Estimation of Transfer Entropy between Discrete and Continuous Random Processes

Juliana M. de Assis, Francisco M. de Assis

***Abstract*—Transfer entropy is a measure of causality that has been widely applied and one of its identities is the sum of mutual information terms. In this article we evaluate two existing methods of mutual information estimation in the specific application of detecting causality between a discrete random process and a continuous random process: binning method and nearest neighbours method. Simulated examples confirm, in the overall scenario, that the nearest neighbours method detects causality more reliably than the binning method.**

***Index Terms*—Transfer entropy, causality, continuous process, discrete process, estimation, nearest neighbours, binning.**

I. Introduction

Transfer entropy (TE), as well as Granger causality and directed information, is a measure of causality. Firstly introduced by Schreiber [1], TE has been proposed as an effective measure of causality in industry [2], in order to detect where was a disturbance in the industrial process. TE has also been proposed as a powerful mean to detect neural connections in neuroscience [3]. Another application of TE in neuroscience has been to detect reliably the cerebral hemisphere containing epileptic focus. This was made without observing actual seizure activity by using TE [4]. There are many other applications of transfer entropy in the literature, specially in neuroscience field [5]–[8]. There is also recent application of TE in medicine [9] and in observational climate data [10].

The relation between TE and directed information has been explored recently. Amblard *et al.* proved that directed information rate is the sum of two parts, one of which is equivalent to a particular mode of transfer entropy, and the other to the instantaneous information exchange rate, for a stationary process [11]. Additionally, Liu and Aviyente proved that if $\mathbf{X}$ and $\mathbf{Y}$ are two stationary processes, without instantaneous information exchange and such that the distribution of the present value of Y_n given the whole past of $\mathbf{X}$ and $\mathbf{Y}$ is equal to the distribution of Y_n given ℓ past values of $\mathbf{X}$ and m past values of $\mathbf{Y}$, then transfer entropy is the upper bound of directed information rate [12]. Directed information has also been applied in diverse fields, such as neuroscience [13]–[15], and economy [16].

Usually, TE has to be estimated from data. This happens because in most cases of interest, probability distributions of the involved time series are not available (we will show later that TE is an information measure which relies on probability distributions of the random processes). TE has been applied to time series assuming discrete or continuous values. However, we are not aware of its application to a mixed case, that is, a case of measuring whether a discrete process causes a continuous (in amplitude) process, or *vice versa*. The purpose of this paper is to evaluate TE estimators for these mixed cases, which may be of interest for those working with mixed processes and with a causality measure necessity. For example, this may be relevant in the context of a Poisson channel with feedback. The model of a discrete time Poisson channel involves the use of a continuous random process as input and a discrete random process as output [17]. Also, it has already been established that when feedback is present directed information gives a tighter bound on the capacity of the channel [18] (directed information is intimately related to TE, as mentioned before).

Two methods of estimation are explored in this paper for this case of mixed processes. Both methods stems from an identity for TE, written as a sum of two mutual information terms. The first method uses a very popular method of estimation of mutual information, which is based in adaptive partitioning of the support of the continuous variable, here called binning. The second method is based on the estimation of mutual information based on the distribution of k nearest neighbours (NN), for a mixed distribution, as proposed by Ross [19].

This paper is organized as follows: Section II establishes some notation and terminology, Section III defines TE mathematically, Section IV brings the development of the TE estimators for mixed cases, and Section V shows the results of the estimation with the TE estimators in different situations. Finally, Section VI concludes the paper.

II. Notation and Terminology

In this paper, we denote random variables by uppercase letters, stochastic processes by uppercase bold letters, and specific values assumed by them in lowercase letters. Subscripts denote the outcome's position in a sequence, for example, X_n generally indicates the n^{th} outcome of the process $\mathbf{X}$. Superscripts on a random variable denote finite length sequences of this random variable, for example, $X^N = \{X_1, X_2, \ldots, X_N\}$, and $X_2^4 = \{X_2, X_3, X_4\}$. Throughout this paper, ln is the logarithm in natural base, $\mathbb{E}(X)$ indicates the mean of X. Shannon entropy of a random variable X is denoted by $H(X)$. $H(X)$ also stands for differential entropy (when X assumes continuous values).

III. Definitions

Firstly, we stress that causality, as used in this paper, is based in Norbert Wiener's concept, which states that one

Juliana M. de Assis was with the Department of Electrical Engineering, UFCG, Brazil, e-mail: juliana.assis@ee.ufcg.edu.br

Francisco M. de Assis was with the Department of Electrical Engineering, UFCG, Brazil, e-mail: fmarcos@dee.ufcg.edu.br

Digital Object Identifier: 10.14209/jcis.2018.1

process causes another if the knowledge of the past of the first process is useful in predicting the future of the second process. An interesting observation is that when defined this way, causality is not equal to the definition of causality we are normally used to. This happens because Wiener's causality does not take into account hidden causes, but is limited to the considered random processes. Granger causality is also based in Wiener's concept. However, Granger causality is developed assuming that the processes present a particular model, specifically that they can be described as autoregressive processes. This does not always hold necessarily.

Considering two random processes $\mathbf{X}$ and $\mathbf{Y}$, Schreiber [1] defined TE from $\mathbf{X}$ to $\mathbf{Y}$ as:

$$TE_n(X \to Y) = \sum_{y_{n-m}^n, x_{n-\ell}^{n-1}} P(y_n, y_{n-m}^{n-1}, x_{n-\ell}^{n-1}) \log \frac{P(y_n|y_{n-m}^{n-1}, x_{n-\ell}^{n-1})}{P(y_n|y_{n-m}^{n-1})}, \quad (1)$$

which measures the deviation from the generalized Markov property:

$$P(y_n|y_{n-m}^{n-1}, x_{n-\ell}^{n-1}) = P(y_n|y_{n-m}^{n-1}). \quad (2)$$

We can see from equation (1) that TE is not symmetric, so it does not constitute a metric (but for a causality measure it is a desirable property). Equation (1) equals the KL distance between distributions $P(y_n|y_{n-m}^{n-1}, x_{n-\ell}^{n-1})$ and $P(y_n|y_{n-m}^{n-1})$. Notice that TE measures how easier it is to predict Y_n when we know past values of $X_{n-\ell}^{n-1}$ and Y_{n-m}^{n-1} than when we know only the past values of Y_{n-m}^{n-1}. The chosen values for past indexes ℓ and m are relevant in the evaluation of TE, and, for computational reasons, a preferable choice for these indexes is simply $\ell = m = 1$ [1]. Notice also that the definition of TE is dependent of the time index n, unless $\mathbf{X}$ and $\mathbf{Y}$ are jointly stationary processes. Despite the fact that in this paper we shall not always use stationary processes, we may drop the term n of the expression $TE_n(X \to Y)$ for simplicity. It will be clear for each example in which time instant n the estimation is performed. Moreover, one particular case of interest here is measuring the asymptotic limit of TE, that is

$$TE_\infty(X \to Y) = \lim_{n\to\infty} TE_n(X \to Y). \quad (3)$$

As mentioned before, here we consider transfer entropy as a sum of mutual information terms. Mutual information between two random variables X and Y is a symmetric measure defined as [20]:

$$I(X;Y) = \sum_{\mathcal{X},\mathcal{Y}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}. \quad (4)$$

Mutual information is also defined when conditioned on the knowledge of one or more random variables. For example, mutual information between random variables X and Y when conditioned on the known random variable W is [20]:

$$I(X;Y|W) = \sum_{\mathcal{X},\mathcal{Y},\mathcal{W}} P(x,y,w) \log \frac{P(x,y|w)}{P(x|w)P(y|w)}. \quad (5)$$

For the case of interest in this paper, when the random variable X is discrete and the random variable Y is continuous, mutual information is defined as:

$$I(X;Y) = \sum_{\mathcal{X}} P(x) \int f(y|x) \ln \frac{f(y|x)}{f(y)} dy. \quad (6)$$

Now, with the definition of the functional $I(.;.)$, it is possible to write the following identity for (1):

$$\begin{aligned} TE_n(X \to Y) &= I(Y_n; X_{n-\ell}^{n-1}|Y_{n-m}^{n-1}) \\ &= I(Y_n, Y_{n-m}^{n-1}; X_{n-\ell}^{n-1}) - \\ &\quad I(Y_{n-m}^{n-1}; X_{n-\ell}^{n-1}), \end{aligned} \quad (7)$$

which will be useful to estimate TE.

