (Chapman and Hall, London, 1989).

[17] W.A. Barnett, A.P. Kirman and M. Salmon, Nonlinear Dynamics and Economics (Cambridge University Press, Cambridge UK, 1996).

[18] G.P. Basharin, On a statistical estimate for the entropy of the sequence of idependent random variables, Theory Prob. App. 4 (1959) 333-338.

[19] C. Beck and F. Schlogl, Thermodynamics of Chaotic Systems, (Cambridge University Press, Cambridge, 1993).

[20] J. Beirlant and M.C.A. van Zuijlen, The empirical distribution function and strong laws for functions of order statistics of uniform spacings, J. Multivar. Anal. 16 (1985) 300-317.

[21] J. Beirlant, Limit theory for spacing statistics from general univariate distributions, Pub. Inst. Stat. Univ. Paris XXXI Fasc. 1 (1986) 27-57.

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

[23] D. Bell, J. Kay and J. Malley, A non-parametric approach to non-linear causality testing, Econ Lett 51 (1996) 7–18.

[24] J.L. Bentley, Multidimensional Divide-and-Concquer, Communications of the ACM, 23 (1980) 214.

[25] J.L. Bentley, K-d trees for semi-dynamic point sets, in: Sixth Annual ACM Symposium on Computational Geometry (San Francisco, 1990) p. 91.

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

[27] J.F. Bercher and C. Vignat, Estimating the entropy of a signal with applications, IEEE Transaction in Signal Processing 48 (2000) 1687-1694.

[28] J. Bhattacharya and H. Petsche, Drawing on mind's canvas: Differences in cortical integration patterns between artists and non-artists. Human Brain Mapping 26 (2005) 1-14.

[29] J. Bhattacharya, E. Pereda and H. Petsche, Efficient detection of coupling between bivariate experimental signals in the state space, IEEE Transaction of Systems Man and Cybernetics - Part B 33 (2003) 85-95.

[30] J. Bhattacharya, H. Petsche and E. Pereda, Long-range synchrony in the gamma band: role in the music perception, Journal of Neuroscience 15 (2001) 6329-6337.

[31] W. Bialek, F. Rieke, R. de Ruyter van Steveninck and D. Warland, Reading a neural code, Science 252 (1991) 1854-1857.

[32] P. Bickel and L. Breiman, Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test, Annals of Statistics 11 (1983) 185-214.

[33] S. Bingham and M. Kot, Multidimensional trees, range searching, and a correlation dimension algorithm of reduced complexity, Phys Lett. A 140 (1989) 327-330.

[34] C.M. Bishop, Neural Networks for Pattern Recognition (Oxford University Press, New York, 1995).

[35] S. Blinnikov and R. Moessner, Expansions for nearly Gaussian distributions, Astronomy and Astrophysics Supplement Series, 130 (1998) 193-205.

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

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

[38] M. Bračič Lotrič and A. Stefanovska, Synchronization and modulation in the human cardiorespiratory system. Physica A 283 (2000) 451-461.

[39] J. Brea, D.F. Russell and A.B. Neiman AB Measuring direction in the coupling of biological oscillators: A case study for electroreceptors of paddlefish, Chaos 16 (2006) 026111.

[40] D.R. Brillinger, Some data analyses using mutual information, Brasilian J. of Probability and Statistics 18 (2004) 163-183.

[41] C. Brooks, Predicting stock index volatility: Can market volume help? J Forecasting 17 (1998) 59–80.

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

[43] A.J. Butte and I.S. Kohane, Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements, Pac Symp Biocomput. (2000) 418-29.

[44] C.J. Cellucci, A.M. Albano and P.E. Rapp, Statistical validation of mutual information calculations: Comparison of alternative numerical algorithm, Physical Review E 71 (2005) 066208.

[45] M. Chávez, J. Martinerie and M. Le Van Quyen, Statistical assessment of nonlinear causality: Application to epileptic EEG signals, Journal of Neuroscience Methods 124 (2003) 113–128.

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

[47] P. Comon, Independent component analysis - a new concept? Signal Processing 36 (1994) 287-314.

[48] I.P. Cornfeld, S.V. Fomin and Y.G. Sinai, Ergodic Theory (Springer, New York, 1982).

[49] J. Costa and A. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, IEEE Transaction on Signal Processing, 25 (2004) 2210-2221.

[50] T. Cover and J. Thomas, Elements of Information Theory (John Wiley and Sons, New York, NY, 1991), Chapter 9.

[51] T. Cover and J. Thomas, Elements of Information Theory (John Wiley and Sons, New York, NY, 1991), Chapter 5.

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

[53] N. Cressie, On the logarithm of higher order spacings, Biometrika 63 (1976) 343-355.

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

[55] G. Darbellay, An estimator of the mutual information based on a criterion of independence, Computational Statistics & Data Analysis 32 (1999) 1-17.

[56] C. Daub, R. Steuer, J. Selbig and S. Kloska, Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data, BMC Bioinformatics 5:118 (2004) 1-12.

[57] M. De Berg, O. Schwarzkopf, M. van Kreveld and M. Overmars, Computational Geometry: Algorithms and Applications (Springer, Berlin, 2000).

[58] C. De Boor, A practical guide to splines (Springer, New York, 1978).

[59] G. Deco and D. Obradovic, An Information-Theoretic Approach to Neural Computing (Springer, New York, 1996).

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

[61] L. Devroye, Lecture Notes in Bucket Algorithms, Progress in Computer Science no. 6 (Birkhäuser, Boston, 1986).

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

[63] C. Diks and M. Mudelsee, Redundancies in the Earth's climatological time series, Phys. Letters A 275 (2000) 407–414.

