An Application of Directed Information to Infer Synaptic Connectivity

Juliana M. de Assis and Francisco M. de Assis

***Abstract*— This paper introduces a review on directed information and its application as a causality measure among two stochastic processes. Jiao's method for estimating directed information from data, using the context tree weighting algorithm is described and then used to infer synaptic connectivity from simulated neurons, from their neural spike trains only. It is observed that positive values of directed information estimates correctly predicted synaptic connections.**

I. Introduction

It is well established that information theory answers fundamental questions to communication purposes, as also brings great contributions to computer science, statistics, physics and probability [1]. The familiar concept of mutual information, primarily developed for communication and valuable to find the capacity of channels, has been extensively used with biomedical data [10] and in different scientific fields, such as neuroscience [2], [3], [4], [5], [6], [7] and genetics [8], [9], [12].

Also born in information theory, but not restricted to its applications, there is the relatively new concept of directed information [11], [16], [17]. While mutual information acts as a dependency measure, directed information acts as a causality measure.

Once it is established as a causality measure, there may be applications of directed information in many areas, and so there are developments in scientific literature to estimate directed information from data. For example, it is possible to investigate whether two neurons are directly connected by examining the directed information between their spike trains [18]. In this case, the spike trains are assumed to be stationary ergodic Markov processes. In addition, the conditional intensity functions, which give instantaneous probability of spiking *per* unit time [21], are used with generalized linear models in the estimation of directed information. In the present study, we investigate the use of a different estimator of directed information in order to detect synaptic neuronal connections. This estimator uses context tree weighting algorithm (CTW) [11].

This paper is organized as follows: section II starts by establishing some notation, and section III reviews the definitions of Granger causality, mutual information and directed information. Section IV presents how directed information may be estimated from data using CTW and section V presents directed information estimates to simulated neural spike trains. Finally, section VI concludes the paper.

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.

II. Notation and Terminology

In this paper, we denote random variables by uppercase letters, stochastic processes by uppercase bold letters, and their alphabets by calligraphic letters (e.g., $\mathcal{X}$ denotes the alphabet of random variable X). Subscripts usually denote the outcome's position in a sequence, for example, X_n generally indicates the nth output of the process $\boldsymbol{X}$. Superscripts on a random variable denote finite length sequences of this random variable, for example, $X^N = \{X_1, X_2, \ldots, X_N\}$. Throughout this paper, $\log$ is base 2, $\mathcal{H}(p)$ indicates the binary entropy function, that is, $\mathcal{H}(p) = -p\log p - (1-p)\log(1-p)$, and $\mathrm{E}(X)$ indicates the mean of X.

III. Directed Information and Causality

Before introducing the concept of directed information, it is interesting to review the concept of mutual information. Mutual information is a trustful measure of dependency between random variables, that measures the quantity of uncertainty of a random variable that is reduced when we know the value of other random variable. It is different from Pearson correlation coefficient, capturing non-linear and higher order dependencies between random variables X and Y [13], [14].

Mutual information is defined in equation (1) [1]:

$$\begin{aligned} I(X;Y) &= \sum_{\mathcal{X},\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, &(1)\\ &= H(Y) - H(Y|X), &(2)\end{aligned}$$

where $H(Y) = -\sum_{\mathcal{Y}} p(y)\log p(y)$ is the entropy of Y and $H(Y|X) = -\sum_{\mathcal{X},\mathcal{Y}} p(x,y)\log p(y|x)$ is the conditional entropy of Y relative to X, in bits.

Mutual information can also be applied to multivariate random variables:

$$I(X^N;Y^N) = H(Y^N) - H(Y^N|X^N). \quad (3)$$

The terms in (3) can be calculated through chain rule as

$$\begin{aligned} H(Y^N) &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}),\\ H(Y^N|X^N) &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}X^N), &(4)\end{aligned}$$

where $H(Y^N)$ is the joint entropy of the random variables $Y_1, \ldots, Y_N$, and $H(Y^N|X^N)$ is the conditional entropy of Y^N relative to X^N. The conditional entropy in equation (4) measures the uncertainty of Y^N conditional on the knowledge of X^N [1].