IV. Estimators

Since generally we do not have access to the probability distributions of the processes whose possible causality relations are investigated, there are many proposed methods to estimate TE. For discrete random processes, the most common alternative to estimate the distribution consists in counting the frequencies of the observed values, which is called plug-in estimation. However, for continuous processes, the estimation is more intricate since each random variable can assume values in an infinite, uncountable, set. In this paper, we explore two established methods for estimating mutual information in mixed cases, in the particular application of TE estimation: binning method and nearest neighbours method.

A. Binning Method

One direct method to estimate probability densities consists in discretizing the continuous valued process, which is called here binning, and then applying the plug-in method. The observed relative frequencies of the discretized values can be then applied in the functional of some information measure. The binning method is applicable when all the involved processes are continuous or when the involved processes are mixed. The binning method is an adaptive partitioning method [21], and it is commonly applied with equipopulated bins.

In the mixed case, let (X_1^N, Y_1^N) be the N samples generated from an underlying distribution $f(x,y) = P(x)f(y|x)$. The samples Y_1^N are put in ascending order, and Q equipopulated intervals (bins) are chosen [22]:

$$\{\tilde{Y}\}_{i=1,2,\ldots,Q} = \{(-\infty, Y_{(1)}], (Y_{(1)}, Y_{(2)}], \ldots, (Y_{(Q-1)}, \infty)\}, \quad (8)$$

where $Y_{(i)}$ is the i-th Q-quantile of the samples Y_1^N. The estimated probability mass function of $\tilde{Y}_i$ is

$$\hat{P}(i) = \frac{n_i}{N} \approx \frac{Q}{N}, \quad (9)$$

where n_i counts the occurrence of $Y \in \tilde{Y}_i$ in the samples Y_1^N.

On the other hand, the estimated probability mass function of X is

$$\hat{P}(x) = \frac{n_x}{N}, \quad (10)$$

in which n_x counts the occurrences of $X = x$ in the samples X_1^N.

Similarly, the joint probability mass function of $(X, \tilde{Y}_i)$ is

$$\hat{P}(x,i) = \frac{n_{x,i}}{N}, \tag{11}$$

where $n_{x,i}$ counts the occurrences of $(X = x, Y \in \tilde{Y}_i)$ jointly in the samples (X_1^N, Y_1^N).

Thus, the mutual information estimation between X and Y with the binning method is

$$\hat{I}(X;Y) = \sum_{\mathcal{X}} \sum_{i=1}^{Q} \hat{P}(x,i) \log \frac{\hat{P}(x,i)}{\hat{P}(x)\hat{P}(i)}. \tag{12}$$

One major issue with the binning method is how to choose the appropriate number Q of bins. In this paper, we usually applied the rule proposed by Paluš [23], if we estimate the mutual information among r random variables, $Q \leq \sqrt[r+1]{N}$, where N is the sample size.

B. Nearest Neighbours Method

Recently, the estimation of information measures with the nearest neighbours method has gained attention. Kraskov *et al.* developed a method to estimate mutual information derived from data, when the random variables present continuous values [24]. Nearest neighbours estimators for mutual information are based in the Kozachenko-Leonenko estimator for differential entropy [25]. The underlying idea behind these estimators is to use the distance of the nearest neighbours to approximate the density of the random variables. Kozachenko and Leonenko derived the following formula to estimate $H(Y) = -\mathbb{E} \ln f(y)$ for a continuous random variable Y:

$$\hat{H}(Y) = -\psi(k) + \psi(N) + \ln c_d + \frac{d}{N} \sum_{n=1}^{N} \ln(\delta_n), \tag{13}$$

where ψ is the digamma function, k is a parameter that indicates the number of neighbours considered, N is the sample size, d is the dimension of Y, and c_d is the volume of the d-dimensional unit ball. The term δ_n corresponds to twice the distance from y_n to its k^{th} neighbour. In this derivation, $f(y)$ is approximated as a uniform distribution over the entire δ_n-ball centered in y_n. Since this assumption does not always hold, it constitutes the main reason for bias in Kozachenko-Leonenko estimators.

The essential idea of Kraskov *et al.* to estimate mutual information was to use a different parameter k to estimate the marginal entropies ($H(X)$ and $H(Y)$) and the joint entropy ($H(X,Y)$) with Kozachenko-Leonenko estimators. With those estimates, it is possible to use the identity [20]

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{14}$$

and estimate mutual information. It would also be possible to estimate mutual information considering the same k for marginal entropies and joint entropy estimates. However, bias is reduced by considering the same neighbour for one of the marginal entropies estimator and for joint entropy estimator (which means considering a different k for estimation in marginal and joint spaces). With this procedure, the last term of equation (13) will be partially cancelled, when subtracted in identity (14).

Inspired in the work of Kraskov *et al.*, Ross developed a similar method to estimate mutual information for a mixed case, that is, to estimate mutual information between discrete and continuous random variables [19]. Ross estimator has been indicated as more efficient than popular binning estimator, even when binning is applied with bias correction [26]. In order to understand Ross estimator, consider the identity [20]

$$I(X;Y) = H(Y) - H(Y|X). \tag{15}$$

Ross estimator essentially applies Kozachenko-Leonenko estimator twice, first to $H(Y)$ and then to $H(Y|X)$. In order to cancel bias in the subtraction, Ross estimator chooses a different parameter k for each entropy estimate, such that both entropy estimates consider the same neighbour (similarly to Kraskov estimator). Ross estimator chooses, among those points that had the same outcome of the discrete variable $X_n = x$, the fixed k^{th} closest neighbour. This neighbour will be at a distance δ_n. Then, Ross estimator counts the number of all j_n neighbours that are in a distance δ_n of point Y_n. In mathematical terms, we write, for each realization n of the pair (X,Y):

$$\hat{I}_n = \psi(N) - \psi(N_{X_n}) + \psi(k) - \psi(j_n), \tag{16}$$

where N_{X_n} is the number of points whose discrete random variable is the particular value assumed by X_n. Mutual information estimate is obtained by the sample mean of $\hat{I}_n$:

$$\hat{I} = \frac{1}{N} \sum_{n=1}^{N} \hat{I}_n. \tag{17}$$

Fig. 1 illustrates the application of Ross estimator.

V. Results

The contribution of this paper is to investigate the application of both binning method and nearest neighbours method described above for mutual information, in equation (7), to estimate TE between discrete and continuous random processes, and detect causality. This has not been done yet in the reviewed literature.

In order to evaluate the performance of these methods, we have developed some examples involving causality in mixed cases. We assume ergodicity of the processes when evaluating TE estimates. In all simulations, when using the nearest neighbours method, we set the number of neighbours as $k = 3$, as recommended in the literature [19], [24]. Besides, when using the binning method, we set the number of bins as

$$Q = \max\left\{2, \left\lfloor \sqrt[m+\ell+1]{N} \right\rfloor\right\}, \tag{18}$$

since the dimensions of vectors in equation (7) are equal to the past indexes of processes $\mathbf{X}$ and $\mathbf{Y}$, ℓ and m, respectively. Thus, the minimum value for Q is limited at 2.

A. First Example: TE from Discrete to Continuous

In the first example, we have a discrete random process $\mathbf{X}$ that is causally influencing a continuous random process $\mathbf{Y}$. $\mathbf{X}$ is a Markov chain whose state diagram is given in Fig. 2.

