# 1 Information measures

## 1.1 Introduction to information theory

In order to get an insight into transfer entropy and active information storage we have to define some key information theoretic concepts. We will follow the definitions according to MacKay [1]. At first we define the Shannon information content and entropy of a discrete random variable $X$.

**Definition 1** *Shannon information content of an outcome $x$ is defined to be*

$$h(x) = \log_2 \frac{1}{P(x)}. \tag{1}$$

*The units are called bits.*

**Definition 2** *The entropy of an ensemble $X$ is defined to be the average Shannon information content of an outcome:*

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{A}_X} P(x) \log \frac{1}{P(x)} && \textit{for } P(x) \neq 0 \\ &= 0 && \textit{for } P(x) = 0 \end{aligned} \tag{2}$$

*since* $\lim_{\theta \to 0^+} \theta \log \frac{1}{\theta} = 0$.

Next we need to define conditional entropy.

**Definition 3** *The conditional entropy of $X$ given $y = b_k$ is the entropy of the probability distribution $P(x \mid y = b_k)$.*

$$H(X \mid y = b_k) = \sum_{x \in \mathcal{A}_X} P(x \mid y = b_k) \log \frac{1}{P(x \mid y = b_k)} \tag{3}$$

**Definition 4** *The conditional entropy of $X$ given $Y$ is average, over $y$, of the conditional entropy of $X$ given $y$.*

$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{A}_y} P(y) \left[ \sum_{x \in \mathcal{A}_X} P(x \mid y) \log \frac{1}{P(x \mid y)} \right] \\ &= \sum_{x \in \mathcal{A}_X, y \in \mathcal{A}_Y} P(x, y) \log \frac{1}{P(x \mid y)}. \end{aligned} \tag{4}$$

The interpretation of conditional transfer entropy is that it measures the average uncertainty that remains about $X$ when $Y$ is known.

Now we can define mutual information and conditional mutual information.

**Definition 5** *The mutual information between $X$ and $Y$ is*

$$I(X;Y) = H(X) - H(X \mid Y). \tag{5}$$

Mutual information measures the reduction of uncertainty about X that results from learning the value of y.

**Definition 6** *The conditional mutual information between $X$ and $Y$ given $z = c_k$ is the mutual information between the random variables $X$ and $Y$ in the joint ensemble $P(x, y \mid z = c_k)$,*

$$I(X;Y \mid z = c_k) = H(X \mid z = c_k) - H(X \mid Y, z = c_k). \tag{6}$$

**Definition 7** *The conditional mutual information between $X$ and $Y$ given $z = c_k$ is the mutual information between the random variables $X$ and $Y$ in the joint ensemble $P(x, y \mid z = c_k)$,*

$$I(X;Y \mid z = c_k) = H(X \mid z = c_k) - H(X \mid Y, z = c_k). \tag{7}$$

**Definition 8** *The conditional mutual information between $X$ and $Y$ given $Z$ is the average over $z$ of the conditional mutual information from Definition 7.*

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z). \tag{8}$$

Entropy and all other measures based on entropy we have just introduced are defined for a discrete set of probabilities. Shannon in [2] analogously defined the entropy of a continuous distribution with the density distribution function $p(x)$ as

$$H(X) = \int_{-\infty}^{\infty} p(x) \log \frac{1}{p(x)} dx. \tag{9}$$

This quantity is sometimes called *differential entropy* although this is not a measure of uncertainty of the random variable X as Shannon intended it to be.

Similarly, from the identity

$$I(X;Y) = \sum_{x \in \mathcal{A}_X, y \in \mathcal{A}_Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{10}$$

we can define mutual information of a continuous variable as:

$$I(X;Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy. \tag{11}$$

## 1.2 Limiting density of discrete points

As it has been already noted in the previous section, Shannon's formulation of entropy for continuous variables is not an information measure. First it lacks many properties of discrete entropy and second it is not a result of any proper derivation. Nevertheless differential entropy has many theoretical applications.