[64] C. Diks and V. Panchenko, A Note on the Hiemstra-Jones Test for Granger Non-causality, Studies in Nonlinear Dynamics and Econometrics, 9 (2005), 4/1-7.

[65] C. Diks and V. Panchenko, A new statistic and practical guidelines for nonparametric Granger causality testing, Journal of Economic Dynamics and Control 30 (2006) 1647-1669.

[66] Y.G. Dmitriev and F.P. Tarasenko, On the estimation functions of the probability density and its derivatives, Theory Probab. Appl. 18 (1973) 628-633.

[67] R.L. Dobrushin, A simplified method of experimentally evaluating the entropy of a stationary sequence. Teoriya Veroyatnostei i ee Primeneniya, 3 (1958) 462-464.

[68] E.J. Dudewitz, E.C. Van der Meulen, Entropy-based test of uniformity. J. Amer. Statist. Assoc. 76 (1981) 967-974.

[69] B. Efron and C. Stein, The jackknife estimate of variance, Annals of Statistics 9 (1981) 586-596.

[70] D. Erdogmus, Information theoretic learning: Renyi's Entropy and its Application to Adaptive System Training, PhD thesis (University of Florida, 2002).

[71] D. Erdogmus and J. Principe, An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems, IEEE Trans. Signal processing 50 (2002) 1780-1786.

[72] D. Erdogmus, K.E. Hild and J.C. Principe, Adaptive blind deconvolution of linear channels using Renyi's entropy with Parzen window estimation. IEEE Trans. On Signal Processing 52 (2004) 1489-1498.

[73] U. Feldmann and J. Bhattacharya, Predictability improvement as an asymmetrical measure of interdependence in bivariate time series, International Journal of Bifurcation and Chaos 14 (2004) 505-514.

[74] M. Fernandes, Nonparametric entropy-based tests of independence between stochastic processes, PhD thesis (Université Libre de Bruxelles, 2000).

[75] R.A. Fisher, On an absolute criterion for fitting frequency curves, Messenger of Mathematics, 41 (1912) 155-160.

[76] R.A. Fisher, On the mathematical foundations of theoretical statistics, Philos. Trans. Roy. Soc. London Ser. A 222 (1922) 309-368.

[77] J. Fisher and J. Principe, Entropy manipulation of arbitrary nonlinear mappings, Proc. IEEE Workshop Neural Networks for Signal Processings (Amelia Island, 1997) 14-23.

[78] A. Fraser and H. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 11341140.

[79] A. Fraser, Information and entropy in strange attractors. IEEE Trans. Information Theory 35 (1989) 245–262.

[80] J.H. Friedman, F. Baskett and L.J. Shustek, An Algorithm for Finding Nearest Neighbor, IEEE Transactions on Computers 10 (1975) 1000-1006.

[81] J. Friedman, Exploratory projection pursuit, J. American Statistical Association, 82 (1987) 249-266.

[82] T. Gautama and M.M. Van Hulle, Surrogate-based test for Granger causality, in: Proceedings IEEE Neural Network for Signal Processing Workshop, Toulouse, France (2003) 799-808.

[83] E. Gaztanga, P. Fosalba and E. Elizande, Gravitational evolution of the large-scale probability density distribution: the Edgeworth and Gamma expansions, Astrophysical Journal 539, Part 1 (2000) 522-531.

[84] A. German, J.B. Carlin, H.S. Stern and D.B. Rubin, Bayesian Data Analysis (Chapman and Hall/A CRC Press Company, Texts in Statistical Science Series, 2004).

[85] J. Geweke, Inference and causality in economic time series models, in: Handbook of Econometrics, eds Z. Griliches and M.D. Intriligator (North-Holland, 1984) vol. 2, 1101–1144.

[86] I.M. Gelfand and A.M. Yaglom, Calculation of the amount of information about a random function contained in another such function, Am. Math. Soc. Translations 12 (1959) 199–236.

[87] M.N. Goria, N.N. Leonenko, V.V. Mergel and P.L. Novi Inverardi, A new class of random vector entropy estimators and its applications in testing statistical hypotheses, Nonparametric Statistics 17 (2005) 277-297.

[88] C.W.J. Granger, Investigating causal relations by econometric and cross-spectral methods, Econometrica 37 (1969) 424-438.

[89] C.W.J. Granger and J-L. Lin, Using the mutual information coefficient to identify lags in nonlinear models, J. Time Series Anal. 15 (1994) 371-384.

[90] C.W.J. Granger and P. Newbold, Forecasting Economic Time Series (Academic Press, New York, 1977).

[91] C.W.J. Granger, Testing for causality: A personal viewpoint, Journal of Economic Dynamics and Control 2 (1980) 329-352.

[92] C.W.J. Granger, Time series analysis, cointegration, and applications. Nobel Lecture, December 8, 2003. In: Les Prix Nobel. The Nobel Prizes 2003, ed. Tore Frängsmyr, [Nobel Foundation] (Stockholm, 2004) pp. 360–366. http://nobelprize.org/nobel_prizes/economics/laureates/2003/granger-lecture.pdf

[93] P. Grassberger, An optimized box-assisted algorithm for fractal dimensions, Phys. Lett. A 148 (1990) 63-68.

[94] P. Grassberger and I. Procaccia, Measuring of strangeness of strange attractors, Physica D (1983) 189-208.

[95] P. Grassberger, Finite sample corrections to entropy and dimension estimates, Phys. Lett. A 128 (1988) 369-373.

[96] P. Grassberger, Entropy Estimates from Insufficient Samplings, Arxiv preprint physics/0307138, (2003) - arxiv.org

[97] M.L. Green and R. Savit, Dependent variables in broad band continuous time series, Physica D 50 (1991) 521-544.

