Estimation of Directed Information to Processes Assuming Continuous Values with CTW Algorithm

Juliana M. de Assis and Francisco M. de Assis

Abstract— **This paper introduces the use of a directed information estimator for discrete-valued processes to continuous-valued processes by previously discretizing the continuous-valued processes according to three different and generally applicable methods. The directed information estimator uses context tree weighting algorithm (CTW). The three discretization methods are called equidistant, equipopulated and symbolic methods. Simulated results indicate that faster and more conservative results are found by using few discretization levels with equipopulated method.**

Keywords— **Directed information, Causality, Estimation, Discretization.**

I. Introduction

Introduced by Massey and Marko [1] [2], directed information (DI) is a model-free measure of causality. It has been recently used in different fields to estimate causal relationships. DI estimation revealed influences in stock markets in economy [3]. It also has been used to determine neuronal connections [4] [5] and to determine epileptic onset zone [6] in neuroscience. DI has been used to determine internet access influences [7] and to gene networks influences [8], among other examples in systems presenting causal links.

Despite the fact that DI correctly identifies causal relationships, DI estimation is not trivial. Initially, DI was well established among processes assuming discrete values, and it is estimated for finite-alphabet stationary ergodic processes, either parametrically as in [5] or using the celebrated context tree weighting algorithm (CTW) [9] as in [3], with Jiao estimators.

Malladi *et al.* present extensions of DI definition from discrete-time, discrete-valued processes to discrete-time, continuous-valued processes [6]. Malladi *et al.* also introduce almost surely convergent estimators for DI between stationary ergodic continuous-valued Markovian processes, using either a model-based approach or a data-driven approach.

In this article, we investigate the use of one of Jiao estimators to estimate DI between stationary continuous-valued processes by previously discretizing the continuous values. The discretization is performed and analyzed with three different methods. The paper is organized as follows. Section II establishes notation and terminology. Section III introduces causality and DI definitions. Section IV introduces briefly CTW algorithm as also the chosen Jiao estimator used in the simulations performed here. Section IV also presents the three methods used to discretize the continuous data. Section V presents the performed simulations and comparisons among the three methods. 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. Subscripts usually denote the outcome's position in a sequence, for example, X_n generally indicates the nth output of the process $\boldsymbol{X}$. Subscripts also denote the first index considered in a sequence. Superscripts on a random variable denote finite length sequences of this random variable, for example, $X_2^N = \{X_2, X_3, \ldots, X_N\}$. Throughout this paper, log is base 2, and $\mathbb{E}(X)$ indicates the expected value of X.

III. Directed Information

One important concept used in this article is the concept of causality. The sense of causality throughout this paper is the same given by Norbert Wiener [10], according to which a random process $\mathbf{X}$ causes a random process $\mathbf{Y}$ if the knowledge of the past of $\mathbf{X}$ improves the prediction of future values of $\mathbf{Y}$.

DI between finite continuous-valued time series X^N and Y^N is defined as [6]:

$$I(X^N \to Y^N) = h(Y^N) - h(Y^N||X^N) \tag{1}$$

where

$$h(Y^N) = \sum_{n=1}^{N} h(Y_n|Y^{n-1}) \tag{2}$$

is differential entropy of the sequence Y^N [11] and

$$h(Y^N||X^N) = \sum_{n=1}^{N} h(Y_n|Y^{n-1}X^n) \tag{3}$$

is the causally conditioned differential entropy of sequence Y^N causally conditioned on sequence X^N.

When dealing with time series or stochastic processes, one is usually interested in how DI increases as the sequence grows, or in its rate - directed information per letter. When we mention finite time series, the DI rate is defined as [12]:

$$I_N(X \to Y) = \frac{1}{N} I(X^N \to Y^N). \tag{4}$$

Moreover, when we mention stochastic processes, the DI rate is defined in the limit as $N \to \infty$, when the limit exists:

$$I_\infty(X \to Y) = \lim_{N\to\infty} I_N(X \to Y). \tag{5}$$

Juliana M. de Assis and Francisco M. de Assis, Department of Electrical Engineering, Federal University of Campina Grande, Campina Grande -PB, Brazil, E-mails: juliana.assis@ee.ufcg.edu.br, fmarcos@dee.ufcg.edu.br. This work was partially supported by CNPq.

