NIH Public Access
Author Manuscript
J Comput Neurosci. Author manuscript; available in PMC 2011 September 13.

Published in final edited form as:
J Comput Neurosci. 2011 February ; 30(1): 17–44. doi:10.1007/s10827-010-0247-2.

Estimating the directed information to infer causal relationships in ensemble neural spike train recordings

Christopher J. Quinn,
Department of Electrical & Computer Engineering, University of Illinois, Urbana, IL, USA quinn7@illinois.edu

Todd P. Coleman,
Neuroscience Program, Department of Electrical & Computer Engineering, University of Illinois, Urbana, IL, USA colemant@illinois.edu

Negar Kiyavash, and
Department of IESE, University of Illinois, Urbana, IL, USA kiyavash@illinois.edu

Nicholas G. Hatsopoulos
Committees on Computational Neuroscience & Neurobiology, Department of Organismal Biology & Anatomy, University of Chicago, Chicago, IL, USA nicho@uchicago.edu

Abstract

Advances in recording technologies have given neuroscience researchers access to large amounts of data, in particular, simultaneous, individual recordings of large groups of neurons in different parts of the brain. A variety of quantitative techniques have been utilized to analyze the spiking activities of the neurons to elucidate the functional connectivity of the recorded neurons. In the past, researchers have used correlative measures. More recently, to better capture the dynamic, complex relationships present in the data, neuroscientists have employed causal measures—most of which are variants of Granger causality—with limited success. This paper motivates the directed information, an information and control theoretic concept, as a modality-independent embodiment of Granger's original notion of causality. Key properties include: (a) it is nonzero if and only if one process causally influences another, and (b) its specific value can be interpreted as the *strength* of a causal relationship. We next describe how the causally conditioned directed information between two processes given knowledge of others provides a network version of causality: it is nonzero if and only if, in the presence of the present and past of other processes, one process causally influences another. This notion is shown to be able to differentiate between true direct causal influences, common inputs, and cascade effects in more two processes. We next describe a procedure to estimate the directed information on neural spike trains using point process generalized linear models, maximum likelihood estimation and information-theoretic model order selection. We demonstrate that on a simulated network of neurons, it (a) correctly identifies all pairwise causal relationships and (b) correctly identifies network causal relationships. This procedure is then used to analyze ensemble spike train recordings in primary motor cortex of an awake monkey while performing target reaching tasks, uncovering causal relationships whose directionality are consistent with predictions made from the wave propagation of simultaneously recorded local field potentials.

Keywords

Causality; Functional connectivity; Point processes; Mutual information

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

1 Introduction

Due to recent advances in multiple electrode recording techniques, neuroscientists are now able to record the simultaneous, individual activity of hundreds of neurons in various regions of the brain. Many researchers are asking questions about how the spiking of individual neurons are influencing and influenced by the spiking activity of specific neighboring neurons, the local ensemble spiking activity, and the spiking activity in other regions of the brain. The dynamic, complex interactions between neurons often make thorough analysis of the large volumes of data, such as identifying the complete, functional topology of the recorded neurons, quite difficult. However, since such analyses could provide great insight into brain function, there has been a concentrated effort by researchers to develop techniques to perform them.

Past work in analyzing the relationships between multiple, simultaneously recorded neurons often involved using correlative measures. Quantitative techniques such as cross correlation (Eguiluz et al. 2005; Diekman et al. 2009), mutual information (Paninski 2003), and coherence (Salvador et al. 2005) have been employed, as they can provide insight to whether two or more spike trains are statistically dependent. Such information can be helpful to discern potential functional connections between pairs of neurons or even groups of neurons. However, since these measures are correlative, they are symmetric, and thus do not capture any of the *directionality* that might be present. For example, even if the spiking activity of neuron A directly affects the spiking activity of neuron B, but there is no influence in the other direction, correlative measures would only be able to identify that the spiking activity of neurons A and B are related.

To address this issue, neuroscientists have recently begun using alternative, *causal* measures. The goal of using these measures is to identify the *types* of relationships between the spiking activity of the recorded neurons, to distinguish if the relationship between neurons A and B are mutual (both influence each other) or unidirectional (A influences B, but B does not influence A). Some of the quantitative techniques that have been used to identify causal relationships include Granger causality (Granger 1969), extensions of Granger causality such as directed transfer function (Kaminski and Blinowska 1991), transfer entropy (Schreiber 2000), and dynamic causal modeling (Friston et al. 2003), among others.

The aforementioned techniques have been used in a variety of settings in the recent past. In some situations, they have provided insight into the underlying causal relationships, and consequently the functional connectivity, of the recorded neurons. However, in other cases, the complicated relationship structures have led to mis-interpretations (Kamiński et al. 2001).