[98] L. Györfi and E.C. Van der Meulen EC. Density-free convergence properties of various estimators of entropy, Comput. Statist. Data Anal. 5 (1987) 425-436.

[99] L. Györfi and E.C. Van der Meulen, An entropy estimate based on a kernel density estimation, in: Limit Theorems in Probability and Statistics, eds I. Berkes, E. Csaki and P. Revesz (North Holland, 1989) pp. 229-240.

[100] L. Györfi and E.C. Van der Meulen, On nonparametric estimation of entropy functionals, in: Nonparametric Functional Estimation and Related Topics, ed G. Roussas (Kluwer Academic Publisher, Amsterdam, 1990) pp. 81-95.

[101] P. Hall, Limit theorems for sums of general functions of m-spacings, Math. Proc. Camb. Phil. Soc. 96 (1984) 517-532.

[102] P. Hall, On powerful distributional tests based on sample spacings, J. Multivariate Statist. 19 (1986) 201-225.

[103] P. Hall, On Kullback-Leibler loss and density estimation. Annals of Statistics 15 (1987) 1491-1519.

[104] P. Hall, The Bootstrap and Edgeworth Expansion (Springer-Verlag, New York, 1992).

[105] P. Hall and S.C. Morton, On the estimation of entropy, Ann. Inst. Statist. Math. 45 (1993) 69-88.

[106] T.C.Halsey, M.H. Jensen, L.P. Kadanoff, I. Procaccia and B. Shraimann, Fractal measures and their singularities: the characterization of strange sets, Phys. Rev. A 33 (1986) 1141.

[107] S. Haykin, Neural Networks: A Comprehensive Foundation, Second Edition (Prentice Hall, Englewood Cliffs, NJ, 1998).

[108] H.G.E. Hentschel and I. Procaccia, The infinite number of dimensions of probabilistic fractals and strange attractors, Physica D 8 (1983) 435.

[109] H. Herzel, Complexity of symbol sequences, Syst. Anal. Model. Simul. 5 (1988) 435-444.

[110] C. Hiemstra and J.D. Jones, Testing for linear and nonlinear Granger causality in the stock pricevolume relation, Journal of Finance 49 (1994) 1639-1664.

[111] H. Hinrichs, H.J. Heinze, M.A. Schoenfeld, Causal visual interactions as revealed by an information theoretic measure and fMRI, NeuroImage 31 (2006) 1051–1060.

[112] G. Hinton and T. Sejnowski, Learning and relearning in Boltzmann machines, in: Parallel Distributed procesing, eds. D. Rumelhart and J. McClelland (MIT Press, Cambridge, 1986) Vol 1, Chapter 7, pp. 282-317.

[113] K. Hlaváčková -Schindler, Some comparisons of the complexity of neural network models and linear methods by means of metric entropy, in: Multiple Participant Decision Making, eds. Andrýsek, Kárný, Kracík, Int. Series on Advanced Intelligence, 9 (2004) 149-160.

[114] M.-Ch. Ho, Y.-Ch. Hung and I-M. Jiang, Phase synchronization in inhomogeneous globally coupled map lattices, Phys. Lett. A 324 (2004) 450-457.

[115] B.P.T. Hoekstra, C.G.H. Diks, M.A. Allessie and J. DeGoede, Non-linear time series analysis: Methods and applications to atrial fibrillation, Ann. Ist. Super. Sanita 37 (2003) 325-333.

[116] H. Hua-Yang and S. Amari, Adaptive on-line learning algorithms for blind separation - Maximum entropy and minimum mutual information, Neural Computation 9 (1997) 1457-1482.

[117] P. Huber, Projection Pursuit, Annals of Statistics, 13 (1985) 435-475.

[118] S. Ihara, Information Theory for Continuous Systems (World Scientific, Singapore, 1993).

[119] A.V. Ivanov and A. Rozhkova, Properties of the statistical estimate of the entropy of a random vector with a probability density, Problems of Information Transmission 10 (1981) 171-178.

[120] J. Jamšek, A. Stefanovska and P.V.E. McClintock, Nonlinear cardio-respiratory interactions revealed by time-phase bispectral analysis. Phys. Med. Biol. 49 (2004) 4407–4425.

[121] H. Joe, On the estimation of entropy and other functionals of a multivariable density., Ann. Inst. Stats. Math 41 (1989) 638-697.

[122] M.C. Jones and R. Sibson, What is projection pursuit? J. Royal Statist. Soc. London, Ser. A. 150 (1987) 1-36.

[123] A.M. Kagan, Y.V. Linnik and C.R. Rao, Characterization Problems in Mathematical Statistics (Wiley, New York, 1973).

[124] A. Kaiser and T. Schreiber, Information transfer in continuous processes, Physica D 166 (2002) 43–62.

[125] S. Kalitzin, B.W. van Dijk, H. Spekreijse and W.A. van Leeuwen, Coherency and connectitvity in oscillating neural networks: linear patialization analysis, Biological Cybernetics 76(1) (1997) 73-83.

[126] M. Kamiński, M. Ding, W.A. Truccolo and S.L. Bressler, Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance, Biol. Cybern. 85 (2001) 145–157.

[127] M. Kaminski, M. Ding, W.A. Truccolo and S.L. Bressler, Evaluating causal relations in neural systems: Granger causality, directed transfer function and statisticall assessment of significance. Biological Cybernetics 85 (2001) 145-57.

[128] M. Kaminski and H. Liang, Causal influence: Advances in neurosignal analysis, Crit. Rev. Biomed. Eng. 33 (2005) 347-430.

[129] J.N. Kapur, Measures of Information and Their Applications (John Wiley Eastern, New Delhi, 1994).