Despite its broad utility, mutual information is not able to detect causality. One used measure of causality, primarily made for analysis in economy, is Granger causality. This quantity measures a statistical dependence between the past of a process and the present of another [17]. It has also been used in neuroscience to detect causal influences of beta oscillations from primary somatosensory and inferior posterior parietal cortices to motor cortex, in monkeys [15]. This measure uses the variances of autoregressive models of random processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ [18], [19]:

$$Y_n = \sum_{m=1}^{p} a_m Y_{n-m} + E_n, \quad (5)$$

$$Y_n = \sum_{m=1}^{p} (b_m Y_{n-m} + c_m X_{n-m}) + \tilde{E}_n, \quad (6)$$

The term p can be fixed a priori or can be evaluated using a model order selection tool [18], E_n and $\tilde{E}_n$ are prediction errors. Granger causality is defined as [18]:

$$G_{\boldsymbol{X}\to\boldsymbol{Y}} = \log \frac{\text{var}(E)}{\text{var}(\tilde{E})}. \quad (7)$$

The lesser the variance of prediction error $\text{var}(\tilde{E})$, the better the prediction performed using $\boldsymbol{X}$, and the greater is its causality over $\boldsymbol{Y}$.

Despite the fact of being a vastly used measure of causality, Granger causality has drawbacks: it cannot be applied to arbitrary joint distributions and imposes a strict probability structure on the data [18].

Directed information comes as another possible causality measure that overcomes these issues. Denoted by $I(X^N \to Y^N)$, directed information is defined as [18], [16]

$$I(X^N \to Y^N) = H(Y^N) - H(Y^N||X^N), \quad (8)$$

$$= \sum_{n=1}^{N} \text{E}\left(\log \frac{P(Y_n|Y^{n-1}X^n)}{P(Y_n|Y^{n-1})}\right) \quad (9)$$

where we have the causally conditioned entropy term:

$$H(Y^N||X^N) = \sum_{n=1}^{N} H(Y_n|Y^{n-1}X^n). \quad (10)$$

The directed information rate from $\boldsymbol{X}$ to $\boldsymbol{Y}$is defined as:

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

Rewriting equations (3) and (8), we notice the subtle difference among them: the index N or n in the X variable, inside the sum.

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

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

This small difference has great impact in the meaning of these measures, since in the case of directed information, the entropy of the process $\boldsymbol{Y}$ is conditioned only to synchronous or past values of the process $\boldsymbol{X}$.

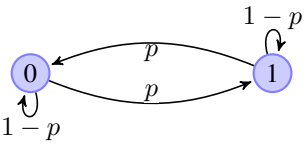

Fig. 1. State diagram for process $\boldsymbol{X}$.

One simple example illustrates the difference between mutual information and directed information and reveals how the latter brings the notion of causality. We can consider a discrete binary Markov chain $\boldsymbol{X}$ as the one illustrated by the state diagram of Fig. 1, and the process $\boldsymbol{Y}$, such that, X_n begins with $n = 1$ and $Y_n = X_{n-1}$, thus $\boldsymbol{Y}$ is causally conditioned on $\boldsymbol{X}$. The mutual information rate is calculated as:

$$\frac{1}{N} I(X^N;Y^N) = \frac{1}{N}\left(H(Y^N) - H(Y^N|X^N)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^N)\right) \quad (11)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(Y_n|Y_{n-1}) - H(Y_n|Y^{n-1}X^N)\right) \quad (12)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(Y_n|Y_{n-1}) - H(Y_n|X_{n-1})\right) \quad (13)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}(p) = \mathcal{H}(p).$$

where (11) comes from the definition, (12) comes from the Markovity of $\boldsymbol{Y}$ and (13) comes from the fact that $Y_n = X_{n-1}$.

On the other hand, we can calculate the rate of directed information from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and from $\boldsymbol{Y}$ to $\boldsymbol{X}$ and see the difference:

$$I_N(X \to Y) = \frac{1}{N}\left(H(Y^N) - H(Y^N||X^N)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^n)\right) \quad (14)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(Y_n|Y_{n-1}) - H(Y_n|X_{n-1})\right) \quad (15)$$

$$= \mathcal{H}(p)$$

$$I_N(Y \to X) = \frac{1}{N}\left(H(X^N) - H(X^N||Y^N)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(X_n|X^{n-1}) - H(X_n|X^{n-1}Y^n)\right) \quad (16)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(H(X_n|X_{n-1}) - H(X_n|X_{n-1})\right) \quad (17)$$

$$= 0.$$

where (14) and (16) come from the definition, (15) comes from the fact that $Y_n = X_{n-1}$ and (17) comes from the Markovity of $\boldsymbol{X}$.