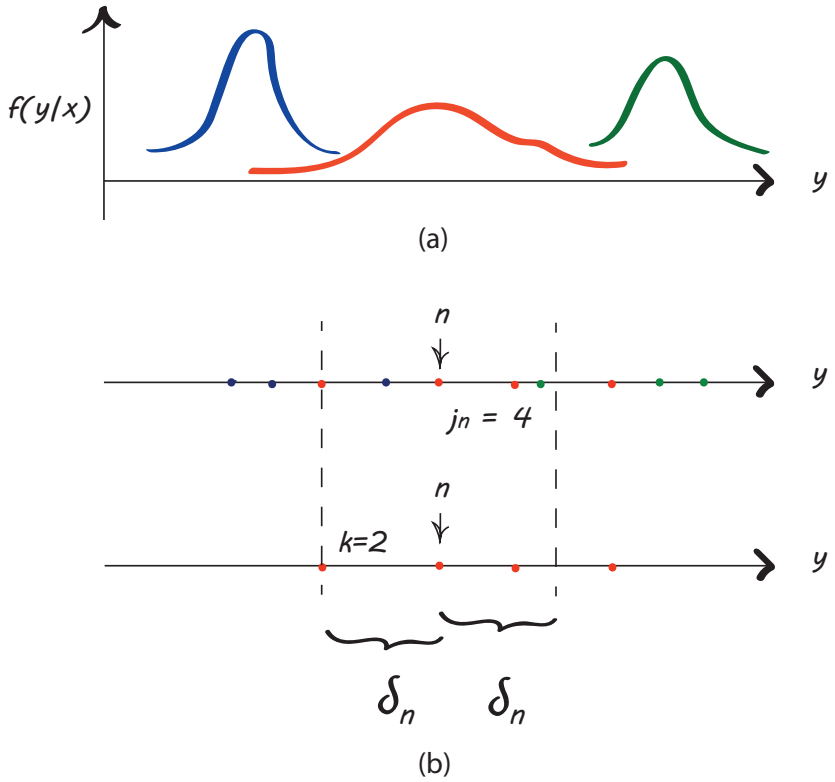


Fig. 1: Illustration of Ross estimator. In panel (a), sketch of three conditioned probability density functions $f(Y|X)$, represented by blue, red and green curves. In panel (b), on the superior axis, pairs of data $Z = (X, Y)$, where the values of Y are represented by the dots position on the y axis and the three possible values of X are represented by the color of the dots (blue, red or green). In panel (b), Z_n is indicated by a vertical arrow. Dashed lines indicate the distance δ_n from Z_n to its 2^{nd} neighbour with X "red", parameter $k = 2$. The 2^{nd} neighbour of Z_n on the lower axis is the 4^{th} ($j_n = 4$) neighbour on the superior axis, which considers all values of X. In this example, $N = 10$, $k = 2$, $N_{X_n} = 4$ and $j_n = 4$ (including the neighbour at a distance δ_n).

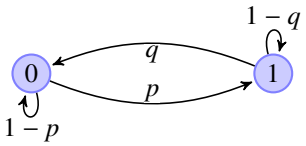


Fig. 2: State diagram for process **X**.

On the other hand, process **Y** is given by the following relation:

$$Y_n = \alpha Y_{n-1} + \gamma X_{n-1} + \epsilon \eta_n, \tag{19}$$

where α, γ and ϵ are fixed parameters and η_n is a standard Gaussian random variable. In our simulations, we set parameters $p = 1/2$ and $q = 3/4$ ($P(X_n = 0) = 0.6$ and $P(X_n = 1) = 0.4$ in steady state). The parameter γ varied in the range of $[-0.5, 0.5]$, in a step of 0.1. For each value set for γ, we simulated 50 trials of these random processes with duration $N = 1000$. We fixed $\alpha = 0.5$ and $\epsilon = 0.1$.

For each γ value, we may evaluate theoretical bounds for $TE_\infty(X \to Y)$, considering $m = \ell = 1$, as follows. Firstly, consider the identity for TE [27]:

$$TE_n(X \to Y) = H(Y_n|Y_{n-1}) - H(Y_n|Y_{n-1}X_{n-1})$$

$$TE_\infty(X \to Y) = \lim_{n\to\infty} [H(Y_n|Y_{n-1}) - H(Y_n|Y_{n-1}X_{n-1})]. \tag{20}$$

The second term of the right side of equation (20) is

$$\begin{aligned} \lim_{n\to\infty} H(Y_n|Y_{n-1}X_{n-1}) &= \\ &= \lim_{n\to\infty} H(Y_n - \alpha Y_{n-1} - \gamma X_{n-1}|Y_{n-1}X_{n-1}) \\ &= \lim_{n\to\infty} H(\epsilon\eta_n) \\ &= \frac{1}{2}\ln(2\pi e\epsilon^2). \end{aligned}$$

There are bounds to the first term of the right side of equation (20). As an upper bound, we write:

$$\begin{aligned} \lim_{n\to\infty} H(Y_n|Y_{n-1}) &= \lim_{n\to\infty} H(\gamma X_{n-1} + \epsilon\eta_n|Y_{n-1}) \\ &\leq \lim_{n\to\infty} H(\gamma X_{n-1} + \epsilon\eta_n), \end{aligned} \tag{21}$$

because conditioning does not increase entropy (notice also that X_{n-1} and Y_{n-1} are not independent, since both depend on X_{n-2}).

As $n \to \infty$, **X** reaches its steady state, so the underlying distribution of $U = \gamma X_{n-1} + \epsilon\eta_n$ in (21) is:

$$f_U(u) = p(X = 0)g_0(u) + p(X = 1)g_1(u), \tag{22}$$

where

$$\begin{aligned} g_0(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-u^2/(2\epsilon^2)}, \text{ and} \\ g_1(u) &= \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-(u-\gamma)^2/(2\epsilon^2)}, \end{aligned} \tag{23}$$

because η is standard Gaussian. Thus, $f_U(u)$ is a Gaussian mixture.

The entropy of a Gaussian mixture does not have an analytical solution [28]. In order to find a numerical approximation for the bound

$$\lim_{n\to\infty} H(Y_n|Y_{n-1}) \leq H(U),$$

we use the trapezoidal rule:

$$\int_a^b f(x)dx \approx (b-a)\frac{f(b)-f(a)}{2}. \tag{24}$$

In this case, we desire an approximation for the integral

$$H(U) = -\int_{-\infty}^{\infty} f_U(u)\ln f_U(u)du. \tag{25}$$

Fig. 3 is the graphic of $-f_U(u)\ln f_U(u)$, with $\gamma = 0.5$. We observed that $f_U(u)$ is approximately 0 for u outside the range $[-1.5, 1.5]$. Thus, we summed the approximation of (24) in intervals of $(b-a) = \Delta_u = 0.001$, from $u = -1.5$ to $u = 1.5$, and we found

$$H(U) \approx -0.2275 \text{ nats}.$$

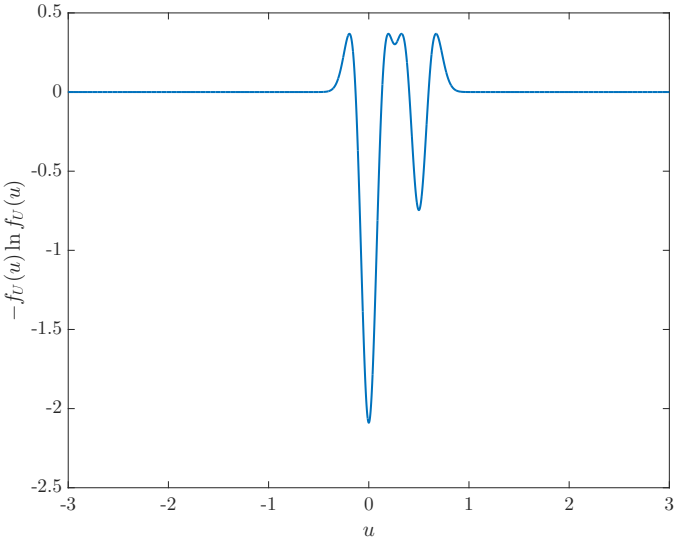


Fig. 3: Graphic for numerical integration of $-f_U(u)\ln f_U(u)$.