[130] T. Katura, N. Tanaka, A. Obata, H. Sato and A. Maki, Quantitative evaluation of interrelations between spontaneous low-frequency oscillations in cerebral hemodynamics and systemic cardiovascular dynamics, Neuroimage 31 (2006) 1592-1600.

[131] M.B. Kennel, J. Shlens and H.D.I. Abarbanel, Estimating entropy rates with Bayesian confidence intervals, Neural Computation, 17 (2005) 1531-1576.

[132] K.H. Knuth, Informed source separation: A Bayesian tutorial, Proceedings of the 13th European Signal Processing Conference (EUSIPCO 2005) ed. E. Kuruoglu (Antalya, Turkey, 2005).

[133] K.H. Knuth, A. Gotera, C.T. Curry, K.A. Huyser, K.R. Wheeler and W.B. Rossow, Revealing relationships among relevant climate variables with information theory, Earth-Sun System Technology Conference 2005, Adelphi, MD, 2005.

[134] A.N. Kolmogorov, Entropy per unit time as a metric invariant of automorphism, Dokl. Akad. Nauk SSSR 124 (1959) 754–755.

[135] L.F. Kozachenko, N.N. Leonenko, Sample estimate of the entropy of a random vector, Problems of Information Transmission 23 (1987) 95-100.

[136] A. Kraskov, H. Stögbauer and P. Grassberger, Estimation mutual information, Physical Review E69 (2004) 066138.

[137] S. Kullback and R.A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79-86.

[138] N. Kwak and C. Choi, Input feature selection by mutual information based on Parzen window, IEEE Trans. On Pattern Analysis and Machine Intellingence 24 (2002) 1667-1671.

[139] N. Leonenko, L. Pronzato and V. Savani, 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.

[140] W. Li, Mutual information functions versus correlation functions, J. Statistical Physics 60 (1990) 823-837.

[141] J.J. Lin, N. Saito and R.A. Levine, Edgeworth approximations of the Kullbach-Leibler distance towards problems in image analysis, 2001, http://www.math.ucdavis.edu/˜saito/publications/saito_ekld2.pdf

[142] Y. Ma and A. Kanas, Testing for nonlinear Granger causality from fundamentals to exchange rates in the ERM, J. Int. Finance Markets, Institutions and Money 10 (2000) 69–82.

[143] D. Marinazzo, M. Pellicoro and S. Stramaglia, Nonlinear parametric model for Granger causality of time series, Physical Review E 73 (2006) 066216.

[144] R. Marschinski and H. Kantz, Analysing the information flow between financial time series - An improved estimator for transfer entropy, European Physical Journal B 30 (2002) 275-281.

[145] G. Miller and W. Madow, On the maximum likelihood estimate of the Shannon-Wiener measure of information, Air Force Cambridge Research Center Technical Report 75 (1954) 54-75.

[146] G. Miller, Note on the bias of information estimates, in: Information theory in psychology II-B (pp.95-100), ed. H. Quastler (Glencoe, IL, Free Press, 1955).

[147] E.G. Miller, A new class of entropy estimators for multi-dimensional densities, International Conference on Acoustics, Speech and Signal Processing, 2003.

[148] R. Moddemeijer, On estimation of entropy and mutual information of continuous distribution, Signal Processing 16 (1989) 233-246.

[149] I.I. Mokhov and D.A. Smirnov, El Nino-Southern Oscillation drives North Atlantic Oscillation as revealed with nonlinear techniques from climatic indices, Geophysical Research Letters 33 (2006) L03708.

[150] A. Mokkadem, Estimation of the entropy and information for absolutely continuous random variables, IEEE Transaction on Information Theory 35 (1989) 195-196.

[151] Y. Moon, B. Rajagopalan and U. Lall, Estimation of mutual information using kernel density estimators, Physical Review E 52 (1995) 2318-2321.

[152] R. Neal and G. Hinton, A view of the EM algorithm that justifies incremental, sparse, and other variants, in: Learning in Graphical Models, ed. Michael I. Jordan (MIT Press, Cambridge, MA, 1999) pp. 355-368.

[153] I. Nemenman, F. Shafee and W. Bialek, Entropy and inference, revisited, in: Advances in Neural Information Processing Systems 14, eds. T.G. Dietterich, S. Becker and Z. Ghahramani (MIT press, Cambridge, MA, 2002).

[154] I. Nemenman, W. Bialek W and R. de Ruyter van Stevenick, Entropy and information in neural spike trains: progress on sampling problem, Physical Review E69 (2004) 056111.

[155] I. Nemenman and W. Bialek, Occam factors and model-independent Bayesian learning of continuous dsitributions, Phys. Rev. E 65 (2002) 026137

[156] K. Otsuka, Y. Miyasaka and T. Kubota, Formation of an information network in a self-pulsating multimode laser, Phys. Rev. E 69 (2004) 046201.

[157] R. Ottes, A critique of Suppe's theory of probabilistic causality, Synthese 48 (1981) 167-189.

[158] M. Paluš, A. Stefanovska, Direction of coupling from phases of interacting oscillators: An information-theoretic approach, Phys. Rev. E 67 (2003) 055201(R).

[159] M. Paluš, V. Komárek, T. Procházka, Z. Hrnčíř and K. Štěrbová, Synchronization and information flow in EEG of epileptic patients, IEEE Engineering in Medicine and Biology Magazine 20 (2001) 65-71.

[160] M. Paluš, V. Komárek, Z. Hrnčíř and K. Štěrbová, Synchronization as adjustment of information rates: Detection from bivariate time series. Phys. Rev. E 63 (2001) 046211.

[161] M. Paluš, Testing for nonlinearity using redudancies: Quantitative and qualitative aspects, Physica D 80 (1995) 186-205.