This paper connects a newly defined information-theoretic concept of “directed information” to the Granger's philosophical relationship between causality and prediction (Granger 1969) in a rigorous manner, operating on arbitrary modalities. The estimation procedure on neural spike trains requires milder assumptions than other techniques and has strong proven consistency properties. Directed information plays a fundamental role in information theory, especially in communication with feedback (Massey 1990; Rissanen and Wax 1987; Kramer 1998; Marko 1973). Within the context of causality, directed information has been used sparingly to infer the causal structure of gene regulatory networks (Rao et al. 2006, 2007, 2008; Mathai et al. 2007) and neural data (Amblard and Michel 2010). However, in all the aforementioned papers, either a plug-in estimator (described in Section 5.1), that is not necessarily statistically consistent was used, or no proposed estimation schemes were discovered. In this paper, we demonstrate a consistent estimation

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

procedure to infer the directed information between two point processes (representing neural spike trains). Moreover, we extend this notion to define "direct" causal relationships that uncover causal relationships between networks of spike trains.

This paper is organized as follows. First, in Section 2, definitions and notations are provided. In Section 3, previously used causal measures will be discussed. Section 4 introduces the directed information as a measure of causal influence. Section 5 develops an estimation paradigm, which is consistent under appropriate assumptions, to estimate the directed information from simultaneous recordings. A measure of the *strength* of each estimated pairwise influence, along with 95% confidence intervals of the directed information, are also presented. Section 6 introduces causal conditioning to infer causal relationships between a *network* of processes. This is particularly important to differentiate between true direct causal influences and common inputs or cascade effects. Section 7 demonstrates results of this estimation paradigm on synthetically constructed neural spike trains, where all pairwise causal relationships (whether there is an influence or not) are correctly identified. This procedure subsequently is used to effectively uncover the network structure of relationships between processes and differentiates direct causal influences from common inputs or cascade effects. This procedure is also used to analyze ensemble spike train recordings in primary motor cortex of an awake monkey while executing movements pertaining to target reaching (Wu and Hatsopoulos 2006). The procedure identified strong structure in the estimated causal relationships, the directionality of which is consistent with predictions made from the wave propagation of simultaneously recorded local field potentials (Rubino et al. 2006).

2 Definitions and notations

In this section, we provide probabilistic notations and information-theoretic definitions that will be used throughout the remainder of the manuscript. Denote definitions with the symbol $\triangleq$.

– For integers $i \leq j$, define $x_i^j \triangleq (x_i, \cdots, x_j)$. For brevity, define $x^n \triangleq x_1^n = (x_1, \ldots, x_n)$

– Throughout this paper, $\mathcal{X}$ corresponds to a measurable space that a random variable, denoted with upper-case letters (X) takes values in, and lower-case values $x \in \mathcal{X}$ correspond to specific realizations.

– We define the probability mass function (PMF) of a discrete random variable by

$$\mathrm{P}_X(x) \triangleq P(X=x),$$

and the probability density function (PDF) of a continuous random variable by

$$f_X(x) \triangleq \lim_{\Delta \to 0} \frac{P(X \in [x, x+\Delta))}{\Delta}$$

– The entropy of a discrete random variable is given by

$$\mathrm{H}(X) = \sum_{x \in \mathcal{X}} -\mathrm{P}_X(x) \log \mathrm{P}_X(x) \tag{1}$$

– The conditional entropy is given by

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mathrm{H}(Y|X) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} -\mathrm{P}_{X,Y}(x,y)\ \log \mathrm{P}_{Y|X}(y|x) \tag{2}$$

– The chain rule for entropy is given by

$$\mathrm{H}(X^n) = \sum_{i=1}^{n} \mathrm{H}\left(X_i|X^{i-1}\right) \tag{3}$$

– For two probability distributions P and Q on $\mathcal{X}$, the Kullback–Leibler (KL) divergence is given by

$$\begin{aligned}\mathrm{D}(\mathrm{P}\parallel \mathrm{Q}) &\triangleq \mathrm{E}_{\mathrm{P}}\left[\log\frac{\mathrm{P}(X)}{\mathrm{Q}(X)}\right] \\ &= \sum_{x\in\mathcal{X}} \mathrm{P}(x)\ \log\frac{\mathrm{P}(x)}{\mathrm{Q}(x)} \geq 0\end{aligned} \tag{4}$$