Jaynes in [12] proposed a different approach to defining entropy of a continuous variable. Let $x_i, i = 1, \ldots, n$ be discrete points such that

$$\lim_{n \to \infty} \frac{1}{n} (\text{number of points in a} < x < \text{b}) = \int_a^b m(x)dx \tag{12}$$

exists. Then the differences $(x_{i+1} - x_i)$ in the neighbourhood of any particular value of $x$ will tend to zero so that

$$\lim_{n \to \infty} n(x_{i+1} - x_i) = m(x_i)^{-1}. \tag{13}$$

The discrete probability $P(x_i)$ from Definition 2 will transform to

$$P(x_i) = p(x_i)(x_{i+1} - x_i). \tag{14}$$

Equivalently from equation (13)

$$P(x_i) \to p(x_i) \frac{1}{nm(x_i)}. \tag{15}$$

Plugging (14) and (15) into the definition of Shannon entropy Definition 2 we get

$$H(X) \to - \int p(x) log \left[ \frac{p(x)}{nm(x)} \right] dx. \tag{16}$$

Following this reasoning, Jaynes defined the continuous information measure as:

$$\overline{H}(X) = \lim_{n\to\infty} [H(X) - \log(n)] = -\int p(x)log\left[\frac{p(x)}{m(x)}\right] dx. \tag{17}$$

## 1.3 Transfer entropy and active information storage

Let $\{x(t)\}$ and $\{y(t)\}$ be the realizations of two processes $\{X(t)\}$ and $\{Y(t)\}$. Paluš in [3] proposed, as an approach to measure the directional information rate, to measure conditional mutual information $I(y(t); x(t+\tau) \mid x(t))$. It is the average amount of information contained in the process $\{Y(t)\}$ about the process $\{X(t)\}$ in its future $\tau$ time units ahead conditioned on $X(t)$, as opposed to the mutual information $I(y(t); x(t+\tau)$ which can contain information about $X(t+\tau)$ in $X(t)$. Similarly Shreiber in [4] defines mutual information rate of two Markov processes $\{I\}$ and $\{J\}$ of order $k$ and $l$ by measuring the deviation from independence given by the Markov property $p(i_{t+1} \mid i_t^{(k)}) = p(i_{t+1} \mid i_t^{(k)}, j_t^{(l)})$ using conditional Kullback-Leibler divergence in the following form

$$D_{KL}(P(I_{t+1} \mid I_t^{(k)}, J_t^{(l)}) \| P(I_{t+1} \mid I_t^{(k)}, J_t^{(l)})) =$$

$$= \sum p(i_{t+1}, i_t^{(k)}, j_t^{(l)}) \log \frac{p(i_{t+1} \mid i_t^{(k)}, j_t^{(l)})}{p(i_{t+1} \mid i_t^{(k)})}.$$

These ideas inspired the current definition of transfer entropy.

**Definition 9** *Let $X_t$ and $Y_t$ be two processes. The transfer entropy from the source $Y$ with the history length $k$ to the target $X$ with history length $l$ is*

$$T_{Y\to X}^{(k,l)} = I(X_t \ ; \mathbf{Y}_{t-1}^{(l)} \mid \mathbf{X}_{t-1}^{(k)}) \tag{18}$$

*where*

$$\mathbf{X}_{t-1}^{(k)} = (X_{t-1}, X_{t-1-\tau_k}, X_{t-1-2\tau_k}, ..., , X_{t-1-(k-1)\tau_k})$$
$$\mathbf{Y}_{t-1}^{(l)} = (Y_{t-1}, Y_{t-1-\tau_l}, Y_{t-1-2\tau_l}, ..., , Y_{t-1-(l-1)\tau_l}).$$

From the identity for conditional mutual information

$$I(X; Y \mid Z) = H(X) - I(X; Z) - H(X \mid Y, Z)$$

it follow that transfer entropy can be expressed as

$$T^{(k,l)}_{Y\to X} = H(X_t) - I(X_t \mid \mathbf{X}^{(k)}_{t-1}) + H(X_t \mid \mathbf{Y}^{(l)}_{t-1}, \mathbf{X}^{(k)}_{t-1}).$$

The second term on the right in the expression above has a special significance for us.