[162] M. Paluš, Coarse-grained entropy rates for characterization of complex time series, Physica D 93 (1996) 64-77.

[163] M. Paluš, Kolmogorov entropy from time series using information-theoretic functional, Neural Network World Vol. 7 No. 3 (1997) 269-292, http://www.uivt.cas.cz/~mp/papers/rd1a.ps.

[164] M. Paluš, Identifying and quantifying chaos by using information-theoretic functionals, in: Time series prediction: Forecasting the future and understanding the past, eds A.S. Weigend, N.A. Gershenfeld (Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol XV., Addison-Wesley, Reading, MA,1993) pp. 387-413.

[165] M. Paluš, Detecting nonlinearity in multivariate time series, Phys. Lett. A 213 (1996) 138–147.

[166] M. Paluš and D. Hoyer, Detecting nonlinearity and phase synchronization with surrogate data, IEEE Engineering in Medicine and Biology 17 (1998) 40–45.

[167] L. Paninski, Estimation of entropy and mutual information, Neural Computation, 15 (2003) 1191-1253.

[168] L. Paninski, Estimating entropy on m bins given fewer than m samples, IEEE Transaction on Information Theory 50 (2004) 2200-2203.

[169] L. Paninski, Asymptotic theory of information-theoretic experimental design, Neural Computation 17 (2005) 1480-1507.

[170] E. Parzen, On estimation of a probability density function and mode, in: Time Series Analysis Papers (Holden-Day, Inc., San Diego, California, 1967).

[171] J. Pearl, Causality: Models, Reasoning and Inference (Cambridge University Press, New York, 2000).

[172] M. Penrose, Central limit theorems for k-nearest neighbor distances, Stochastic Processes and their Applications 85 (2000) 295-320.

[173] K. Petersen, Ergodic Theory (Cambridge University Press, Cambridge, 1983).

[174] M.A. Peters and P.A. Iglesias, Minimum entropy control for discrete–time varying systems, Automatica 33 (1997) 591-605.

[175] M.E. Pflieger, Time-lagged causual information: A new metric for effective connectivity analysis, Neuroimage 19 (2003), a poster on CD ROM (HBM 2003).

[176] M.E. Pflieger, R.E. Greenblatt, Using conditional mututal information to approximate causality for mutlivariate physiological time series. Int. J. Bioelectromagnetism 7 (2005) 285-288.

[177] A. Pikovsky, M. Rosenblum and J. Kurths, Synchronization. A Universal Concept in Nonlinear Sciences (Cambridge University Press, Cambridge, 2001).

[178] M. S. Pinsker, Information and Information Stability of Random Processes (Holden Day, San Francisco, 1964).

[179] B. Pompe, Measuring statistical dependencies in a time series, J. Stat. Phys. 73 (1993) 587-610.

[180] T. Poschel, W. Ebeling, and H. Rose, Dynamic entropies, long-range correlations and fluctuations in complex linear structures, J. Stat. Phys. 80 (1995) 1443-1452.

[181] B.L.S. Prasaka Rao, Nonparametric Functional Estimation (Academic Press, New York, 1983).

[182] D. Prichard and J. Theiler, Generalized redundancies for time series analysis, Physica D 84 (1995) 476–493.

[183] J.C. Principe and D. Xu, Information-theoretic learning using Rényi quadratic entropy, in: Proc. 1st Int. Workshop on Independent Component Analysis (ICA'99), Aussois, France (1999) 395-400.

[184] J.C. Principe, J.W. Fischer and D. Xu, Information theoretic learning, in: Unsupervised Adaptive Filtering, ed. S. Haykin (John Wiley and Sons, New York, 2000) pp. 265-319.

[185] J.C. Principe, D. Xu, Q. Zhao and J.W. Fischer, Learning from examples with information theoretic criteria, Journal of VLSI Signal Processing Systems, Special Issue on Neural Networks (2000) 61-77.

[186] L. Pronzato, H.P. Wynn and A.A. Zhigljavsky, Using Renyi entropies to measure: Uncertainty in search problems, Lect. Appl. Math. AMS 33 (1997) 253-268.

[187] M. Le Van Quyen, J. Martinerie, C. Adam, and F.J. Varela, Nonlinear analyses of interictal EEG map the brain interdependences in human focal epilepsy, Physica D 127 (1999) 250-266.

[188] R. Quian Quiroga, J. Arnhold and P. Grassberger, Learning driver-response relationships from synchronization patterns, Phys. Rev. E 61(5) (2000) 5142–5148.

[189] B. Rajagopalan and U. Lall, Seasonality of precipitation along meridian in the western United States, Geophysical Research Letters 22 (1997) 1081-1084.

[190] N. Reid, Asymptotic Expansions, In: Encyclopedia for Statistical Sciences, Update Volume, eds S. Kotz, C.B. Read and D.L. Banks (Wiley, New York, 1996), pp. 32-39.

[191] A. Rényi, 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.

[192] A. Rényi, Some fundamental questions of information theory, Selected Papers of Alfred Rényi, vol.2, Akademia Kiado, Budapest (1976) 526-552.

[193] G. Rigoll, Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition systems, IEEE Transactions on Speech and Audio Processing, Special Issue on Neural Networks for Speech 2 (1994) 175-184.

[194] J. Rissanen, T.P. Speed, B. Yu, Density estimation by stochastic complexity, IEEE Transactions on Information Theory 38 (1992) 315-323.

[195] F. Rossi, A. Lendasse, D. Francois, V. Wertz and M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modeling, in: Chemometrics and Intelligent Laboratory Systems, Lille, France, November 30-December 1, (2005) 52-55.

[196] M. Roulston, Estimating the errors on measured entropy and mutual information, Physica D 125 (1999) 285-294.