DI exhibits some properties when the involved processes are discrete-valued, which can be extended to the case when the involved processes are continuous-valued. One property is DI is always non-negative [6]. Another property is that, differently than mutual information, DI is generally not symmetric, that is, $I(X^N \to Y^N) \neq I(Y^N \to X^N)$ in general. This constitutes one important characteristic, since causality is usually directed, and requires also a directed measure.

Here we also present a property derived for discrete-valued processes which can easily be extended to continuous-valued processes: DI is upper bounded by mutual information [12]. To observe this fact, we write the definition of mutual information between two finite time series X^N and Y^N and we rewrite DI definition to explicit the apparently slight difference among them two:

$$I(X^N; Y^N) = \sum_{n=1}^{N} [h(Y^n|Y^{n-1}) - h(Y^n|Y^{n-1}X^N)] \tag{6}$$

$$I(X^N \to Y^N) = \sum_{n=1}^{N} [h(Y^n|Y^{n-1}) - h(Y^n|Y^{n-1}X^n)] \tag{7}$$

which is the superscript N or n on the X conditioning in the second entropy term inside the sum. Since conditioning always reduces entropy (even in the differential case [11]), $h(Y^n|Y^{n-1}X^N)$ is always less than or equal to $h(Y^n|Y^{n-1}X^n)$, then $I(X^N; Y^N)$ is always greater than or equal to $I(X^N \to Y^N)$.

IV. Estimation of Directed Information

In this section, we present the DI estimators analyzed in this paper for continuous processes. These estimators have essentially two major steps in their application: firstly they discretize the data and secondly they apply one of Jiao estimators [3] to the discretized processes. Jiao estimators estimate DI between two finite-alphabet stationary ergodic processes and use CTW algorithm. In the following subsection, we describe the three proposed methods of discretization.

A. Discretization Methods

Discretization methods are often used in the estimation of mutual information between two continuous-valued random variables [13]. Even though the discretization may lead to biased estimates [14], they are very appealing due to its simplicity.

As already mentioned, in this paper we analyze three methods of discretization. The first method consists in segmenting the support of the processes in L equidistant segments, according to Euclidean distance, from the minimum value to maximum value. The second method also segments the support of the processes, but in L equipopulated segments. The third method is inspired in reference [15]. It consists in ordering k consecutive values of the time series, representing this transition order by one number that represents one of $k!$ possible permutations. For example, consider the time series:

$$X^7 = [0.5\ 0.75\ -0.1\ -0.23\ 0.05\ 0.52\ 0.49]$$

TABLE I
Permutation values

Transition	Discretized value
012	1
021	2
102	3
120	4
201	5
210	6

If we choose $k = 2$, we have the corresponding transitions:

$$[01\ 10\ 10\ 01\ 01\ 10].$$

Labelling the symbol 01 as 1 and the symbol 10 as 2, we obtain the following discretized sequence (beginning in index $n = 2$):

$$\tilde{X}_2^7 = [1\ 2\ 2\ 1\ 1\ 2].$$

On the other hand, if we choose $k = 3$ we have the corresponding transitions:

$$[120\ 210\ 102\ 012\ 021].$$

Labelling each transition as in table I, we obtain the following discretized sequence (beginning in index $n = 3$):

$$\tilde{X}_3^7 = [4\ 6\ 3\ 1\ 2].$$

Thus, in this third discretization method, if we choose parameter k, we discretize the continuous process in $L = k!$ symbols.

B. Jiao Estimator

As mentioned in section I, Jiao estimators use the CTW algorithm. So, in order to understand Jiao estimators, we briefly explain CTW algorithm. CTW assumes that we have a finite memory tree source to build a context tree with depth D, here considered larger than the memory of the source. The context tree has nodes, each one labelled by the string s, which is also a context, of at most D symbols, with the counts of how many times s preceded each of the symbols. The next procedure is to attribute weighted probabilities to the nodes based on these counts. The weighted probabilities P_w are based on the Krichevsky-Trofimov (KT) estimated probabilities, P_e. There is a sequential formula to compute P_e for each node of the context tree of a source emitting M symbols, where b_i is the counting of symbol i, $i \in \{0, \ldots, M-1\}$ [9], [3], [4]:

$$\begin{aligned} P_e(b_0, b_1, \ldots b_{i-1}, b_i + 1, b_{i+1}, \ldots, b_{M-1}) &= \\ \frac{b_i + 1/2}{b_0 + \cdots + b_i + \cdots + b_{M-1} + M/2} &\times \\ P_e(b_0, b_1, \ldots b_{i-1}, b_i, b_{i+1}, \ldots, b_{M-1}). \end{aligned}$$

The root node is denoted by λ and P_w^λ is the universal probability assignment by CTW. References [9], [3], [4] show examples of CTW implementation applied to sequences.