It is clear in this example that $I_N(X \to Y)$ measures the rate of information bits of $\boldsymbol{Y}$ that is causally given by $\boldsymbol{X}$ and

$I_N(Y \to X)$ gives the amount of information in the reverse direction.

IV. Estimation of Directed Information

Jiao *et. al* developed four estimators for a pair of jointly stationary ergodic finite-alphabet processes [11]. He applied the estimators to measure causality in stock markets. As we can observe in equation (9), to calculate directed information between two processes, one may need their probability distributions. Frequently this is not clear from usual data, but has to be estimated. There are methods to estimate mutual information between random variables, when their probability distribution is unknown, such as the binning method and the kth neighbour method (these methods are developed especially to continuous random variables)[9], [12]. For large stochastic processes assuming discrete values and their estimation of directed information, a more suitable method is CTW [11].

We explain briefly how the CTW works. 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, were b_i is the counting of symbol i, $i \in \{0, \ldots, M-1\}$ [20], [11]:

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

The root node is denoted by λ and P_w^λ is the universal probability assignment by CTW. Fig. 2 brings an example of a context tree with depth $D = 2$, given for a binary sequence with size $n = 6$: $x_{-1}x_0x_1 \ldots x_6 = 00100110$. The tree begins in the root that counts the number of occurrences of zeros and ones in the sequence $x_1x_2 \ldots x_n$. The counting continues by observing how many times the context indicated following the tree branches precedes zeros (a_s) and ones (b_s), forming the pair (a_s, b_s) in each node. For instance, counting how many times the context $s = 11$ precedes 0 and 1 is given in the superior leaf of the tree, resulting in the values $a_{11} = 1$ and $b_{11} = 0$.

In order to compute P_w^s, there is the following formula:

$$P_w^s = \begin{cases} \frac{1}{2}P_e^s(x^n) + \frac{1}{2}\prod_{i=0}^{M-1} P_w^{is}(x^n), & \text{for } 0 \le l(s) < D, \\ P_e^s(x^n), & \text{for } l(s) = D, \end{cases}$$

where s indicates the node/context of the tree, $l(s)$ indicates the length of the context.

For estimation of directed information, one of the estimators of Jiao (estimator 2) [11] uses the following formulas, for each outcome of length N of the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$:

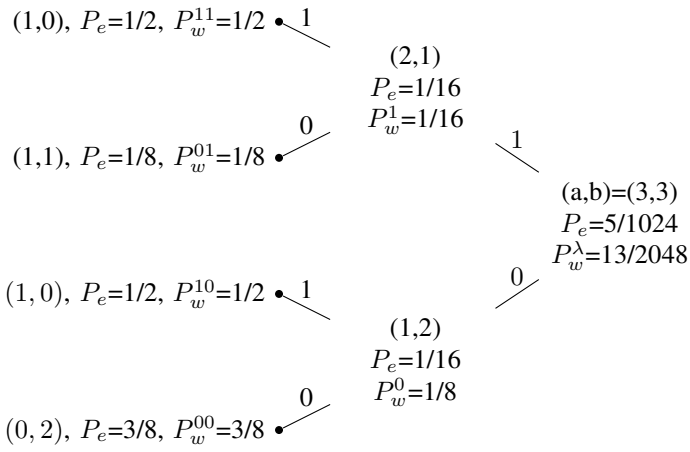

Fig. 2. Context tree example of depth $D = 2$, with estimated probabilities and weighted probabilities, for sequence $x_{-1}x_0x_1 \ldots x_6 = 00100110$. Each node counts how many times the context indicated by following the tree branches precedes zeros and ones. For example, the counts for how many times the context $s = 00$ precedes zeros and ones on the sequence $x_1x_2 \ldots x_6 = 100110$ is given in the inferior leaf of the tree, giving the values 0 and 2, respectively.

$$\begin{aligned} \hat{I}(X^N \to Y^N) &= \hat{H}(Y^N) - \hat{H}(Y^N||X^N), \text{ where} \\ \hat{H}(Y^N) &= \frac{1}{N}\sum_{n=1}^{N}\sum_{\mathcal{Y}} Q(y_{n+1}|Y^n)\log\frac{1}{Q(y_{n+1}|Y^n)} \\ \hat{H}(Y^N||X^N) &= \sum_{n=1}^{N} f(Q(x_{n+1}, y_{n+1}|X^n, Y^n)), \\ f(P) &= -\sum_{\mathcal{X},\mathcal{Y}} P(x,y)\log P(y|x), \\ Q(x_{N+1}|x^N) &= \frac{P_w^\lambda(x^{N+1})}{P_w^\lambda(x^N)} \end{aligned}$$

Fig. 3 illustrates this estimator implementation for the example of section III. Here we establish $p = 0.1$. The directed information rates in the direct and reverse order are plotted as a function of sample size N.