[197] D.E. Rumelhart, G.E. Hinton and J.R. Wiliams, Learning representations by back-propagating errors, Nature (London) 323 (1986) 533-536.

[198] B. Russel, On the notion of cause, Proceedings of the Aristotelian Society, New Series 13 (1913) 1-26.

[199] C. Schäfer, M.G. Rosenblum, J. Kurths and H.H. Abel, Heartbeat synchronized with ventilation, Nature 392 (1998) 239-240.

[200] C. Schäfer, M.G. Rosenblum, J. Kurths and H.H. Abel, Synchronization in the human cardiorespiratory system, Phys. Rev. E 60 (1999) 857-870.

[201] S.J. Schiff, P. So, T. Chang, R.E. Burke and T. Sauer, Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble, Phys. Rev. E 54 (1996) 6708-6724.

[202] A.O. Schmitt, H. Herzel, and W. Ebeling, A new method to calculate higher-order entropies from finite samples, Europhys. Lett. 23 (1993) 303-309.

[203] A. Schmitz, Measuring statistical dependence and coupling of subsystems, Phys. Rev. E 62 (2000) 7508-7511.

[204] N. Schraudolph, Optimization of entropy with neural networks, PhD Thesis (University of California, San Diego, USA, 1995).

[205] N. Schraudolph, Gradient-based manipulation of non-parametric entropy estimates, IEEE Trans. On Neural Networks 14 (2004) 828-837.

[206] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. 85 (2000) 461-464.

[207] T. Schreiber, Efficient neighbor searching in nonlinear time series analysis, Int. Journal of Bifurc. and Chaos 5 (1995) 349-358.

[208] T. Schürmann and P. Grassberger, Entropy estimation of symbol sequences, Chaos 6 (1996) 414-427.

[209] T. Schürmann, Bias analysis in entropy estimation, J. Phys. A: Math. Gen. 37 (2004) L295-L301.

[210] C. Selltiz, L.S. Wrightsman, and S.W. Cook, Research Methods in Social Relations (Holt, Rinehart and Winston, New York, 1959).

[211] C.E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27 (1948) 379-423.

[212] R. Shaw, Strange attractors, chaotic behavior and information flow. Z. Naturforsch 36A (1981) 80-112.

[213] R.S. Shaw, The dripping faucet as a model chaotic system (Aerial, Santa Cruz, 1985).

[214] P. Silvapulle and J.S. Choi, Testing for linear and nonlinear Granger causality in the stock price-volume relation: Korean evidence, Quarterly Review of Economics and Finance 39 (1999) 59-76.

[215] B.W. Silverman, Density Estimation (Chapman and Hall, London, 1986).

[216] Y.G. Sinai, On the concept of entropy for a dynamic system, Dokl. Akad. Nauk SSSR 124 (1959) 768-771.

[217] Y.G. Sinai, Introduction to Ergodic Theory (Princeton University Press, Princeton, 1976).

[218] R.R. Sokal and F.J. Rohlf, Biometry (W.H.Freeman and Company, New York, 2003).

[219] A. Sorjamaa, J. Hao and A. Lendasse, Mutual Information and k-Nearest Neighbors Approximator for Time Series Prediction, Proceedings of ICANN 2005, LNCS 3697 (2005) 553-558.

[220] A. Stefanovska, H. Haken, P. V. E. McClintock, M. Hožič, F. Bajrović and S. Ribarič, Reversible transitions between synchronization states of the cardiorespiratory system. Phys. Rev. Lett. 85 (2000) 4831-4834.

[221] R. Steuer, J. Kurths, C.O. Daub, J. Weise and J. Selbig, The mutual information: detecting and evaluating dependencies between variables, Bioinformatics 18 (Suppl.2) (2002), S231-240.

[222] S.P. Strong, R. Koberle, R.R. de Ruyter van Steveninck and W. Bialek, Entropy and information in neural spike trains, Phys. Rev. Lett. 80 (1998) 197-202.

[223] H.J. Su and H. White, A nonparametric Hellinger metric test for conditional independence, Technical Report (2003) Department of Economics, University of California, San Diego.

[224] F. Takens, In: D.A. Rand, D.S. Young, eds., Dynamical Systems and Turbulence (Warwick 1980), Lecture Notes in Mathematics 898. Springer, Berlin, 1981, p.365.

[225] F. Takens, Invariants related to dimension and entropy, in: Atas do 13 Coloquio Brasileiro de Mathematica, Rio de Janeiro, Brasil, 1983.

[226] A. Taleb and C. Jutten, Entropy optimization - application to source separation, LNCS 1327 (1997) 529-534.

[227] F.P. Tarasenko, On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable, and the distribution-free entropy test of goodness-of-fit, in: Proc. IEEE 56 (1968) 2052-2053.

[228] P. Tass, M.G. Rosenblum, J. Weule, J. Kurths, A. Pikovsky, J. Volkmann, A. Schnitzler, and H.-J. Freund, Detection of n:m phase locking from noisy data: Application to Magnetoencephalography, Phys. Rev. Lett. 81 (1998) 3291-3294.

[229] T. Tersyrita, in: Handbook of Applied Economic Statistics, eds A. Ullah and D.E.A. Gilles (Marcel Dekker, New York, 1988).

[230] J. Theiler, Efficient algorithm for estimation the correlation dimension from a set of discrete points, Physical Review A 36 (1987) 4456-4462.

[231] J. Theiler, Estimating fractal dimension, J. Opt. Soc. Amer. A 7 (1990) 1055-1071.

[232] S. Theodoridis and K. Koutroumbas, Pattern Recognition (Academic Press, San Diego, 1999).

[233] K. Torkkola, Feature extraction by non-parametric mutual information maximization, Journal of Machine Learning Research 3 (2003) 1415-1438.