– The mutual information between random variables X and Y is given by

$$\mathrm{I}(X;Y) \triangleq \mathrm{D}(P_{XY}(\cdot,\cdot)\parallel \mathrm{P}_X(\cdot)\mathrm{P}_Y(\cdot)) \tag{5a}$$

$$=\mathrm{E}_{\mathrm{P}_{XY}}\left[\log\frac{\mathrm{P}_{Y|X}(Y|X)}{\mathrm{P}_Y(Y)}\right] \tag{5b}$$

$$=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathrm{P}_{X,Y}(x,y)\ \log\frac{\mathrm{P}_{Y|X}(y|x)}{\mathrm{P}_Y(y)} \tag{5c}$$

$$=\mathrm{H}(Y) - \mathrm{H}(Y|X) \tag{5d}$$

The mutual information is known to be symmetric: $\mathrm{I}(X;Y) = \mathrm{I}(Y;X)$.

– The chain rule for mutual information is given by

$$\mathrm{I}(X^n;Y^n) = \sum_{i=1}^{n} \mathrm{I}\left(Y_i;X^n|Y^{i-1}\right) \tag{6}$$

where the conditional mutual information is given by

$$\mathrm{I}(X;Y|Z) = \mathrm{E}_{\mathrm{P}_{XYZ}}\left[\log\frac{\mathrm{P}_{Y|X,Z}(Y|X,Z)}{\mathrm{P}_{Y|Z}(Y|Z)}\right]. \tag{7}$$

– We denote a random process by $\mathbf{X} = (X_i : i \geq 1)$, with associated $\mathrm{P}_{\mathbf{X}}(\cdot)$ which induces the joint probability distribution of all finite collections of $\mathbf{X}$.

– We denote the set of kth order Markov chains as

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\mathcal{M}_k(\mathcal{X}) = \left\{ P_{\mathbf{X}} : \mathrm{P}_{X^n}(x^n) = \prod_{i=1}^{n} \mathrm{P}_{X_i|X_{i-k}^{i-1}}\left(x_{i-k}^{i-1}\right) \right\}$$

with $X_j \triangleq \varnothing$ for $j<0$.

– We denote the set of all finite-memory random processes on $\mathcal{X}$ as

$$\mathcal{M}(\mathcal{X}) = \bigcup_{k\geq 1} \mathcal{M}_k(\mathcal{X})$$

– We denote the set of stationary and ergodic random processes on $\mathcal{X}$ as SE ($\mathcal{X}$)

– The entropy rate and mutual information rate, assuming they exist, are given as follows

$$\mathcal{H}(Y) \triangleq \lim_{n\to\infty} \frac{1}{n} H(Y^n) \tag{8}$$

$$\mathcal{I}(X;Y) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n) \tag{9}$$

– Within the context of point processes, consider the time interval $(0, T]$ as the time window for which our neural spike train is observed. In this context, define $\mathcal{Y}_T$ to be the set of functions $y : (0, T] \to \mathbb{Z}_+$ that are non-decreasing, right-continuous, and $y_0 = 0$. In other words, $\mathcal{Y}_T$ is the set of point processes on $(0, T]$. Succinctly, we can represent a point process as a sample path $y \in \mathcal{Y}_T$ where each jump in y corresponds to the occurrence of a spike (at that time)

– Consider two random processes $\mathbf{X} = (X_\tau : 0 \leq \tau \leq T)$ and $\mathbf{Y} = (Y_\tau : 0 \leq \tau \leq T) \in \mathcal{Y}_T$. Define the *histories* at time t for the point process $Y \in \mathcal{Y}_T$ as the σ-algebra generated by appropriate random processes up to time t as:

$$\mathcal{F}_t = \sigma\left(X_\tau : \tau \in [0,t],\quad Y_\tau : \tau \in [0,t)\right) \tag{10a}$$

$$\mathcal{F}_t' = \sigma\left(Y_\tau : \tau \in [0,t)\right) \tag{10b}$$

It is well known that the conditional intensity function (CIF) completely characterizes the statistical structure of all well-behaved point processes used in statistical inference of neural data (Brown et al. 2003). The CIF is defined as (Daley and Vere-Jones 1988):

$$\lambda(t \,\|\, \mathcal{F}_t) \triangleq \lim_{\Delta\to 0} \frac{P(Y_{t+\Delta} - Y_t = 1 | \mathcal{F}_t)}{\Delta}, \tag{11a}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\lambda\left(t \parallel \mathcal{F}_t^{'}\right) \triangleq \lim_{\Delta \to 0} \frac{P\left(Y_{t+\Delta} - Y_t = 1 | \mathcal{F}_t^{'}\right)}{\Delta} \tag{11b}$$

Succinctly, the conditional intensity specifies the instantaneous probability of spiking per unit time, given *previous* neural spiking (and, in the scenario when using $\mathcal{F}_t$, also previous exogenous inputs X). Almost all neuroscience point process models (Brown et al. 2002) implicitly use this *causal* assumption in the definition of $\mathcal{F}_t$ given by Eq. (10a). Examples of how $\mathcal{F}_t$ is interpreted will appear in the experimental results section.