**Definition 10** *The active information storage of the process $X$ with history length $k$ is*

$$A^{(k)}_X = I(\mathbf{X}^{(k)}_{t-1}; X_t) = H(X_t) - H(X_t \mid \mathbf{X}^{(k)}_{t-1}). \tag{19}$$

The active information storage measures the amount of information in the past state of $\mathbf{X}^{(k)}_t$ of $X$ about its next value $X_t$.

## 1.4 Transfer entropy and Granger causality

Transfer entropy is closely related to another concept of dependency, namely Granger causality. This relationship provides a different interpretation of the concept of transfer entropy. Lets look take a closer look at it.

We take the definition of Granger causality from [5]

**Definition 11** *Let us use the notation from Definition 9. Let $F(x_t|\mathbf{x}^{(k)}_{t-1}, \mathbf{y}^{(l)}_{t-1})$ be the distribution function of the target variable $X_t$ conditional on $\mathbf{X}^{(k)}_{t-1}$, $\mathbf{Y}^{(l)}_{t-1}$ and $F(x_t|\mathbf{x}^{(k)}_{t-1})$ be the distribution function of $X_t$ conditional on its own past, then variable $Y$ is said to Granger-cause variable $X$ if*

$$F(x_t|\mathbf{x}^{(k)}_{t-1}, \mathbf{y}^{(l)}_{y-1}) \neq F(x_t|\mathbf{x}^{(k)}_{t-1}). \tag{20}$$

Now, consider tow linear regression models

$$X_t = X_{t-1}A_1 + \ldots + X_{t-k}A_k + Y_{t-1}B_1 + \ldots + Y_{t-l}B_l + \epsilon_t \tag{21}$$

$$X_t = X_{t-1}A'_1 + \ldots + X_{t-k}A'_k + \epsilon'_t \tag{22}$$

with parameters of the models $A_i, A'_i, B'_j$. Geweke in [10] defined the measure of Granger causality from $Y$ to $X$ the following way,

$$F^{(k,l)}_{Y\to X} = \log(|\mathrm{var}(\epsilon'_t)|/|\mathrm{var}(\epsilon_t)|) \tag{23}$$

where $|\cdot|$ denotes the determinant. We can observe that this is actually log-likelihood ratio test under the null hypothesis

$$H_0 : B_1 = \cdots = B_l = 0 \tag{24}$$

when the residuals of the both model is gaussian.

From what has already been written it can be seen that both transfer entropy and Granger causality are measures of predictive causality. The similarity is even closer when we consider the joint process $X_t, Y_t$ as multivariate gaussian. Barnett et. al. in [11] proved that under these conditions that Granger causality and transfer entropy are equivalent up to factor 2 i.e.

$$F^{(k,l)}_{Y\to X} = 2T^{(k,l)}_{Y\to X}. \tag{25}$$

## 1.5 Estimation of transfer entropy and active information storage

Estimation of entropy measures is an open problem. Currently there are few available classes of estimators that one can choose from based on the properties of the observed data. Transfer entropy is no exception and the best estimator given some specific criteria is yet to be determined. An overview of transfer entropy estimators can be found in [5] or [6].

We are going to focus on one specific estimator of the class of estimators based on k-nearest neighbour search.

### 1.5.1 Kozachenko-Leonenko Shannon entropy estimator

We are going to put the basic idea of behind Kozachenko-Leonenko differential entropy estimator as was presented in [7].

Let $X$ be a continuous random variable, $f(x)$ be its density and its differential entropy as defined in (9). Specifically

$$H(X) = \int_{-\infty}^{\infty} f(x) \log \frac{1}{f(x)} dx. \tag{26}$$

Then the Monte-Carlo estimate of $H(X)$ is

$$\widehat{H}(X) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{1}{f(x_i)}. \tag{27}$$

Since we don't know $f(x_i)$ it has to be substituted by an estimate $\widehat{f}(x_i)$ which we was found using k-nearest neighbours of $x_i$.