[234] T. Trappenberg, J. Ouyang and A. Back, Input variable selection: Mutual information and linear mixing measures, IEEE Trans. on Knowledge and Data Engineering 18 (2006) 37-46.

[235] A. Trever and S. Panzeri, The upward bias in measures of information derived from limited data samples, Neural Computation 7 (1995) 399-407.

[236] U. Triacca, Is Granger causality analysis appropriate to investigate the relationships between atmospheric concentration of carbon dioxide and global surface air temperature? Theor. Appl. Climatol. 81 (2005) 133-135.

[237] A.B. Tsybakov and E.C. van Meulen, Root-n consistent estimators of entropy for densities with unbounded support, Scand. J. Statist. 23 (1994) 75-83.

[238] http://mathworld.wolfram.com/UlamMap.html

[239] B. Van Es, Estimating functional related to a density by a class of statistics based on spacings, Scand. J. Statist. 19 (1992) 61-72.

[240] M.M. Van Hulle, Entropy-based kernel mixture modeling for topographic map formation, IEEE Trans. Neural Networks (Special Issue on Information Theoretic Learning), 15 (2004) 850-858.

[241] M.M. Van Hulle, Edgeworth approximation of multivariate differential entropy, Neural Computation 17 (2005) 1903-1910.

[242] M.M. Van Hulle, Multivariate Edgeworth-based entropy estimation, in: Proc. IEEE Workshop on Machine Learning for Signal Processing, Mystic, Connecticut, USA, September 28-30, 2005.

[243] O. Vašíček, A test for normality based on sample entropy, J. R. Stat. Soc Ser. B Methodol. 38 (1976) 54-59.

[244] M. Vejmelka and K. Hlaváčková-Schindler, Mutual Information Estimation in Higher Dimensions: A Speed-Up of a k-Nearest Neighbor Based Estimator, submitted to ICANNGA'07.

[245] P.F. Verdes, Assessing causality from multivariate time series, Physical Review E 72 (2005) 026222.

[246] A.C.G. Verdugo Lazo and P.N. Rathie, On the entropy of continuous probability distributions, IEEE Transactions on Information Theory 24 (1978) 120-122.

[247] J. Victor, Asymptotic bias in information estimates and the exponential (Bell) polynomials, Neural Computation 12 (2000) 2797-2804.

[248] J. Victor, Binless strategies for estimation of information from neural data, Physical Review E 66 (2002) 051903.

[249] J. Victor and K. Purpura, Metric-space analysis of spike trains: Theory, algorithms and application. Network 8 (1997) 127-164.

[250] P. Viola, Alignment by maximization of mutual information, PhD thesis, MIT, 1995.

[251] P. Viola, N. Schraudolph and T. Sejnowski, Empirical entropy manipulation for real-world problems, in: Advances in Neural Information Processing Systems (NIPS 8) (The MIT Press, Cambridge, MA, 1996) pp. 851857.

[252] N. Wiener, The theory of prediction, in: Modern Mathematics for Engineers, ed. E.F. Beckenbach (McGraw-Hill, New York, 1956).

[253] http://en.wikipedia.org/wiki/Causality

[254] L. Xu, C. Cheung, H. Yang and S. Amari, Maximum equalization by entropy maximization and mixture of cumulative distribution functions, in: Proc. of ICNN97, Houston (1997) 1821-1826.

[255] J. Xu, Z.R. Liu, R. Liu and Q.F. Yang, Information transmission in human cerebral cortex, Physica D 106 (1997) 363-374.

[256] H. Yang and S. Amari, Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information, Neural Computation 9 (1997) 1457-1489.

[257] J. Yang and P. Grigolini, On the time evolution of the entropic index, Physics Letters A 263 (1999) 323-330.

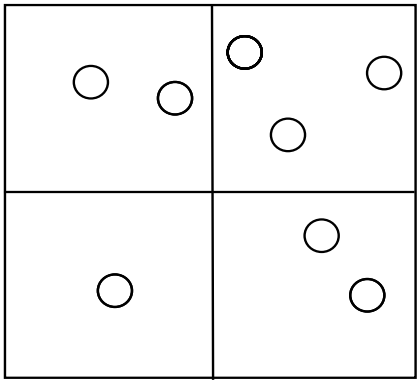

Figure 1: An example of equidistant binning (the method from Butte and Kohane) - the bins have the same size

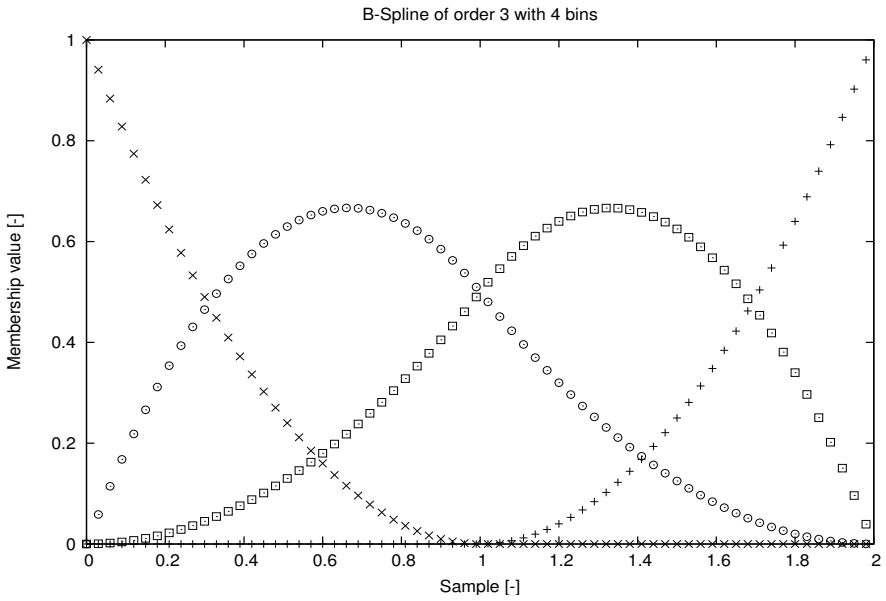

Figure 2: Generalized binning with B-splines from Daub et al. - an example of B-splines of order 3 for four bins.