– For a point process $Y \in \mathcal{Y}_T$ with conditional intensity functions $\lambda\ (t \parallel \mathcal{F}_t)$ and $\lambda\left(t \parallel \mathcal{F}_t^{'}\right)$, the likelihood or density of Y at y given x is given by (Brown et al. 2003)

$$f_{Y \| X}\ (y \parallel x; \lambda) = \exp\left\{\int_0^T \log \lambda\ (t \parallel \mathcal{F}_t)\ dy_t - \lambda\ (t \parallel \mathcal{F}_t)\ dt\right\}, \tag{12}$$

and analogously, the marginal likelihood or density of Y at y is given by

$$f_Y\ (y; \lambda) = \exp\left\{\int_0^T \log \lambda\left(t \parallel \mathcal{F}_t^{'}\right) dy_t - \lambda\left(t \parallel \mathcal{F}_t^{'}\right) dt\right\}. \tag{13}$$

We use the ‖ notation to explicitly speak to how these conditional probabilities in Eq. (11b) are taken with respect to causal histories, specified in Eq. (10b). By discretizing $(0, T]$ into $n = T/\Delta$ intervals of length $\Delta << 1$ so that $dy = (dy_1, \ldots, dy_n)$ with $dy_i \triangleq y_{(i+1)\Delta} - y_{i\Delta} \in \{0, 1\}$, we can approximate Eqs. (12) and (13) by

$$-\log\ f_{Y \| X}\ (y \parallel x; \lambda) \simeq \sum_{i=1}^{n} -\log \lambda\ (i \parallel \mathcal{F}_i)\ dy_i + \lambda\ (i \parallel \mathcal{F}_i)\ \Delta. \tag{14}$$

$$-\log\ f_Y\ (y; \lambda) \simeq \sum_{i=1}^{n} -\log \lambda\left(i \parallel \mathcal{F}_i^{'}\right) dy_i + \lambda\left(i \parallel \mathcal{F}_i^{'}\right) \Delta. \tag{15}$$

where the discrete time index i corresponds to the continuous interval $(0, T]$ at time $i\Delta$.

– Denote the set of *GLM* point processes with discrete-time (Δ) conditional likelihood pertaining to a generalized linear model of the conditional intensity: by the function h to be

$$\mathrm{GLM}_{J,K}\ (h) = \left\{\lambda : \log \lambda\ (i \parallel \mathcal{F}_i) = \alpha_0 + \sum_{j=1}^{J} \alpha_j dy_{i-j} + \sum_{k=1}^{K} \beta_k h_k\ (x_{i-(k-1)})\right\}$$

The function $(h_1, \ldots, h_K)$ operate on the extrinsic covariate X in the recent past. We subsequently define GLM (h) as

$$\mathrm{GLM}\ (h) = \bigcup_{J \geq 1, K \geq 1} \mathrm{GLM}_{J,K}\ (h).$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3 Previous approaches to identify causal relationships in neural data

3.1 Granger causality and DTF

Granger causality (Granger 1969) has been perhaps the most widely-established means of identifying causal relations between two times series (Hesse et al. 2003). It operates by calculating the variances to correction terms for autoregressive models. Given two time series $\mathbf{X} = \{X_i : i \geq 1\}$ and $\mathbf{Y} = \{Y_i : i \geq 1\}$, to determine whether **X** causally influences **Y, Y** is first modeled as an univariate autoregressive series with error correction term V_i:

$$Y_i = \sum_{j=1}^{p} a_j Y_{i-j} + V_i$$

Then **Y** is modeled again, but this time using the **X** series as causal side information:

$$Y_i = \sum_{j=1}^{p} \left[b_j Y_{i-j} + c_j X_{i-j} \right] + \tilde{V}_i$$

with $\tilde{V}_i$ as the new error correction term. The value of p can be fixed a priori or determined using a model order selection tool (Akaike 1976; Barron and Cover 1991). The Granger causality is defined as

$$G_{X \to Y} \triangleq \log \frac{\mathrm{var}(V)}{\mathrm{var}(\tilde{V})}. \tag{16}$$

This technique examines the ratio of the variances of the correction terms. If including X in the modeling improves the model, then the variance of the correction term $\tilde{V}_I$ will be lower, and thus $G_{X \to Y} > 0$. Usually $G_{X \to Y}$ and $G_{Y \to X}$ are compared, and the larger term is taken to be the direction of causal influence.