Let $P_k(\epsilon)$ be the probability distribution of the distance between $x_i$ and its $k$th nearest neighbour. Then $P_k(\epsilon)d\epsilon$ is the probability that there is one point in the $r$ distance from $x_i$ where $r \in< \frac{\epsilon}{2}; \frac{\epsilon}{2} + \frac{d\epsilon}{2} >$, $k-1$ points are at distances less than $r$ and $N-k-1$ points are at distances greater than $k$th nearest neighbour. Using multinomial distribution formula we get

$$P_k(\epsilon)d\epsilon = \frac{(N-1)!}{1!(k-1!)(N-k-1)!}\left(\frac{dp_i(\epsilon)}{d\epsilon}d\epsilon\right)(p_i(\epsilon))^{k-1}(1-p_i(\epsilon))^{N-k-1} \tag{28}$$

where $p_i(\epsilon)$ is the probability mass of $\epsilon$ ball centered at $x_i$, that is

$$p_i(\epsilon) = \int_{\|\xi - x_i\| < \frac{\epsilon}{2}} f(\xi)d\xi. \tag{29}$$

It follows from (28) and (29) that the expectation value $\log p_i(\epsilon)$ is given by

$$\begin{aligned} E(\log p_i(\epsilon)) &= \int_0^\infty \log p_i(\epsilon) P_k(\epsilon)d\epsilon \\ &= \psi(k) - \psi(N) \end{aligned} \tag{30}$$

where $\psi(x)$ is the digamma function.

If we assume that $f(x)$ is constant in the entire $\epsilon$ ball we can approximate $p_i(\epsilon)$ by

$$p_i(\epsilon) \approx c_d \epsilon^d f(x_i), \tag{31}$$

where d is the dimension of $x$ and $c_d$ is the volume of the d-dimensional unit ball. For the maximum norm $c_d = 1$.

Finally taking the logarithm and expectation of (31), and combining it with (30) and (27) we get Kozachenko-Leonenko entropy estimator

$$\widehat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N}\sum_{i=1}^{N}\log \epsilon(i), \tag{32}$$

where $\epsilon(i)$ is twice the distance from $x_i$ to its $k$th nearest neighbour.

### 1.5.2 Kraskov-Stögbauer-Grassberger mutual information estimator

Kraskov, Stögbauer and Grassberger in [7] came with a method of using Kozachenko-Leoneko entropy estimator to estimate mutual information. Using the identity

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{33}$$

we are going to obtain an estimator $\widehat{I}(X,Y)$.

First we directly apply Kozachenko-Leonenko entropy estimator to joint random variable $Z=(X,Y)$ with maximum norm and we get

$$\widehat{H}(X,Y) = -\psi(k) + \psi(N) + \frac{d_X + d_Y}{N} \sum_{i=1}^{N} \log \epsilon(i) \tag{34}$$

where $\epsilon(i)$ is the $\frac{\epsilon}{2}$ from $z_i$ to its $k$th nearest neighbour and $d_Z = d_X + d_Y$.

For the estimate of $H(X)$ we take the distance $\epsilon(i)$ from (34) as an approximation of $[n_x(i)+1]$st nearest neighbour of $x_i$, where $n_x(i)$ is the number of points in within $\|x_j - x_i\| < \frac{\epsilon}{2}$, and we get

$$\widehat{H}(X) = -\frac{1}{N} \sum_{i=1}^{N} \psi(n_x(i)+1) + \psi(N) + \frac{d_X}{N} \sum \log \epsilon(i). \tag{35}$$

Analogously for the marginal space $Y$ we get

$$\widehat{H}(Y) = -\frac{1}{N} \sum_{i=1}^{N} \psi(n_y(i)+1) + \psi(N) + \frac{d_Y}{N} \sum_{i=1}^{N} \log \epsilon(i). \tag{36}$$

Combining (33), (34), (35) and (36) we get the first Kraskov, Stögbauer and Grassberger estimator

$$I^{(1)}(X,Y) = \psi(k) - \langle \psi(n_x+1) + \psi(n_y+1) \rangle + \psi(N) \tag{37}$$

where $\langle \cdots \rangle = \frac{1}{N} \sum_{i=1}^{N} (\cdots)$.