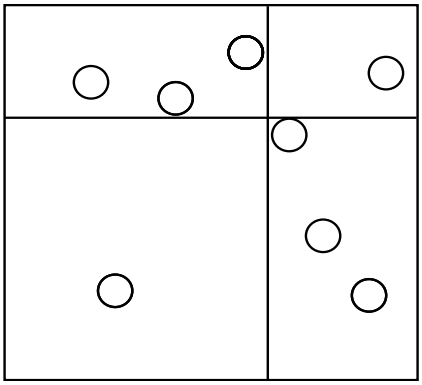

Figure 3: An example of marginally equiquantized binning method from Paluš - the bins can have different size but contain the same number of data points in their marginals

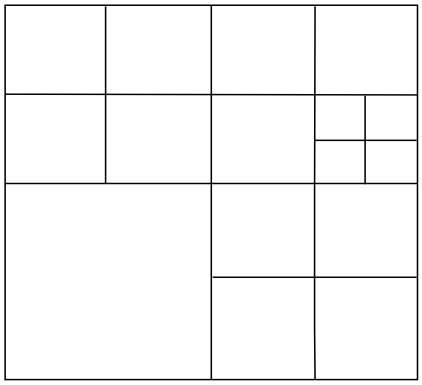

Figure 4: An example of a partition that can arise from the splitting procedure defined by the Darbellay & Vajda algorithm.

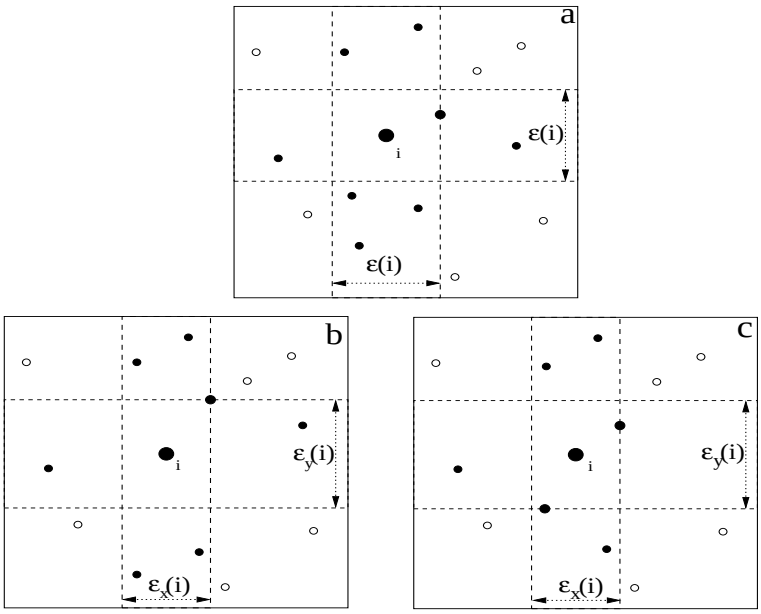

Figure 5: Illustration of the ϵ-neighborhoods for KSG estimators $I^{(1)}$ and $I^{(2)}$ from Kraskov et al. Detailed description is given in the text. Used with permission.

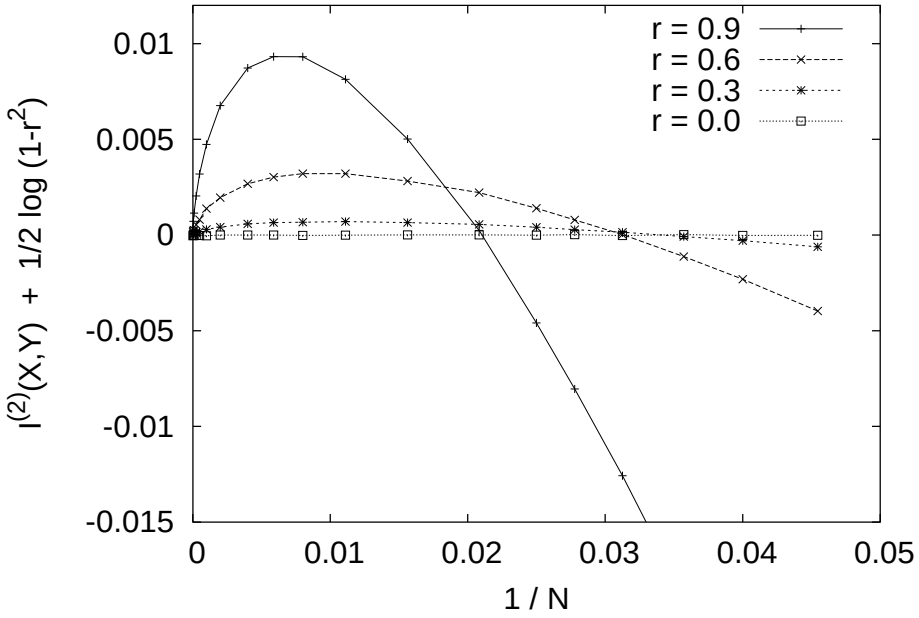

Figure 6: Convergence of the KSG estimator from Kraskov et al. for Gaussians with given correlation coefficient. Used with permission.