The directed transfer function (DTF) (Kaminski and Blinowska 1991) is related to Granger causality, with the principle difference being that it transforms the autoregressive model into the spectral domain (Kamiński et al. 2001). Instead of working with univariate and bivariate models, DTF works with multivariate models for each time series, and so in theory should improve the modeling, since it can take into account the full covariance matrix for each of the time series (Kamiński et al. 2001).

These and derivative techniques have been used extensively (Hesse et al. 2003; Uddin et al. 2009; Goebel et al. 2003; Roebroeck et al. 2005; Rogers et al. 2007; Dhamala et al. 2008; Abler et al. 2006; Korzeniewska et al. 2003; Wang et al. 2007; Brovelli et al. 2004). These approaches can be attractive because they are generally fast to compute and easy to interpret. However, because of the sample-variance calculations, they are not necessarily statistically appropriate for statistical inference questions pertaining to neural spike train data—which are generally modeled as point processes. Autocorrelations and spectral transforms on point processes often do not accurately provide meaningful, conceptual interpretations. Moreover, these approaches do not have strong statistical guarantees of correctly identifying causal relations. Another issue is that even in cases where they can detect a causal influence, these approaches do not necessarily identify the *extent* of the influence (whether A fully causes B or only partially). It is not clear that the actual values obtained through these methods, $G_{X \to Y}$, have a physical meaning beyond comparison with the opposite direction (e.g. $G_{X \to Y}$ v.s. $G_{Y \to X}$).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

3.2 Transfer entropy

Transfer entropy was developed by Schreiber (2000). It assumes two stochastic processes $\mathbf{X} = (X_i : i \geq 1)$ and $\mathbf{Y} = (Y_i : i \geq 1)$ satisfy a Markov property:

$$\mathrm{P}_{Y_{n+1}|Y^n,X^n}\left(y_{n+1}|y^n,x^n\right)=\mathrm{P}_{Y_{n+1}|Y_{n-J+1}^n,X_{n-K+1}^n}\left(y_{n+1}|y_{n-J+1}^n,x_{n-k+1}^n\right)$$

for some known constants J and K. Schreiber defined transfer entropy as:

$$T_{X\to Y}(i)=I\left(Y_{i+1};X_{i-k+1}^i|Y_{i-J+1}^i\right)$$

This term is part of a sum of terms (Eq. (21)) that will be defined as the directed information (with a Markov assumption applied). Some studies have employed this measure (Chávez et al. 2003; Gourevitch and Eggermont 2007; Kraskov 2008). This has not been as widely employed as Granger causality and related measures, principally due to the lack of convergence properties (Pereda et al. 2005). As no model for the underlying distribution is suggested, the straightforward approach to estimate the transfer entropy is to use plug-in estimates, which in general do not have statistical convergence guarantees.

3.3 Dynamic causal modeling

Dynamic causal modeling (DCM) (Friston et al. 2003) is a recently developed procedure which differs in its approach from previously discussed techniques. DCM models the brain as a *deterministic*, causal, dynamic multiple-input and multiple-output (MIMO) system, with a priori unknown coupling coefficients. Through a series of perturbations and observations, the potentially time varying coefficients of the system are estimated using Bayesian inference (Friston et al. 2003). By incorporating dynamic coefficients, DCM could potentially capture the effects of plasticity, which the aforementioned procedures, which assume static coefficients, cannot. DCM has been applied to both fMRI studies (Stephan et al. 2008; Grefkes et al. 2008; Hamandi et al. 2008; Schuyler et al. 2009; Bitan et al. 2005), and EEG and MEG studies (David et al. 2006). While it has been applied with some success to certain brain imaging studies, to the authors' knowledge, it has not been shown to robustly characterize causal relationships in local recording data such as data obtained with large electrode arrays. Also, although there are asymptotic convergence results for some of the coefficients through properties of EM estimation (Friston et al. 2003), the model as a whole does not have statistically guaranteed convergence properties.

4 Directed information as a robust measure of statistical causality

We will now motivate and introduce a more general measure of statistical causality—directed information—and discuss how it has a meaningful interpretation of statistically causal influences in a variety of settings.