Another approach Kraskov et al. have taken in [7] to estimate mutual information was to replace $P_k(\epsilon)$ in (28) in the Kozachenko-Leonenko estimattion method by a two-dimensional density:

$$P_k(\epsilon_x, \epsilon_y) = P_k^{(b)}(\epsilon_x, \epsilon_y) + P_k^{(c)}(\epsilon_x, \epsilon_y) \tag{38}$$

with

$$P_k^{(b)}(\epsilon_x, \epsilon_y) = \binom{N-1}{k} \frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y} (1-p_i)^{N-k-1} \tag{39}$$

and

$$P_k^{(c)}(\epsilon_x, \epsilon_y) = (k-1) \binom{N-1}{k} \frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y} (1-p_i)^{N-k-1} \tag{40}$$

where $q_i = q_i(\epsilon_x, \epsilon_y)$ is the probability mass of $\epsilon_x \times \epsilon_y$ rectangle centered at $(x_i, y_i)$ and $p_i$ is the probability mass of the $\epsilon = \max\{\epsilon_x, \epsilon_y\}$ ball using the maximum norm. With this modifications (30) takes the the following form

$$\begin{aligned} E(\log q_i(\epsilon_x, \epsilon_y)) &= \int_0^\infty \int_0^\infty \log q_i(\epsilon_x, \epsilon_y) P_k(\epsilon_x, \epsilon_y) d\epsilon_x d\epsilon_y \\ &= \psi(k) - \frac{1}{k} - \psi(N). \end{aligned} \tag{41}$$

Next, following the same reasoning as in the case of $I^{(1)}(X, Y)$ we arrive at a second version of Kraskov, Stögbauer and Grassberger estimator, namely

$$I^{(2)}(X, Y) = \psi(k) - \frac{1}{k} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \tag{42}$$

where in this case $n_x(i)$ and $n_y(i)$ are number of points within $\|x_i - x_j\| <= \frac{\epsilon_x(i)}{2}$ and $\|y_i - y_j\| <= \frac{\epsilon_y(i)}{2}$, respectively.

### 1.5.3 Estimating transfer entropy and active information storage estimator

Since active information storage of a process is defined in Definition 10 as mutual information between its current and past state, it follows from (42) that the active information estimator is written as

$$\widehat{A}_X^{(k)} = \psi(k) + \psi(N) - \frac{1}{k} - \left\langle \psi(n_{\mathbf{x}_{t-1}^{(k)}}) + \psi(n_{x_t}) \right\rangle. \tag{43}$$

As for an estimate of transfer entropy Gomez-Herrero et. al. [8] generalized the idea of Kraskov et. al. [7]. Let $V = (V_1, \ldots, V_m)$ be a random $m$-dimensional vector. Then the entropy combination is defined as

$$C(V_{\mathcal{L}_1}, \ldots, V_{\mathcal{L}_p}) = \sum_{i=1}^{p} s_i H(V_{\mathcal{L}_1}) - H(V) \tag{44}$$

where $\forall i \in < i, p >$: $\mathcal{L}_i \subset < 1, m >$ and $s_i \in \{-1, 1\}$ such that $\sum_{i=1}^{p} s_i \chi_{\mathcal{L}_i} = \chi_{<1,m>}$ where $\chi_S$ is characteristic function of $S$ and the entropy combination estimator is given by

$$\widehat{C}(V_{\mathcal{L}_1}, \ldots, V_{\mathcal{L}_p}) = F(k) - \sum_{i=1}^{p} s_i \langle F(k_i(j)) \rangle \tag{45}$$

where $F(k) = \psi(k) - \psi(N)$ and $k_i(j) = n_{V_{\mathcal{L}_i}}(j) + 1$ for $j = 1, \ldots, N$ as was formulated in section 1.5.2.