In order to find a lower bound, we write

$$\begin{aligned} \lim_{n\to\infty} H(Y_n|Y_{n-1}) &= \lim_{n\to\infty} H(\gamma X_{n-1} + \epsilon\eta_n|Y_{n-1}) \\ &\geq \lim_{n\to\infty} H(\gamma X_{n-1} + \epsilon\eta_n|X_{n-2}), \end{aligned} \tag{26}$$

because the sum $\gamma X_{n-1} + \epsilon\eta_n$ depends directly on X_{n-2} (and noise η_n is i.i.d.).

The underlying density $f_{U|X_{n-2}}(\gamma x_{n-1} + \epsilon\eta_n|x_{n-2})$ is given by:

$$\begin{aligned} f_{U|X_{n-2}}(u|x_{n-2}) &= f_{U|X_{n-2}}(\gamma x_{n-1} + \epsilon\eta_n|x_{n-2}) \\ &= \sum_{x_{n-1}} f_{U,X_{n-1}|X_{n-2}}(\gamma x_{n-1} + \epsilon\eta_n, x_{n-1}|x_{n-2}) \\ &= f_{U,X_{n-1}|X_{n-2}}(\epsilon\eta_n, X_{n-1}=0|x_{n-2}) \quad + \\ &+ f_{U,X_{n-1}|X_{n-2}}(\gamma+\epsilon\eta_n, X_{n-1}=1|x_{n-2}) \\ &= f_{U|X_{n-2}^{n-1}}(\epsilon\eta_n|X_{n-2}=x_{n-2}, X_{n-1}=0)P(X_{n-1}=0|x_{n-2}) \quad + \\ &+ f_{U|X_{n-2}^{n-1}}(\epsilon\eta_n+\gamma|X_{n-2}=x_{n-2}, X_{n-1}=1)P(X_{n-1}=1|x_{n-2}) \\ &= g_0(u)P(X_{n-1}=0|x_{n-2}) + g_1(u)P(X_{n-1}=1|x_{n-2}). \end{aligned} \tag{27}$$

Thus, when $X_{n-2} = 0$,

$$f_{U|X_{n-2}}(u|X_{n-2}=0) = g_0(u)(1-p) + g_1(u)p. \tag{28}$$

Analogously, when $X_{n-2} = 1$,

$$f_{U|X_{n-2}}(u|X_{n-2}=1) = g_0(u)q + g_1(u)(1-q), \tag{29}$$

and the lower bound may be evaluated through

$$\lim_{n\to\infty} H(U|X_{n-2}) =$$

$$\lim_{n\to\infty} -\sum_{x_{n-2}} P(x_{n-2}) \int_{-\infty}^{\infty} f_{U|X_{n-2}} \quad (u|x_{n-2}) \ln f_{U|X_{n-2}}(u|x_{n-2})du,$$

which, as done to find the upper bound, may be evaluated by numerical integration, with the trapezoidal rule, for each γ.

Fig. (4) shows the estimates $\widehat{TE}_N(X \to Y)$ medians, with the approximate theoretical bounds.

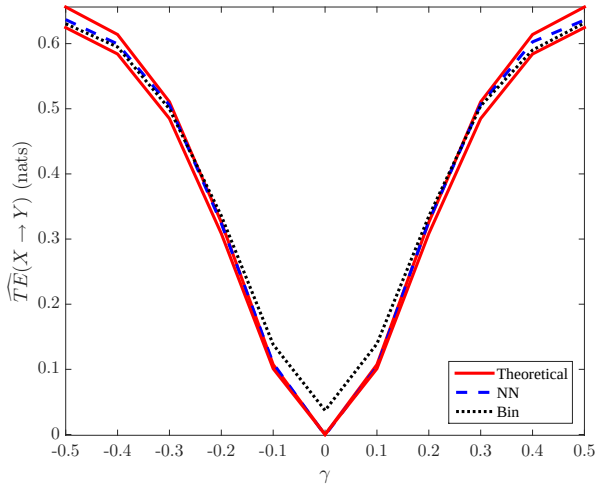


Fig. 4: Medians of TE estimates as a function of the causality coupling parameter γ. Dashed blue curve indicates medians of NN estimates, dotted black curve indicates medians of binning estimates, and red solid curves indicate theoretical bounds, for each γ. Statistics evaluated over 50 trials, each trial with duration of $N = 1000$.

We can see from Fig. 4 that the binning method overestimates TE when the coupling parameter absolute value was low ($|\gamma| = \{0, 0.1, 0.2\}$), while NN estimates are mainly within theoretical bounds. Both methods are within theoretical bounds

for larger $|\gamma|$ values. We also estimated $\widehat{TE}(X \to Y)$ with the fixed coupling parameter $\gamma = 0.5$, for different duration of the processes ($N = \{50, 100, 500, 1000, 5000\}$). Fig. 5 shows that the estimates converge to the same value as N increases, and that value is within the theoretical bounds.

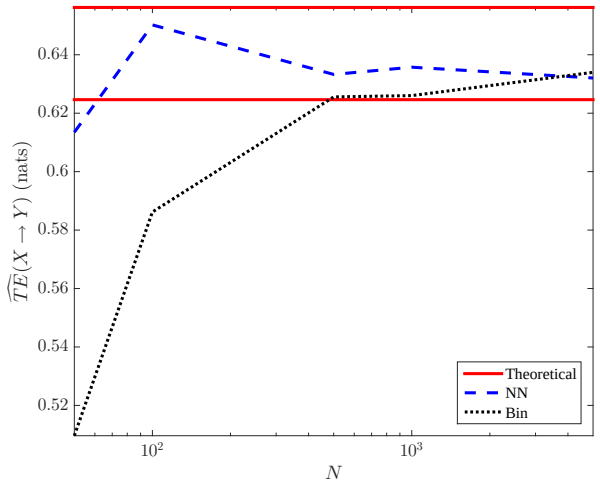


Fig. 5: Medians of TE estimates as a function of the process duration N. Dashed blue curve indicates median values of NN estimates, dotted black curve indicates binning estimates, for each duration N, and red solid lines indicate theoretical bounds. Statistics evaluated over 50 trials, parameter $\gamma = 0.5$.

Fig. 6 illustrates the performance of the estimators according to sample variance of the estimates, for the case of $\gamma = 0.5$, varying duration N. Binning estimates present smaller variance in all cases, showing improved performance over NN method in this criterion. However, we see in both cases that variance diminishes as N increases.

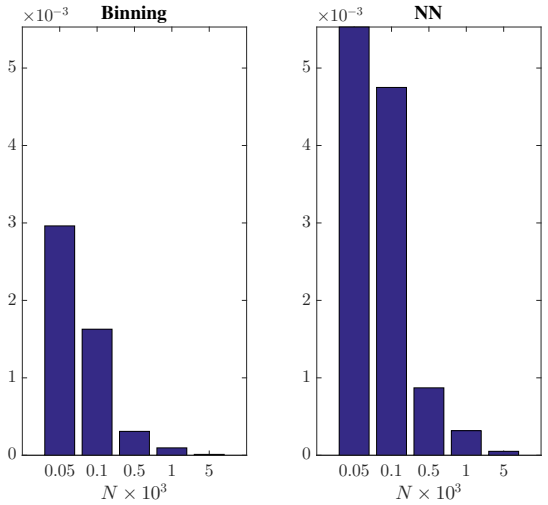


Fig. 6: Sample variances of TE estimates as a function of the process duration N. Statistics evaluated over 50 trials, parameter $\gamma = 0.5$.

B. Second Example: TE from Continuous to Discrete

This example illustrates that it is also possible to estimate TE from a continuous process to a discrete process.