4.1 Motivation and setup

4.1.1 Background—To help address the aforementioned issues, consider the construction of Granger causality. In his original paper, Granger defined causality as "We say that X_t is causing Y_t if we are better able to predict Y_t, using all available information [up to time t] than if the information apart from X_t had been used" (Granger 1969). Despite the generality of this conceptual definition, his functional definition was restricted to linear models for the ease of computation and used variances of correction terms in quantifying causality because variance is easy to compute and understand (Granger 1969).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Two decades later, Rissanen and Massey, both Shannon award winners, independently introduced a different functional definition of causality (Rissanen and Wax 1987; Massey 1990). Massey, whose work is based on earlier work by Marko (1973), named the quantity *directed information*. Directed information is philosophically grounded on the same principle as Granger causality: the extent to which X statistically causes Y is measured by how helpful causal side information of process X is to predicting the future of Y, given knowledge of Y's past. Unlike Granger causality, directed information is not tied to any particular statistical model. It operates on log likelihood ratios—which exist for an arbitrary modality. If we imagine the random processes are all discrete, one interpretation of directed information is the reduction in the minimum number of bits required to sequentially specify a source Y given causal knowledge of X. Specifically, the Shannon codelengths are the "ideal" codelengths (description lengths) of a random source Y (Cover and Thomas 2006)—using such a code length mapping results in the average description length in bits being within one bit of its fundamental limit: the entropy $H(Y)$. Shannon codes are a function of the random sequence's probability distribution:

$$l_{Shannon}(x) \triangleq \lceil \log \frac{1}{P_X(x)} \rceil.$$

For example, if a source Y has distribution P_Y and is jointly distributed with X according to $P_{X,Y}$, then the reduction in the minimum number of bits required to specify Y given knowledge of X is the *mutual information*:

$$\begin{aligned} \mathrm{E}_{P_{X,Y}}\left[\log \frac{1}{P_Y(Y)} - \log \frac{1}{P_{Y|X}(Y|X)}\right] &= \mathrm{D}\left(P_{Y|X} \,\|\, P_Y\right) \\ &= I(X;Y) \end{aligned} \tag{17}$$

where Eq. (17) follows from Eq. (5d). The mutual information is symmetric, and nonzero if and only if the two random variables are statistically independent. From Eq. (6), for vectors X^n and Y^n, the mutual information can be written as:

$$\mathrm{I}(X^n;Y^n) = \sum_{i=1}^{n} \mathrm{I}\left(X^n;Y_i|Y^{i-1}\right) \tag{18}$$

$$= \mathrm{E}\left[\sum_{i=1}^{n} \log \frac{\mathrm{P}_{Y_i|Y^{i-1},X^n}\left(Y_i|Y^{i-1},X^n\right)}{\mathrm{P}_{Y_i|Y^{i-1}}\left(Y_i|Y^{i-1}\right)}\right] \tag{19}$$

$$= \sum_{i=1}^{n} \mathrm{D}\left(\mathrm{P}_{Y_i|Y^{i-1},X^n} \,\|\, \mathrm{P}_{Y_i|Y^{i-1}}\right) \tag{20}$$

where Eq. (18) follows from Eq. (6); Eq. (19) follows from Eq. (7); and Eq. (20) follows from Eq. (4). The symmetry $I(X^n; Y^n) = I(Y^n; X^n)$ implies that the mutual information only measures the correlation between random processes (in the colloquial sense of statistical dependence), and will be unable to capture directionality of causation.

4.1.2 Definition of directed information—Shannon-award winners Rissanen and Massey separately modified the mutual information to capture causal influences (Rissanen

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and Wax 1987; Massey 1990), and this new quantity is known as the *directed information*, denoted I(X → Y), between two stochastic processes X and Y. With similar work as above:

$$\mathrm{I}(X^n \to Y^n) \triangleq \sum_{i=1}^{n} \mathrm{I}\left(X^i; Y_i | Y^{i-1}\right) \tag{21}$$

$$= \mathrm{E}\left[\sum_{i=1}^{n} \log \frac{\mathrm{P}_{Y_i|Y^{i-1},X^i}\left(Y_i|Y^{i-1}, X^i\right)}{\mathrm{P}_{Y_i|Y^{i-1}}\left(Y_i|Y^{i-1}\right)}\right] \tag{22}$$

$$= \sum_{i=1}^{n} \mathrm{D}\left(\mathrm{P}_{Y_i|Y^{i-1},X^i} \,\|\, \mathrm{P}_{Y_i|Y^{i-1}}\right) \tag{23}$$

where Eq. (22) follows from Eq. (7) and Eq. (23) follows from Eq. (4). Alternatively by applying the chain rule for entropy, the directed information can be written as:

$$\mathrm{I}(X^n \to Y^n) = \mathrm{H}(Y^n) - \mathrm{H}(Y^n \| X^n), \tag{24}$$

where the causally conditioned entropy $\mathrm{H}(Y^n \| X^n)$, introduced by Kramer (1998), is defined as:

$$\mathrm{H}(Y^n \| X^n) \triangleq \sum_{i=1}^{n} \mathrm{H}\left(Y_i | Y^{i-1}, X^i\right). \tag{25}$$

The difference between mutual information (Eq. (19)) and directed information (Eq. (22)) is the change from X^n to X^i: the directed information takes into the account the *causal* influence of process X on the current Y_i at each time i.