For example, from equation (33) it can be seen that mutual information is an entropy combination. Similarly from the one of the expressions for conditional mutual information

$$I(X; Y \mid Z) = H(X, Z) + H(Y, Z) - H(Z) - H(X, Y, Z) \tag{46}$$

it follows that $I(X; Y \mid Z)$ is also an entropy combination and combining (45) and (46) we get the Kraskov, Stögbauer and Grassberger estimator for conditional mutual information

$$\widehat{I}(X; Y \mid Z) = \psi(k) - \langle \psi(n_{xz} + 1) + \psi(n_{yz} + 1) - \psi(n_z + 1) \rangle \,. \tag{47}$$

Finally directly from the definition 9 we get the Kraskov, Stögbauer and Grassberger transfer entropy estimator

$$\widehat{T}^{(k,l)}_{Y \to X} = \psi(k) - \left\langle \psi(n_{x_t, \mathbf{x}^k_{t-1}} + 1) + \psi(n_{\mathbf{y}^l_{t-1}, \mathbf{x}^k_{t-1}} + 1) - \psi(n_{\mathbf{x}^k_{t-1}} + 1) \right\rangle \,. \tag{48}$$

## 1.6 Estimators parameter determination

The Kraskov, Stögbauer and Grassberger belongs to a class of nonparametric estimation methods. Nevertheless, since its based on $k$-nearest neighbour searches, $k$ in the nearest neighbour algorithm is one of the parameters we have to chose.

This parameter has to be empirically determined. Kraskov et. al. [7] suggest $k > 1$ but not too large because of the increase of systematic errors. Generally they propose to use $k = 2 - 4$. Bossomaier et. al. in [5] writes that for $k \geq 4$ the estimator is robust and Wibral in [6] writes that $k = 4$ has been determined as a good choice for ECoG data.

Another problem is determining the embedding vector as was defined in Definition 9. For example, in case of transfer entropy we would like as much of the information that is contained in the target process to condition out. If we won't do this, the measured information transfer might be due to self prediction of the target process and not the dependance of the source and target processes. The method for finding the optimal values for the embedding vector is also called state space reconstruction.

There are multiple methods to determine the state space vectors. One way is to set $(m, \tau)$ such that it maximizes the active information storage.

Wibral et. al. [6] propose to optimize the embedding parameters using the Ragwitz-Kantz criterion.

Ragwitz and Kantz in [9] proposed a method to extract a Markov process of order $m > 1$ from observed scalar time series using locally constant predictors. In the following paragraphs we will briefly describe this method.

Let $\vec{s}_n = (s_n, s_{n-\tau}, \ldots, s_{n-(m-1)\tau})$ be time delay embedding vector such that $s_{n+1} = g(\vec{s}_n)$ exists. Then the locally constant predictor for the unobserved $s_{n+1}$ is

$$\widehat{s_{n+1}} = \frac{1}{|\mathcal{U}_n|} \sum_{\vec{s}_k \in \mathcal{U}_n} s_{k+1} \tag{49}$$

where $\mathcal{U}_n = \{\vec{s}_k : \|\vec{s}_k - \vec{s}_n\| \leq \epsilon\}$. In other words, the mean of the immediate futures of the $\epsilon$-neighbours of $\vec{s}_n$. Ragwitz and Kantz in [9] argue that locally constant predictor is a predictor based on Markov transition probability assumption $p(s_{n+1}|s_n, s_{n-\tau}, \ldots, s_\tau) = p(s_{n+1}|s_n, s_{n-\tau}, \ldots, s_{n-(m-1)\tau})$. To get the transition the probabilities $p(s_{n+1}|\vec{s}_n)$, a locally constant approximation is used in the form of $p(s_{k+1}|\vec{s}_k) \approx \widehat{p}(s_{n+1}|\vec{s}_n)\ \forall\ \vec{s}_k \in \mathcal{U}_n$. The justification of locally constant predictor follows then from the idea that if we use the mean square error of predictions $e^2 = \sum_{i=1}^{N}(s_{i+1} - \widehat{s}_{i+1})^2$ then the best estimator of $s_{n+1}$ is $\widehat{s}_{n+1} = \int_{\vec{s}_k \in \mathcal{U}_n} s_{k+1} p(s_{k+1}|\vec{s}_n) ds_{k+1}$. Thus we take those time delay embedding vector parameters as optimal which minimize the locally constant prediction error i.e.