V. Directed Information for Detection of Synaptic Connections

In this section, we investigate the use of Jiao's estimator (estimator 2) [11] for directed information for the application of detecting connections among neural spike trains. In order to do so, we simulate 5 neurons spiking, one of these produces inhibitory synapses with stronger synaptic connections, while the other 4 produce excitatory synapses, as observed in mammalian cortex [22]. Fig. 4 illustrates the connections, where neuron 5 is the one presenting inhibitory connections. Fig. 4 was obtained and adapted from reference [18]. Each neuron receives a noisy thalamic input, besides the synaptic input.

Directed information estimates were performed with a duration of 100000ms, each millisecond is a time bin with the possibility of 1 or 0 spike. The synaptic connection weights were varied randomly from a uniform distribution in 50 trials, but with fixed constants multiplying them. The

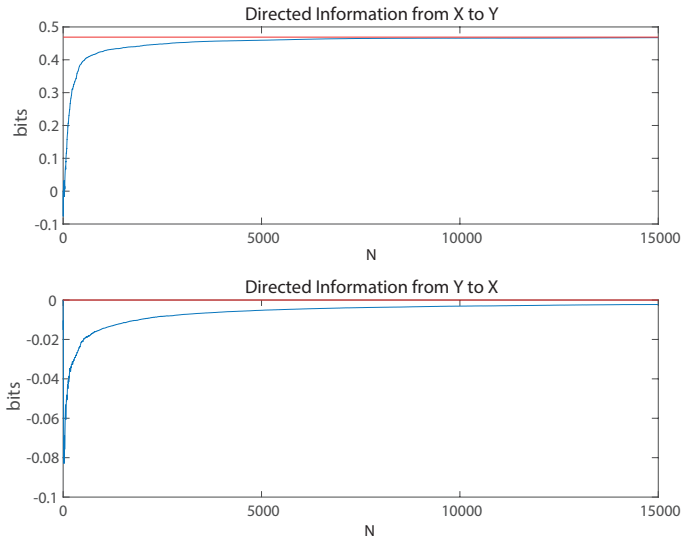

Fig. 3. Estimation of directed information rates for the example cited in section III, from $\boldsymbol{X}$ to $\boldsymbol{Y}$ in the upper panel and from $\boldsymbol{Y}$ to $\boldsymbol{X}$ in the lower panel. The transition probability of $\boldsymbol{X}$ was $p = 0.1$. The estimates blue curves vary according to the size N of the process, red lines indicate the true directed information rates.

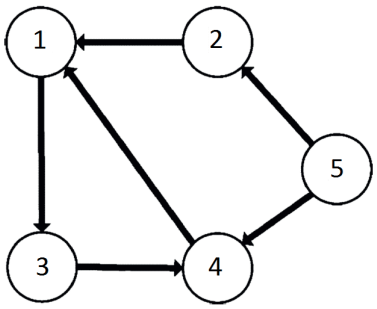

Fig. 4. Diagram to show connections among simulated neurons. Neuron 5 presents inhibitory synaptic connections. Figure obtained and adapted from reference [18].

constants multiplying inhibitory connections were in double measure greater than those multiplying excitatory connections. For more details on this simulation we refer to Izhikevich [22]. The outcome of spike trains for a trial over a 1000ms window is shown in Fig. 5.

For every trial we estimated the directed information rate between every pair of neural spike trains. After that, we calculated the normalized directed information rates [18], given as in equation (18):

$$\frac{I_N(X \to Y)}{H_N(Y)}, \tag{18}$$

where $H_N(Y)$ is the entropy rate of the spike train Y. When the normalized directed information rate is close to one, there is an indication of strong causal relationship, while normalized directed information close to zero indicates a weak causal relationship. Fig. 6 illustrates the synaptic connection weights between each pair of neurons for a trial, and Fig. 7 indicates mean normalized directed information rates over 50 trials.