Jiao *et al.* [3] present four slightly different estimators of DI based in CTW algorithm. All four of them present almost sure consistency under some assumptions of the underlying probability distribution of the involved stochastic processes. The first two of them present the advantage of established rates of convergence. The other two of them present the advantage of always being non-negative. In this paper, we chose to use one of these last two algorithms (“E_4” estimator).

V. Results

In this section we perform the DI calculation of one baseline example where we have a stochastic continuous-valued process $\mathbf{X}$ causing another stochastic continuous-valued process $\mathbf{Y}$. Additionally, we perform many subsequent simulations using the three proposed discretization methods and the chosen DI Jiao estimator.

A. Baseline Example

The “driving” process $\mathbf{X}$ is an i.i.d. Gaussian stochastic process, with zero mean and unit variance (denoted by σ_X^2), while the “response” process $\mathbf{Y}$ is given according to the equation:

$$Y_n = \beta X_{n-2} + Z_n \tag{8}$$

where β is a coupling parameter and Z_n is also an i.i.d. Gaussian stochastic process with zero mean and unit variance (denoted by σ_Z^2). We may evaluate the true DI rate value in this baseline example by evaluating the terms $h(Y^N||X^N)$ and $h(Y^N)$ separately:

$$\begin{aligned}\frac{1}{N}h(Y^N||X^N) &= \frac{1}{N}\sum_{n=1}^{N} h(Y_n|Y^{n-1}X^n)\\ &= \frac{1}{N}\sum_{n=1}^{N} h(\beta X_{n-2}+Z_n|Y^{n-1}X^n)\\ &= \frac{1}{N}\sum_{n=1}^{N} h(\beta X_{n-2}+Z_n|X^{n-2})\\ &= \frac{1}{N}\sum_{n=1}^{N} h(Z_n)\\ &= \frac{1}{2}\log(2\pi e\sigma_Z^2)\\ &= \frac{1}{2}\log(2\pi e),\end{aligned}$$

and

$$\begin{aligned}\frac{1}{N}h(Y^N) &= \frac{1}{N}\sum_{n=1}^{N} h(Y_n|Y^{n-1})\\ &= h(Y_n),\end{aligned}$$

because $\mathbf{Y}$ does not depend on its own past.

In order to compute $h(Y_n)$, since its an zero mean Gaussian random variable, we need to evaluate its variance:

$$\begin{aligned}\operatorname{var}(Y_n) &= \mathbb{E}(\beta X_{n-2}+Z_n)^2\\ &= \mathbb{E}(\beta^2X_{n-2}^2+2\beta X_{n-2}Z_n+Z_n^2)\\ &= \beta^2\mathbb{E}(X_{n-2}^2)+\sigma_Z^2\\ &= \beta^2\sigma_X^2+\sigma_Z^2=\beta^2+1.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{1}{N}h(Y^N) &= \frac{1}{2}\log(2\pi e(1+\beta^2)), \text{ and}\\ I_N(X\to Y) &= \frac{1}{2}\log(1+\beta^2).\end{aligned}$$

B. Simulation

With the purpose of evaluating the estimation methods proposed, we simulated the continuous-valued stochastic processes. For each case of estimation, that is:

- Equidistant discretization followed by application of Jiao estimator or;
- Equipopulated discretization followed by application of Jiao estimator or;
- Symbolic discretization followed by application of Jiao estimator;

we simulated 50 trials of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ with duration $N = 10^5$, context tree depth parameter $D = 2$, for each parameter $\beta = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For all proposed estimation methods, we selected $L = 2$ discretization levels or $L = 6$ discretization levels. Additionally, we performed simulation with $L = 4$ discretization levels for the equidistant and equipopulated dicretization methods. The symbolic method does not allow a discrezation of $L = 4$ levels because there is no integer k such that $k! = 4$.