An important difference between directed information and Granger causality is that directed information itself is a sum of divergences (Eq. (23)) and thus is well-defined for arbitrary joint probability distributions (for example, of point processes (Bremaud 1981; Sundaresan and Verdú 2006)). Moreover, calculation of directed information does not impose any strict probabilistic structure on the (such as an autoregressive model used for Granger causality). Consequently, directed information is more flexible as a metric that can be directly applicable to many modalities, including neural spike trains. As one can determine a "degree of correlation" (statistical interdependence) by computing the mutual information in bits, one can also compute the directed information to determine a "degree of causation" in bits. This quantification allows for an unambiguous interpretation of *how much* Y is statistically causally influenced by X.

4.1.3 Directed information and prediction—Directed information has an important "information gain" interpretation of the divergence with respect to prediction (Cover and Thomas 2006), related to that of mutual information. The mutual information quantifies the expected reduction in the total description cost (Shannon code length) of predicting X and Y separately, as compared to predicting them together (Eq. (5a)). Alternatively,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\begin{aligned} \mathrm{I}(X;Y) &= \mathrm{D}\left(P_{X,Y} \parallel P_X P_Y\right) \\ &= \mathrm{D}\left(P_{Y|X} \parallel P_Y\right) = \mathrm{D}\left(P_{X|Y} \parallel P_X\right), \end{aligned}$$

the mutual information is equivalent to the description penalty of predicting Y with knowledge of X as compared to Y by itself. Using the chain rule (Eq. (6)),

$$\mathrm{I}(X^n;Y^n) = \sum_{i=1}^{n} \mathrm{D}\left(P_{Y_i|X^n,Y^{i-1}}(\cdot) \parallel P_{Y_i|Y^{i-1}}(\cdot)\right),$$

the mutual information between sequences X^n and Y^n (from P_{X^n,Y^n}) measures the total expected reduction in codelength from sequentially predicting (or compressing) the Y^n with full knowledge of the X^n sequence and causal knowledge of the past of Y^n as opposed to just causal knowledge of the past of Y^n.

The directed information has a similar interpretation for prediction with sequences X^n and Y^n (from P_{X^n,Y^n}). It is also a sum of KL-divergences (Eq. (23)):

$$\mathrm{I}(X^n \to X^n) = \sum_{i=1}^{n} \mathrm{D}\left(P_{Y_i|X^i,Y^{i-1}}(\cdot) \parallel P_{Y_i|Y^{i-1}}(\cdot)\right).$$

However, it quantifies the total expected reduction in bits by sequentially encoding Y_i using *causal* side information of both processes, X^i and Y^{i-1}, as compared to encoding Y_i given only Y^{i-1}. This expected log-likelihood ratio follows directly from Granger's original philosophical viewpoint with a Shannon codelength measure of prediction, but differs operationally from Granger's mathematical measure because it operates on arbitrary modalities and statistical models. Other ways of statistically measuring causality from a prediction viewpoint beyond the Shannon code length have recently been discussed in Al-khassaweneh and Aviyente (2008), but for the remainder of this manuscript, we will adhere to the Shannon codelength viewpoint on prediction.

4.1.4 Example of measuring causal influences—To demonstrate that directed information can identify the statistically causal influences between relationships which correlation (as measured by mutual information) cannot, we next present a simple example discussed by Massey and Massey (2005). The example involves two random processes $\mathbf{X} = (X_i : i \geq 0)$ and $\mathbf{Y} = (Y_i : i \geq 1)$ where the X_i random variables are independent, identically distributed (*i.i.d.*) binary (Bernoulli) equiprobable random variables. For $i \geq 1$: let $Y_i = X_{i-1}$, so that $\mathbf{X}$ causally influences $\mathbf{Y}$. Figure 1 depicts the relationship between the processes. Calculating the normalized mutual information between X and Y,

$$\begin{aligned} \tfrac{1}{n}\mathrm{I}(X^n;Y^n) &= \tfrac{1}{n}\mathrm{E}\left\{\log \frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)}\right\} \\ &= \tfrac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n} \log \frac{P_{Y_i|X^n,Y^{i-1}}(Y_i|X^n,Y^{i-1})}{P_{Y_i|Y^{i-1}}(Y_i|Y^{i-1})}\right\} \\ &= \tfrac{1}{n}\left\{\sum_{i=2}^{n} \log 1 - \log\left(\tfrac{1}{2}\right)\right\} \\ &= \tfrac{n-1}{n} \end{aligned} \tag{26}$$

where Eq. (26) follows from Eq. (6), $Y_i = X_{i-1}$, and that X_i's are *i.i.d.*. Taking the limit, $\lim_{n\to\infty} \tfrac{1}{n}\mathrm{I}(X^n;Y^n) = 1$. The mutual information detects a strong relationship, but offers no