We observed that positive values of the normalized directed information rates always indicated true synaptic connections among neurons, over the 50 trials. We also observed that zero or negative values of estimates could identify the absence of

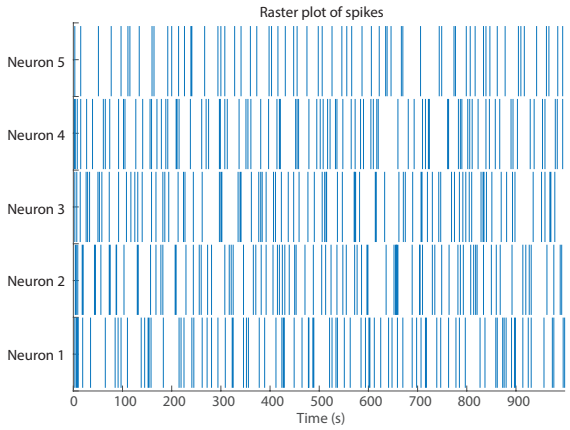

Fig. 5. Raster plot of a trial shows spiking of each of 5 neurons in a time window of 1000ms.

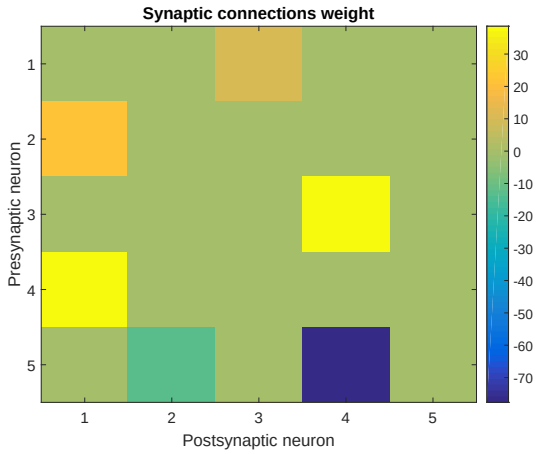

Fig. 6. Synaptic connection weights among neurons in a trial.

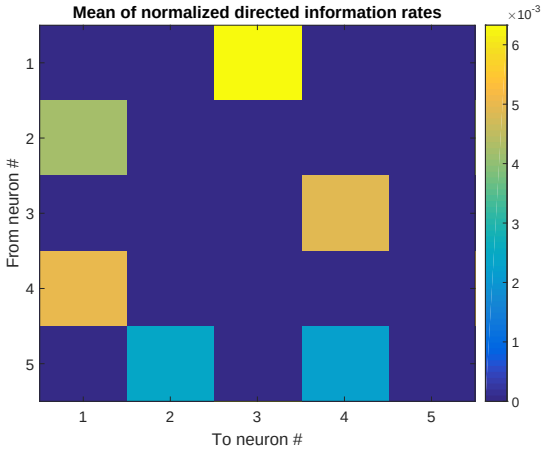

Fig. 7. Mean normalized directed information rates over 50 trials. The values in the principal diagonal were set to zero for visual perception of the other estimates in the scale of color bar, but their true values are actually equal to 1. Negative values of normalized directed information estimates were set to zero.

one synaptic connection or not. That is, even small, positive values of normalized directed information rates, indicating a weak causal relation (the values found here were in the order of 10^{-3}), in our simulations correctly identified synaptic connections, but negative or zero values could not reliably measure the absence of synaptic connections. We counted the number of correct connections detected and show these in Fig. 8.

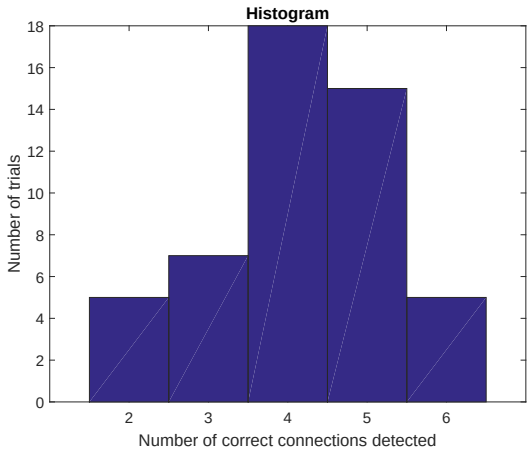

Fig. 8. Histogram counting the number of correct connections detected by positive values of normalized directed information rates estimates over 50 trials. The true number of synaptic connections was 6.