Fig. 1, 2 and 3 show the results for $L = 2$. Fig. 4 and 5 show the results for $L = 6$. Fig. 6 and 7 show the remaining results for $L = 4$. In all figures, red dotted lines indicate the analytical DI rate value, while continuous blue lines indicate the median DI rate estimates and dashed blue lines indicate from 10% to 90% of the estimates.

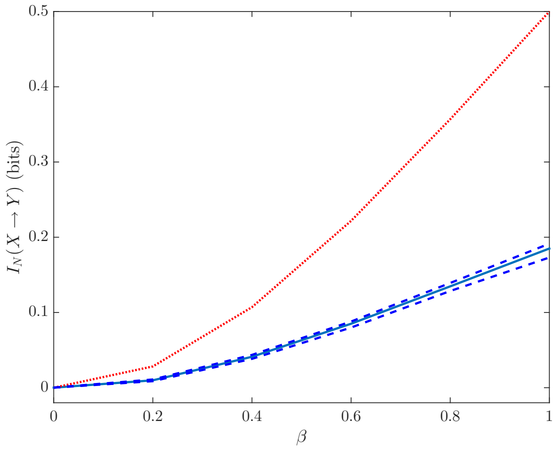

Fig. 1. Equidistant method, $L = 2$

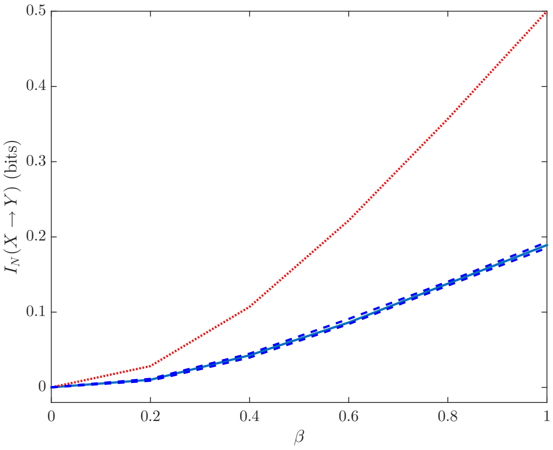

Fig. 2. Equipopulated method, $L = 2$

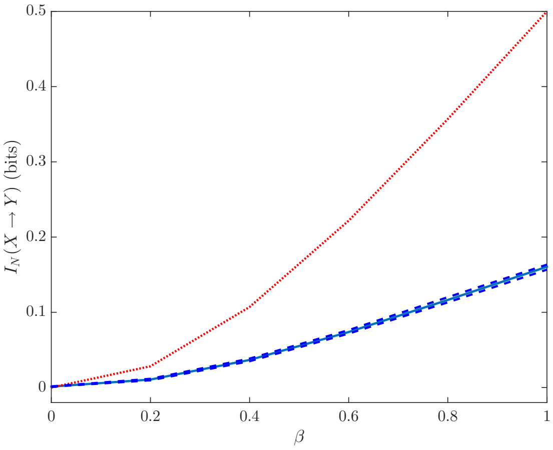

Fig. 3. Symbolic method, $L = 2$

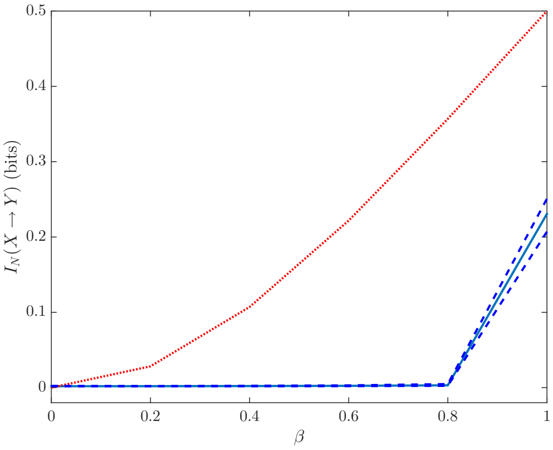

Fig. 4. Equipopulated method, $L = 6$

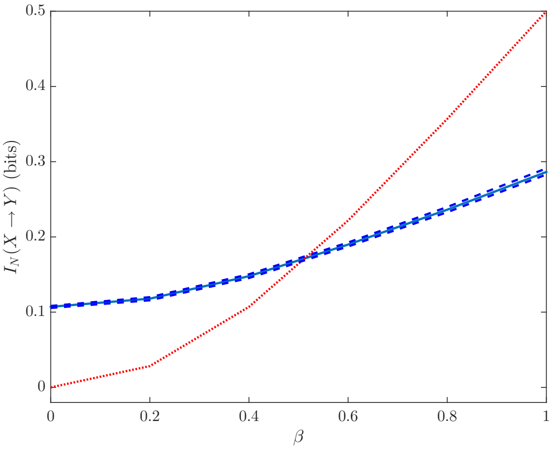

Fig. 5. Symbolic method, $L = 6$

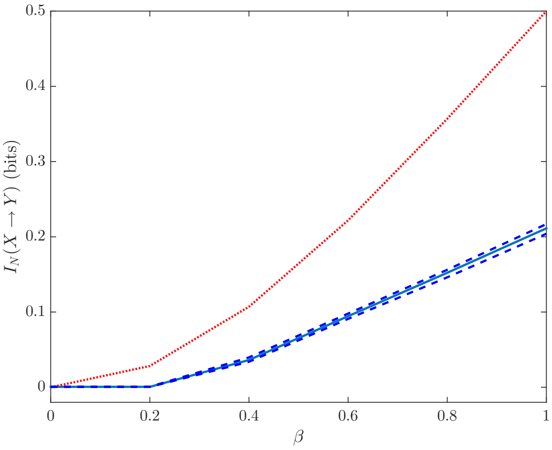

Fig. 6. Equidistant method, $L = 4$

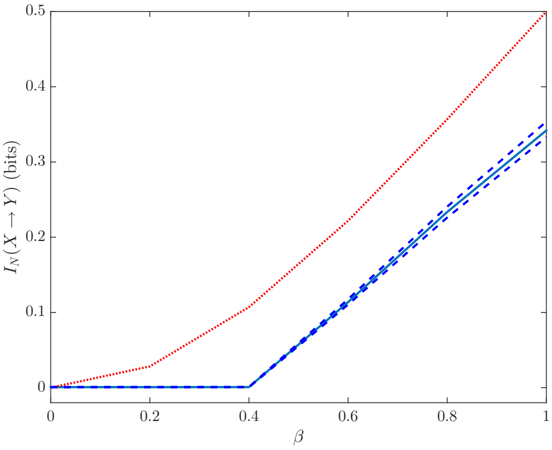

Fig. 7. Equipopulated method, $L = 4$

We highlight some interesting features of the proposed estimation methods. Firstly, concerning the estimates values, we observed that for $L = 2$, all three methods have similar performance. They always underestimate DI values and present very small variance, as shown by dashed lines. On the other hand, for $L = 6$, we observed that equidistant and equipopulated methods do not generally capture the causality for all β parameter values, except for equipopulated method with $\beta = 1$ (results not shown for equidistant method and $L = 6$ due to space limitations, but estimates were approximately 0 for all β values). This may happen because the context tree probabilities become smaller and smaller when using greater alphabet (L). Unfortunately, the symbolic method overestimates DI for small β values while underestimates DI for large β values. However, DI rate estimates in this case follow the behaviour of the analytical DI rate values. For the cases where $L = 4$, we observed that the equipopulated method outperforms the equidistant method, but it did not capture the causality for small β values ($\beta = 0.2, 0.4$) and still underestimates DI, although the underestimation for large β values is less than the underestimation with the equidistant method.

Secondly, concerning the time of the estimation processes, all discretization methods consume little time, with an increment in symbolic method. Nevertheless, the DI rate estimation using Jiao estimator is more time consuming. Each trial takes approximately 35s for $L = 2$ or $L = 4$ levels (alphabet taking 2 or 4 values), while each trial takes approximately 50s for $L = 6$ levels (alphabet taking 6 values). We ran these estimates on a computer with a 2.67GHz processor.