$$(m, \tau) = \operatorname*{arg\,min}_{m \in \mathbb{Z}, \tau \in \mathbb{Z}} \left\| \vec{s} - \widehat{\vec{s}} \right\| . \tag{50}$$

## 1.7 Transfer entropy significance testing

One of the problems that plague all transfer entropy estimators is bias. Either systematic errors or statistical errors are both properties of estimators that has to be dealt with. If we observe non-zero value of transfer entropy it can be due to bias or variance and the transfer entropy estimator (48) is no exception. Since Kraskov, Stögbauer, Grassberger method is non-parametric one of the suggested statistical testing options is to use a permutation test [6] [5].

We are going to test the null hypothesis that there is no relationship between the source and target variables. First we have to come with surrogate source variable under to assumption that the null hypothesis is true. One way to do it is to by permuting $\mathbf{y}_{t-1}^{(l)}$ in the joint probability space $\{x_t, \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_{t-1}^{(l)}\}$ thus destroying any predictive dependence of $Y$ to $X$, and computing $T_{Y_{sur} \to X}$

of the surrogate source variable $Y_{sur}$. Repeating this procedure multiple times we can compute one-sided p-value of the test by the following equation:

$$P(T_{Y_{sur}\to X} \geq T_{Y\to X}) = \sum_{Y_{sur}:T_{Y_{sur}\to X}\geq T_{Y\to X}} P(Y_{sur}). \tag{51}$$

If the p-value is less 0.05 we reject the null hypothesis at the 5% significance level.

# References

[1] D. MacKay, Information Theory, Inference, and Learning Algorithms (fourth printing), Cambridge University Press, 2005.

[2] C. E. Shannon, "A mathematical theory of communication", Bell Systems Technical Journal, vol. 27, pp. 379–423, 1948.

[3] M. Paluš, V. Komárek, Z. Hrnčíř, K. Šteřbová, "Synchronization as adjustment of information rates: Detection from bivariate time series", Physical Review E, vol. 63, p. 046211, 2001.

[4] T. Schreiber, "Measuring information transfer", Physical Review Letters, vol. 85, no. 2, pp. 461–464, 2000.

[5] T.Bossomaier, L.Barnett, J. T.Lizier, An Introduction to Transfer Entropy, Springer, 2016.

[6] M. Wibral, R. Vicente, J. T.Lizier, Eds., Directed Information Measures in Neuroscience. Springer, 2014.

[7] A. Kraskov, H. Stgbauer, P. Grassberger, "Estimating mutual information,: Physics Reviews E, vol. 69, p. 066138, 2004.

[8] G. Gómez-Herrero, W. Wu, K. Rutanen, M. C. Soriano, G. Pipa, R. Vicente, "Assessing coupling dynamics from an ensemble of time series", Entropy, vol. 17, no.4, pp. 1958–1970.

[9] M. Ragwitz, H. Kantz, "Markov models from data by simple nonlinear time series predictors in delay embedding spaces",Physics Reviews E, vol. 65, no. 5, p. 056201, 2002.

[10] J. Geweke, ”Measurement of linear dependence and feedback between multiple time series”, Journal of American Statistical Association, vol. 77, no. 378, pp. 304–313, 1982.

[11] L. Barnett, A. B. Barrett, A. K. Seth, ”Granger causality and transfer entropy are equivalent for Gaussian variables”, Physical Review Letters, vol. 103, no. 23, p. 238701, 2009.

[12] E. T. Jaynes, Probability Theory: The Logic of Science, Cambridge University Press, 2003.