Consider $\mathbf{X}$ an i.i.d. process such that each X_n is uniformly distributed in the range $[\alpha, \beta]$ ($X_n \sim U(\alpha, \beta)$, $0 < \alpha < \beta$). Consider the process $\mathbf{Y}$ which is causally influenced by $\mathbf{X}$ as follows:

$$P(Y_n = y|X_{n-1} = x) = \frac{(x)^y e^{-x}}{y!}. \tag{30}$$

Analogously to the example of the last subsection, we may evaluate TE for this example, considering $\ell = m = 1$, and using identity (20). $P(Y_n = y)$ can be evaluated as:

$$\begin{aligned} P(Y_n = y) &= \int_\alpha^\beta P(Y_n = y|X_{n-1}) f_{X_{n-1}}(x) dx \\ &= \int_\alpha^\beta \frac{x^y e^{-x}}{y!} \frac{1}{\beta - \alpha} dx \\ &= \frac{1}{y!(\beta - \alpha)} \int_\alpha^\beta x^y e^{-x} dx. \end{aligned} \tag{31}$$

Therefore,

$$P(Y_n = y) = \begin{cases} -e^{-x}|_\alpha^\beta = e^{-\alpha} - e^{-\beta}, & \text{if } y = 0, \\ (-x-1)e^{-x}|_\alpha^\beta, & \text{if } y = 1, \\ -x^y e^{-x}|_\alpha^\beta + y \int_\alpha^\beta x^{y-1} e^{-x} dx, & \text{if } y > 1. \end{cases} \tag{32}$$

When $y > 1$, it is possible to evaluate $P(Y_n = y)$ recursively through (32). Fig. 7 illustrates $P(Y_n = y)$, with $\alpha = 25$ and $\beta = 55$ (the choice of these values for those parameters will be explained later). Notice that for $Y_n \geq 100$, $P(Y_n)$ is negligible, in the order of 10^{-9} or less. Thus, it is possible to approximate entropy $H(Y_n|Y_{n-1}) = H(Y_n)$, because Y_n and Y_{n-1} are independent, for every n. The approximation of $H(Y_n)$ was found by considering only the integers from 0 to 100 ($0 \leq Y_n \leq 100$).

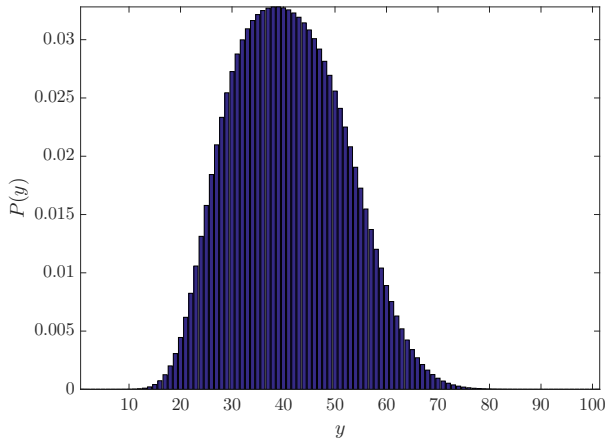


Fig. 7: Probability mass function of Y_n.

In order to find entropy $H(Y_n|Y_{n-1}X_{n-1}) = H(Y_n|X_{n-1})$, consider the fact that Y_n has a Poisson distribution when conditioned to the value of X_{n-1}, with rate $X_{n-1} = x$,

$x > 0$. There are available approximations for the entropy of a Poisson distribution, when its rate is $x > 10$ [29]:

$$H(Y_n|X_{n-1} = x) \approx \frac{1}{2}\ln(2\pi e x) - \frac{1}{12x} + O(x^{-2}). \quad (33)$$

Neglecting the term $O(x^{-2})$, it is possible to reach an approximate value for $H(Y_n|X_{n-1})$:

$$\begin{aligned} H(Y_n|X_{n-1}) &= \int_\alpha^\beta \frac{1}{\beta-\alpha} H(Y_n|X_{n-1}=x)dx \\ &\approx \int_\alpha^\beta \frac{1}{\beta-\alpha}\left(\frac{1}{2}\ln(2\pi e x) - \frac{1}{12x}\right)dx \\ &= \frac{1}{2(\beta-\alpha)}\left(x\ln(2\pi e x) - x - \frac{1}{12}\ln(x)\right)\Big|_\alpha^\beta \end{aligned} \quad (34)$$

By selecting the values for $\alpha = 25$ and $\beta = 55$, for instance, which guarantee the condition $x > 10$ used in the approximation (34), we find

$$TE_\infty(X \to Y) \approx 0.589 \text{ nats}.$$

The estimates for this example are presented in Fig. 8, which shows that the NN method converges to the approximation of the theoretical value as $N \to \infty$, while the binning method diverges. We observe some negative estimates for TE with the NN method, particularly for smaller duration N. This is an undesirable result, since TE is a KL distance. However, negative estimates for mutual information using NN method are reported in reference [24], due to systematic errors (mutual information is also a KL distance). Since we use the identity of a sum of mutual information estimates to evaluate TE, these negative results may be obtained.

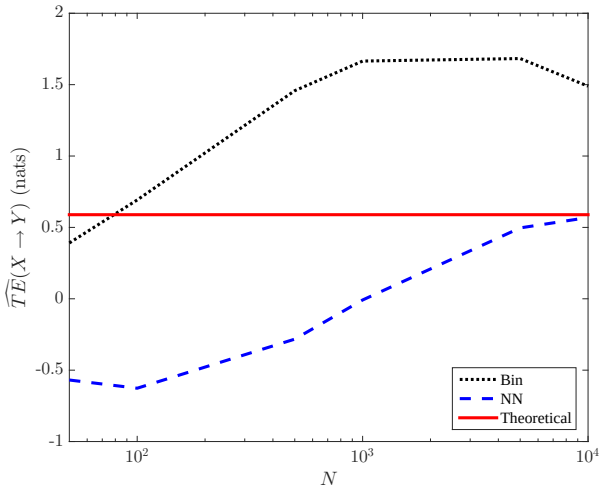


Fig. 8: Medians of TE estimates of a continuous process to a discrete process with the binning method (Bin) and the NN method (NN), as a function of N, in 50 trials. Approximation of theoretical value in continuous red line.

The sample variances obtained in this example are indicated in Fig. 9. Again, the binning method has improved performance over the NN method in the criterion of presenting reduced variances.

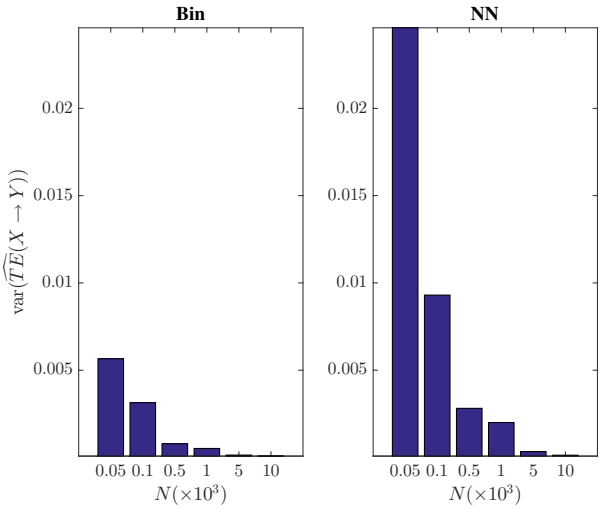


Fig. 9: Sample variances of the TE estimates from a continuous process to a discrete process with the binning and the NN method, as a function of the duration N of the processes, in 50 trials.

We also built a similar example of a continuous process causally influencing a discrete process, but such that $\mathbf{X}$ is an autoregressive process given by:

$$X_n = \alpha X_{n-1} + \eta_n, \quad (35)$$

where η_n is a standard Gaussian random variable. Analogously to the first example of this subsection, $\mathbf{Y}$ is a discrete random process. For each time index n, the probability distribution of the random variable Y_n is given by

$$P(Y_n = y|X_{n-1} = x) = \frac{|x|^y}{y!}e^{-|x|}, \quad (36)$$

which is a Poisson distribution whose rate is given by a past value of the process $\mathbf{X}$.