With the view to evaluate the presence of spurious detection of causality between processes that presented no causality relation, we also simulated 50 trials of two independent i.i.d. Gaussian processes with duration $N = 10^5$ and estimated DI rates according to the three discretization methods. In this case, $I_N(X^N \to Y^N) = 0$. Again, we set the parameter of context tree depth $D = 2$, and discretized in $L = 2, 4$ or 6 levels. Table II shows the estimates medians of DI for these simulations

TABLE II

DI RATE ESTIMATE MEDIANS ACCORDING TO THE DISCRETIZATION METHOD AND LEVELS (L), WHEN ANALYTICAL VALUE $I_N(X^N \to Y^N) = 0$.

Discretization method / Levels	$L = 2$	$L = 4$	$L = 6$
Equidistant	$\approx 10^{-5}$	$\approx 10^{-4}$	$\approx 10^{-3}$
Equipopulated	$\approx 10^{-5}$	$\approx 10^{-4}$	$\approx 10^{-3}$
Symbolic	$\approx 10^{-3}$	-	$\approx 10^{-1}$

Apparently, there is a tendency that the estimates medians present a small increase when we use larger alphabets of discretization. However, the only case with rather large DI rate estimates, despite the fact that there was no causality indeed between the processes, was when using symbolic discretization with $L = 6$.

We should also remark that there are many other possible ways to discretize the continuous-valued processes. For instance, in reference [3], stock market causality is estimated by discretizing continuous values in three values. Value -1 indicated that the stock market went down in one day by more than 0.8%, value 1 indicated that stock market went up in one day by more than 0.8%, while value 0 indicated that the absolute change is less than 0.8%. However, we thought that this method has the disadvantage of a necessity to choose an appropriate value for the absolute change (the settled 0.8% in reference [3]). Further works may evaluate a mean to use this method in a proper and general manner.

VI. Conclusion

In this paper we performed simulations to evaluate the DI rate estimation between continuous-valued processes by using an universal DI estimator using CTW for discrete-valued processes (here called Jiao estimator). To perform the estimation, we used previously three different discretization methods. We observed that, for $L = 2$ discretization levels, all three methods perform similarly. For $L = 4$ discretization levels, the equipopulated method outperforms the equidistant method. Finally, for $L = 6$ discretization levels, we may state that the symbolic method best captures the general dynamics as we increased the coupling parameter β. However, in general cases where we desire faster and conservative estimates using Jiao estimator to detect causality between continuous-valued processes, we recommend the use of equipopulated discretization method with a small number of levels ($L = \{2, 3, 4\}$).

Acknowledgments

We would like to thank Jiantao Jiao for gently making his code on estimation of directed information available.

References

[1] J. Massey, “Causality, feedback and directed information,” in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*. Citeseer, 1990, pp. 303–305.

[2] H. Marko, “The bidirectional communication theory–a generalization of information theory,” *Communications, IEEE Transactions on*, vol. 21, no. 12, pp. 1345–1351, 1973.

[3] 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.

[4] J. M. de Assis and F. M. de Assis, “An application of directed information to infer synaptic connectivity,” in *Anais do XXXIV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, Santarém, Brazil, 2016, pp. 528–532.

[5] 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.

[6] 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.

[7] C. J. Quinn, N. Kiyavash, and T. P. Coleman, “Directed information graphs,” *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6887–6909, 2015.

[8] Y. Liu, S. Aviyente, and M. Al-khassaweneh, “A high dimensional directed information estimation using data-dependent partitioning,” in *Statistical Signal Processing, 2009. SSP’09. IEEE/SP 15th Workshop on*. Cardiff, United Kingdom: IEEE, 2009, pp. 606–609.

[9] 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.

[10] N. Wiener, “The theory of prediction,” *Modern mathematics for engineers*, vol. 1, pp. 125–139, 1956.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[12] G. Kramer, “Directed information for channels with feedback,” Ph.D. dissertation, Swiss Federal Institute of Technology, Zurich, 1998.

[13] G. A. Darbellay, I. Vajda *et al.*, “Estimation of the information by an adaptive partitioning of the observation space,” *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[14] 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.

[15] C. Bandt and B. Pompe, “Permutation entropy: a natural complexity measure for time series,” *Physical review letters*, vol. 88, no. 17, p. 174102, 2002.