The results encountered here with Jiao's estimator are different from those found in [18]. There, positive values for directed information are frequently found, which usually happen because of indirect influences of one neuron over another. For example, in Fig. 4, neuron 1 indirectly causes spike train of neuron 4. However, in this work, the major issue found was that of not detecting causal relationship when there was. For this reason, the estimator used here did not always detect a causal relationship, but when it did, it was a reliable evidence of synaptic connection.

We should stress that the encountered values of directed information may vary according to simulated synaptic connection weights. Also, we observed in this study that despite having stronger synaptic connection weights in the mean, the inhibitory connections presented lower mean values of directed information estimates between neural spike trains (Fig. 7).

VI. Conclusion

We conclude from this paper that directed information is a measure with great applicability to measure causality between two stochastic processes, which has been used in areas such as neuroscience and economy. For this applicability, sometimes it is necessary to infer probability distributions, which may be done by the CTW algorithm. Moreover, we observed that Jiao's estimator (estimator 2) may detect true synaptic neuronal connections, which happened when there were positive values of directed information between the spike trains of presynaptic and postsynaptic neurons.

Acknowledgments

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

References

[1] Cover T and Thomas J. Entropy, Relative Entropy, and Mutual Information. *Elements of Information Theory*. New York: Wiley and Sons, 2006.

[2] G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck. "Neural coding of naturalistic motion stimuli", *Comput. Neural Syst.*, v. 12, Mar 2001.

[3] C. Kayser, M. A. Montemurro, N. K. Logothetis, S. Panzeri. "Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns".*Neuron*, v. 61, Feb 2009.

[4] I. Nelken, G. Chechick, T. D. Mrsic-Flogel, A. J. King and J. W. H. Schnupp. "Encoding stimulus information by spike numbers and mean response time in primary auditory cortex". *J Comput Neurosci*, v. 19, Oct 2005.

[5] E. Arabzadeh, S. Panzeri and M. E. Diamond. "Whisker vibration information carried by rat barrel cortex neurons". *J Neurosci*, v. 24, Jun 2004.

[6] R. Pavão, C. E. Piette, V. Lopes-dos-Santos, D. B. Katz and A. B. L. Tort. "Local field potentials in the gustatory cortex carry taste information". *J Neurosci*, v. 34, Jun 2014.

[7] R. Q. Quiroga and S. Panzeri. "Extracting information from neuronal populations: information theory and decoding approaches". *Nat Rev Neurosci*. v. 10, Mar 2009.

[8] I. Grosse. Applications of Statistical Physics and Information Theory to the Analysis of DNA Sequences. Ph.D. dissertation, Boston University. 2000.

[9] A. Kraskov, H. Stögbauer, P. Grassberger. "Estimating mutual information". *Physical Review E*, v. 69, Jun 2004.

[10] J. Seok and Y. S. Kang. "Mutual information between discrete variables with many categories using recursive adaptive partitioning". *Sci Rep.*, v. 5, Jun 2015.

[11] J. Jiao, H. H.Permuter, L. Zhao, Y. Kim, T. Weissman. "Universal estimation of directed information". *IEEE Trans. Inf. Theory*, v. 59, Oct 2013.

[12] B. Ross. "Mutual information between discrete and continuous data sets". *PLoS One*, v. 9, Feb 2014.

[13] S. Lu. Measuring Dependence via Mutual Information. M.Sc. Thesis, Queen's University, Kingston, 2001.

[14] W. Li. "Mutual information functions versus correlation functions". *J. Stat. Phys.*, v. 60, Sep 1990.

[15] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. "Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality". *Proceedings of the National Academy of Sciences of the United States of America*, v. 101, May 2004.

[16] G. Kramer. Directed Information for Channels with Feedback. Ph.D. Dissertation, Swiss Federal Institute of Technology, Zurich, 1998.

[17] P. O. Amblard, and O. J. J. Michel. "The relation between Granger causality and directed information theory: a review.". *Entropy*, v. 15, Dec 2012.

[18] C. 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*, v. 30, Feb 2011.

[19] M. Kamiński, M. Ding, W. Truccolo and S. Bressler. "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance". *Biol Cybern*, v. 85, Aug 2001.

[20] F. M. J. Willems, Y. M. Shtarkov and T. J. Tjalkens. "The context-tree weighting method: basic properties". *IEEE Trans. Inf. Theory*, v. 41, May 1995.

[21] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, E. N. Brown. "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects". *J. Neurophysiol*, v. 93, Sep 2005.

[22] E. M. Izhikevich. "Simple model of spiking neurons". *IEEE Trans. Neural Networks*, v. 14, Nov 2003.