In our simulations of this example, we set $\alpha = 0.5$. In each of 50 trials of the experiment, we evaluated four TE estimations: two of them were $\widehat{TE}(X \to Y)$ with both binning and NN methods. The other two were $\widehat{TE}(X \to Y_{test})$ estimates built without any dependency or causality relation between $\mathbf{X}$ and $\mathbf{Y}_{test}$, with both estimation methods. More specifically, $\mathbf{Y}_{test}$ was generated by an i.i.d. Gaussian process $\mathbf{Z}$, $Z_n \sim \mathcal{N}(0,1)$, in the same manner as in equation (36), by substituting x_{n-1} by z_{n-1}. The idea was to compare the estimation methods in a scenario where there was causality with a scenario where causality was absolutely absent (but with processes with similar statistics). Fig. 10 shows the results.

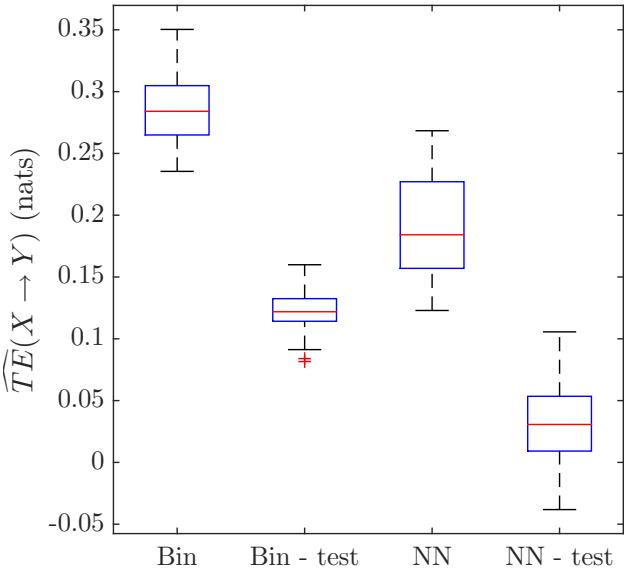


Fig. 10: TE estimates boxplots with the binning method (Bin) over data, with the binning method over processes without causality (Bin - test), with the NN method and with the NN method over processes without causality (NN - test). Statistics evaluated over 50 trials, duration of processes $N = 500$.

It is clear from Fig. 10 that there is a significant difference among estimates from original processes and estimates from processes with no causality, which was guaranteed with a t-test (level of confidence set in 5%). This means that, despite not converging to same median value, both estimation methods indicate faithfully a difference when there was a causality relation. Moreover, Fig. 11 indicates that the estimates converge as the duration of the processes increases ($N \to \infty$).

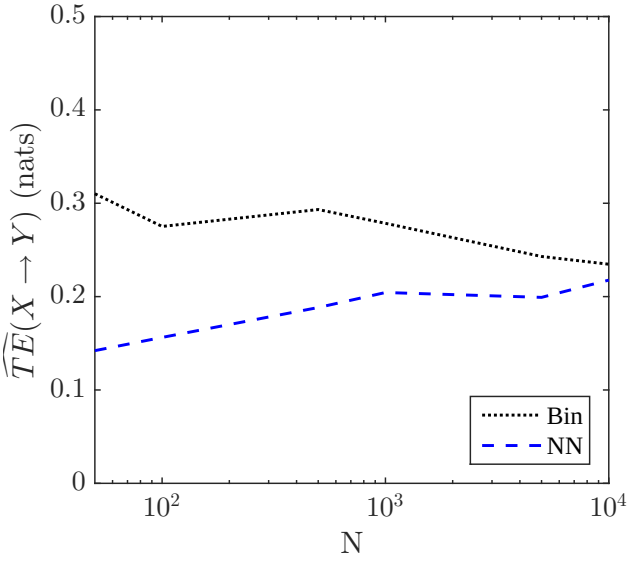
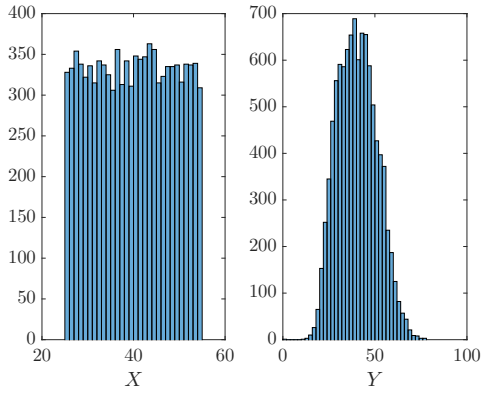
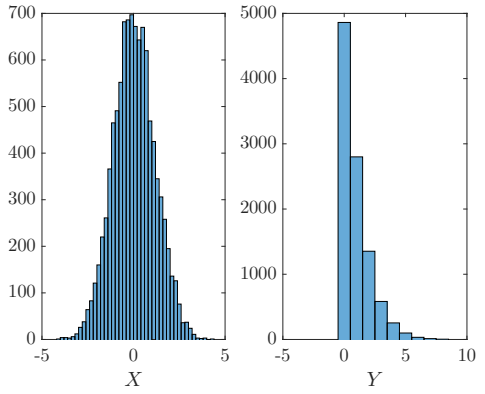


Fig. 11: TE estimates with the binning method and NN method as a function of N.

Interestingly, in the first example of this subsection, the binning method does not converge to the approximation of the theoretical value, while the NN method converges. However, in the second example of this subsection, both methods converge to the same value as N increases. This fact may happen because, in the first example, the rate $X_{n-1} = x$ of the Poisson process varies in a large interval. Thus, the values assumed by $\mathbf{Y}$ in this example vary in a larger interval than in the second example. Even though the alphabet of $\mathbf{Y}$ in both examples is (countably) infinite theoretically, in practice, the realizations of $\mathbf{Y}$ in the second example present a reduced alphabet than the realizations of $\mathbf{Y}$ in the first example. Notice that the binning method uses the plug-in method. Moreover, the bias of mutual information for the plug-in method can be approximated as a function of the alphabet of $\mathbf{Y}$ [30], [31]. This may be the reason for the poor performance of the binning method in the first example of this subsection. Fig. 12 illustrates histograms of $\mathbf{X}$ and $\mathbf{Y}$, from the first and the second examples of this subsection, in one trial of the processes with duration $N = 10000$.



(a) $\mathbf{X}$ uniform



(b) $\mathbf{X}$ autoregressive

Fig. 12: Histograms of the values assumed by processes $\mathbf{X}$ and $\mathbf{Y}$ of this subsection, according to the example where $\mathbf{X}$ is uniformly distributed and the example where $\mathbf{X}$ is autoregressive. Duration of the processes $N = 10000$.

C. Third example: Different Coupling Time

In this section we consider the effect of a different coupling time of processes $\mathbf{X}$ and $\mathbf{Y}$ in the proposed estimation methods.

In order to do so, we generated $\mathbf{X}$ as the Markov process in reference [32], whose state diagram is shown in Fig. 13. The transition probabilities in the state diagram are $\theta_1 = P(X_n = 1|X_{n-1} = 1)$, $\theta_{10} = P(X_n = 1|X_{n-2}^{n-1} = 10)$ and $\theta_{00} = P(X_n = 1|X_{n-2}^{n-1} = 00)$.

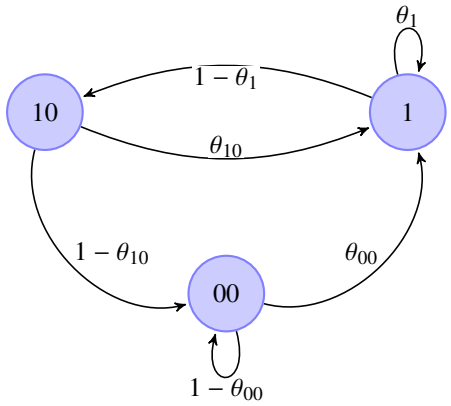


Fig. 13: State diagram for process $\mathbf{X}$ in the example that considers different coupling time.

For each state of process $\mathbf{X}$, which are 1, 10 or 00, we associate another discrete random process $\mathbf{Z}$, such that

$$Z_n = \begin{cases} 1, & \text{if } X_{n-1} = 1, \\ 0, & \text{if } X_{n-2}^{n-1} = 10, \\ -1, & \text{if } X_{n-2}^{n-1} = 00. \end{cases} \tag{37}$$

Then, we defined the process $\mathbf{Y}$ as follows:

$$Y_n = \alpha Y_{n-m} + \gamma Z_n + \epsilon \eta_n. \tag{38}$$

In this example, we evaluated TE estimates of

$$TE_n(X \to Y) \quad = \quad H(Y_n|Y_{n-m}^{n-1}) - H(Y_n|Y_{n-m}^{n-1} X_{n-2}^{n-1})$$

considering the past index $\ell = 2$ and varying past index m. The past index m used for TE estimation was the same coupling time m used for simulating process $\mathbf{Y}$, in equation (38). Also, we set the parameters $\alpha = 0.5$, $\gamma = 0.5$, $\epsilon = 0.1$, and the conditioned transition probabilities $\theta_1 = 0.1$, $\theta_{10} = 0.3$ and $\theta_{00} = 0.5$.

Fig. 14 and Fig. 15 reveal the influence of m in TE estimates (Fig. 14 indicates the results through medians of the estimates and Fig. 15 indicates the results through boxplots). We also generated a process $\mathbf{Y}_{test}$ which was not causally influenced by $\mathbf{X}$, in order to compare the results (just like in the second example of subsection V-B). More specifically, process $\mathbf{Y}_{test}$ for testing was generated with the same equation (38), but with $\mathbf{Z}$ as an i.i.d. process, discrete and uniformly distributed in alphabet $\{-1, 0, 1\}$.

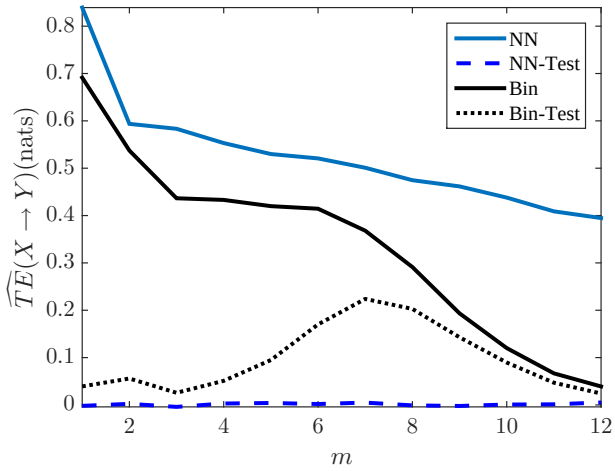


Fig. 14: Estimates $\widehat{TE}(X \to Y)$ medians as a function of coupling time/past index m. Continuous blue curve indicate NN estimates medians, dashed blue curves indicate NN estimates medians over data with no causal relation, continuous black curve indicates binning estimates medians and dotted black curve indicates binning estimates medians over data with no causal relation. Statistics evaluated over 50 trials, duration of processes $N = 500$.

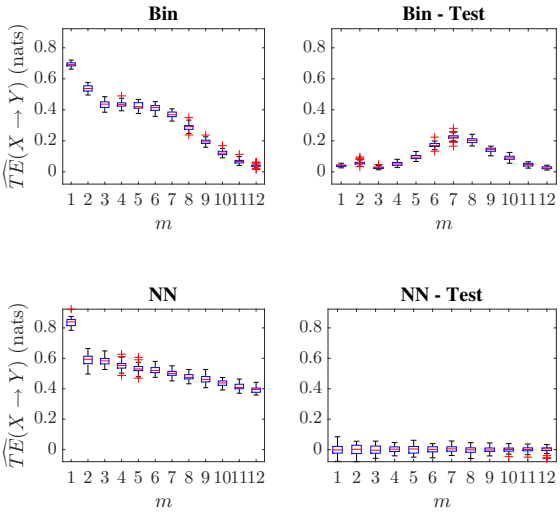


Fig. 15: Boxplots of the estimates $\widehat{TE}(X \to Y)$, in nats, as a function of coupling time/past index m, with the binning method (Bin), with the binning method over processes without causality (Bin - Test), with the NN method and with NN method over processes without causality (NN - Test). Statistics evaluated over 50 trials, duration of processes $N = 500$.

We can see from Fig. 14 and Fig. 15 that the NN method detected no causality in the tested case of absent causality. NN estimates for $\widehat{TE}(X \to Y_{test})$ were mainly null. However, binning estimates for $\widehat{TE}(X \to Y_{test})$ are always greater than zero, indicating a spurious causality detection. Moreover, there is a more pronounced difference between $\widehat{TE}(X \to Y)$ and

$\widehat{TE}(X \rightarrow Y_{test})$ with the NN method than with the binning method, especially for $m \geq 6$. Thus, NN method presented a more distinct difference to the tested case of absent causality. However, when executing a t-test, both methods presented a significant difference between estimates $\widehat{TE}(X \rightarrow Y)$ and $\widehat{TE}(X \rightarrow Y_{test})$, for each m used (level of confidence set in 5%). Additionally, we can see that with both methods, as m increases, it becomes more difficult to notice the causal influence of $\mathbf{X}$ over $\mathbf{Y}$ (the estimates diminish, even though there is still an underlying causal relation). This happens because the dimension of the variable Y_{n-m}^{n-1} in the estimation also increases, while keeping the same sample size (duration of processes, $N = 500$).

D. Speed Performance

We registered the estimation time for TE between the processes of the first example (subsection V-A) as a function of the duration N of the processes. The time with the binning method was always less than 0.1s. Fig. 16 indicates median time took by NN method normalized by median time took by the binning method, in 50 realizations of the processes. It is clear that the NN method is more time consuming, as mentioned in reference [19].

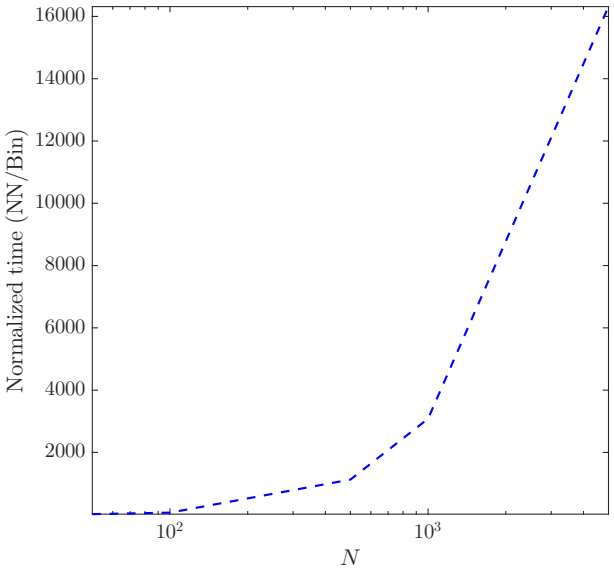


Fig. 16: Estimation median time of NN method normalized by the estimation median time of the binning method, in 50 realizations of the processes of subsection V-A, as a function of the duration N of the processes.

VI. Conclusion

In this paper we investigated the use of two estimators of transfer entropy in mixed cases, that is, in cases when one process is continuous and the other is discrete. The two estimators use the identity of transfer entropies as a sum of two mutual information terms, assuming ergodicity of the analysed processes.

We evaluated estimates in situations with TE approximate tight bounds, or even with an approximation of the theoretical TE value. In those situations, the binning method yielded positively biased results when there was actually no causality. This method was within theoretical bounds for TE when there was actually a causal relation and the alphabet size of the discrete random process is small (for instance, alphabet size equals 2) with a relatively small duration of the processes ($N = 500$). However, when the size of the alphabet of the discrete process is large (for example, alphabet size equals 40) we observed that the binning method did not converge to the theoretical value for TE, for a large duration of the processes ($N = 10000$). On the other hand, the NN method yields results that converge to (or are within) the theoretical approximation (or bounds) in the those situations.

We also evaluated estimates in situations without approximate TE bounds or approximate theoretical TE value. In those situations, we compared the results with test cases of processes with no underlying causality (but with similar statistics). A t-test could detect a difference between the estimates from processes with an underlying causal relation and estimates from processes without this underlying causality, with both methods. This difference was also detected when we increased the coupling time among the analysed processes (and the past index of TE). However, in this case there was a more pronounced difference with the NN method.

Thus, we conclude in this paper that, in the overall scenario, NN method estimates TE more reliably than the binning method, in this particular application of detecting causality between a discrete random process and a continuous random process. In other words, NN method detects less false positives than the binning method. The binning method presents the advantage of faster performance, though. Therefore, the NN method achieves an improved performance over the binning method, at the expense of a higher complexity.

Acknowledgment

This work was partially supported from the Brazilian funding agencies CNPq (30539/2009.9), and COPELE-UFCG.

References

[1] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000, doi: 10.1103/PhysRevLett.85.461.

[2] P. Duan, F. Yang, S. L. Shah, and T. Chen, "Transfer zero-entropy and its application for capturing cause and effect relationship between variables," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 3, pp. 855–867, 2015, doi: 10.1109/TCST.2014.2345095.

[3] S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs, "Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model," *PLoS one*, vol. 6, no. 11, p. e27431, 2011, doi: 10.1371/journal.pone.0027431.

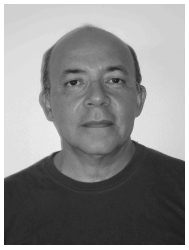
[4] M. Staniek and K. Lehnertz, "Symbolic transfer entropy," *Physical Review Letters*, vol. 100, no. 15, p. 158101, 2008, doi: 10.1103/PhysRevLett.100.158101.

[5] B. Gourévitch and J. J. Eggermont, "Evaluating information transfer between auditory cortical neurons," *Journal of Neurophysiology*, vol. 97, no. 3, pp. 2533–2543, 2007, doi: 10.1152/jn.01106.2006.

[6] O. Stetter, D. Battaglia, J. Soriano, and T. Geisel, "Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals," *PLoS Comput Biol*, vol. 8, no. 8, p. e1002653, 2012, doi: 10.1371/journal.pcbi.1002653.

[7] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy - a model-free measure of effective connectivity for the neurosciences," *J Comput Neurosci*, vol. 30, pp. 45–67, 2011, doi: 10.1007/s10827-010-0262-3.
[8] M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente, "Measuring information-transfer delays," *PLoS one*, vol. 8, no. 2, p. e55809, 2013, doi: 10.1371/journal.pone.0055809.
[9] J. Runge, M. Riedl, A. Müller, H. Stepan, J. Kurths, and N. Wessel, "Quantifying the causal strength of multivariate cardiovascular couplings with momentary information transfer," *Physiological measurement*, vol. 36, no. 4, p. 813, 2015, doi: 10.1088/0967-3334/36/4/813.
[10] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, "Escaping the curse of dimensionality in estimating multivariate transfer entropy," *Physical review letters*, vol. 108, no. 25, p. 258701, 2012, doi: 10.1103/PhysRevLett.108.258701.
[11] P.-O. Amblard and O. J. Michel, "On directed information theory and granger causality graphs," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 7–16, 2011, doi: 10.1007/s10827-010-0231-x.
[12] Y. Liu and S. Aviyente, "The relationship between transfer entropy and directed information," in *Statistical Signal Processing Workshop (SSP), 2012 IEEE*. IEEE, 2012, pp. 73–76, doi: 10.1109/SSP.2012.6319809.
[13] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *J Comput Neurosci*, vol. 30, pp. 17–44, 2011, doi: 10.1007/s10827-010-0247-2.
[14] R. Malladi, G. Kalamangalam, N. Tandon, and B. Aazhang, "Identifying seizure onset zone from the causal connectivity inferred using directed information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 7, pp. 1267–1283, 2016, doi: 10.1109/JSTSP.2016.2601485.
[15] Y. Murin, J. Kim, and A. Goldsmith, "Tracking epileptic seizure activity via information theoretic graphs," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 583–587, doi: 10.1109/ACSSC.2016.7869109.
[16] J. Jiao, H. H.Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Transactions on Information Theroy*, vol. 59, no. 10, pp. 6220–6242, 2013, doi: 10.1109/TIT.2013.2267934.
[17] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in poisson channels," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1837–1849, 2008, doi: 10.1109/TIT.2008.920206.
[18] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, 1990, pp. 303–305.
[19] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLoS One*, vol. 9, no. e87357, 2014, doi: 10.1371/journal.pone.0087357.
[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.
[21] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007, doi: 10.1016/j.physrep.2006.12.004.
[22] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064–3074, 2005, doi: 10.1109/TIT.2005.853314.
[23] M. Paluš, "Testing for nonlinearity using redundancies: Quantitative and qualitative aspects," *Physica D: Nonlinear Phenomena*, vol. 80, no. 1, pp. 186–205, 1995, doi: 10.1016/0167-2789(95)90079-9.
[24] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review*, vol. 69, 2004, doi: 10.1103/PhysRevE.69.066138.
[25] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.
[26] J. M. de Assis, M. O. Santos, and F. M. de Assis, "Auditory stimuli coding by postsynaptic potential and local field potential features," *PLoS One*, vol. 11, no. 8, 2016, doi: 10.1371/journal.pone.0160089.
[27] A. Kaiser and T. Schreiber, "Information transfer in continuous processes," *Physica D: Nonlinear Phenomena*, vol. 166, no. 1, pp. 43–62, 2002, doi: 10.1016/S0167-2789(02)00432-3.
[28] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, "On entropy approximation for gaussian mixture random vectors," in *Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*. IEEE, 2008, pp. 181–188, doi: 10.1109/MFI.2008.4648062.
[29] M. S. Klamkin, Ed., *Problems in applied mathematics: selections from SIAM review*. Society for Industrial and Applied Mathematics, 1990.
[30] G. A. Miller, "Note on the bias of information estimates," *Information theory in psychology: Problems and methods*, vol. 2, no. 95, p. 100, 1955.
[31] R. A. Ince, R. Senatore, E. Arabzadeh, F. Montani, M. E. Diamond, and S. Panzeri, "Information-theoretic methods for studying population codes," *Neural Networks*, vol. 23, no. 6, pp. 713–727, 2010, doi: 10.1016/j.neunet.2010.05.008.
[32] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995, doi: 10.1109/18.382012.

Juliana Martins de Assis was born in Brasília, Distrito Federal, Brazil, in 1988. She received the B.Sc. degree in electrical engineering from the Federal University of Campina Grande, Paraíba, Brazil, in 2012. She received her M. Sc. in neuroscience from the Federal University of Rio Grande do Norte, Rio Grande do Norte, Brazil, in 2014, and her Ph.D degree in electrical engineering from the Federal University of Campina Grande, Paraíba, Brazil, in 2017. Her research interests include estimation of information measures and its applications to neuroscience.



Francisco Marcos de Assis was born in João Pessoa, Paraíba, Brazil, in 1954. He received the B.Sc. and M.Sc. degrees in electrical engineering from the Military Institute of Engineering, Rio de Janeiro, Brazil, in 1984 and 1992, respectively, and the Ph.D. degree in electrical engineering from Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, in 1994. He is currently a Professor with the Federal University of Campina Grande, Paraíba, Brazil. His research interests include coding and information theory.