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

evidence as to what kind of a relationship it is (is there only influence from one process to another or is there crosstalk?). The normalized directed information from Y to X is

$$\begin{aligned}\frac{1}{n}\mathrm{I}(Y^n \to X^n) &= \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}\log\frac{P_{X_i|Y^i,X^{i-1}}(X_i|Y^i,X^{i-1})}{P_{X_i|X^{i-1}}(X_i|X^{i-1})}\right\}\\ &= \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}\log\frac{P_{X_i|X^{i-1}}(X_i|X^{i-1})}{P_{X_i|X^{i-1}}(X_i|X^{i-1})}\right\}\\ &= 0, \end{aligned} \tag{27}$$

where Eq. (27) holds because $Y_i = X_{i-1}$. The normalized directed information in the reverse direction is:

$$\begin{aligned}\frac{1}{n}\mathrm{I}(X^n \to Y^n) &= \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}\log\frac{P_{Y_i|X^i,Y^{i-1}}(Y_i|X^i,Y^{i-1})}{P_{Y_i|Y^{i-1}}(Y_i|Y^{i-1})}\right\}\\ &= \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}\log\frac{P_{X_{i-1}|X^i}(X_{i-1}|X^i)}{P_{X_{i-1}|X^{i-2}}(X_{i-1}|X^{i-2})}\right\}\\ &= \frac{1}{n}\left\{\sum_{i=2}^{n}\log 1 - \log\left(\frac{1}{2}\right)\right\}\\ &= \frac{n-1}{n}\end{aligned} \tag{28}$$

where Eq. (28) follows because $Y_i = X_{i-1}$, for $i \geq 2$ and the X_is are i.i.d. Therefore, $\lim_{n\to\infty}\frac{1}{n}\mathrm{I}(X^n \to Y^n) = 1$. This example the demonstrates merit of directed information in causal inference as it correctly characterizes the direction of information flow while mutual information fails to do so.

4.2 Interpretations of directed information

Mutual information has a canonical role in a variety of problems (Cover and Thomas 2006). For instance, it characterizes the maximum data rate ("capacity") for reliable communication over a memoryless channel without feedback. Next, we will examine some of the important roles that directed information has in similar settings (in addition to prediction and causal inference).

4.2.1 Communication with feedback—First, consider communication of a message W across a noisy channel using n channel uses, where an encoder maps W to channel inputs X^n and the decoder $d(\cdot)$ receives the channel outputs Y^n to construct an estimate $\hat{\mathrm{W}}$. Under the performance criterion that the error probability (P (() $W \neq \hat{\mathrm{W}}$) tend to 0 in n, the fundamental limit (or capacity) is governed by the maximum possible reduction in uncertainty about the message W given knowledge of the channel outputs, Y^n, which is $I(W; Y^n)$. When the encoder does not have feedback, so that each $X_i = e_i(W)$, then it can be easily shown that $I(W; Y^n) = I(X^n; Y^n)$ (Cover and Thomas 2006). However, if there is causal feedback of the outputs of the channel, then the encoder design paradigm is now $X_i = e_i(W, Y^{i-1})$. See Fig. 2. Here, $\mathrm{I}(W; Y^n)$ can be rewritten as:

$$\begin{aligned}\mathrm{I}(W;Y^n) &= \mathrm{E}\left[\log\frac{P_{Y^n|W}(Y^n|W)}{P_{Y^n}(Y^n)}\right]\\ &= \sum_{i=1}^{n}\mathrm{E}\left[\log\frac{P_{Y_i|W,Y^{i-1}}(Y_i|W,Y^{i-1})}{P_{Y_i|,Y^{i-1}}(Y_i|,Y^{i-1})}\right]\end{aligned} \tag{29}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$= \sum_{i=1}^{n} \mathrm{E}\left[\log \frac{P_{Y_i|W,Y^{i-1},X^i}\left(Y_i|W, Y^{i-1}, X^i\right)}{P_{Y_i|,Y^{i-1}}\left(Y_i|, Y^{i-1}\right)}\right] \tag{30}$$

$$= \sum_{i=1}^{n} \mathrm{E}\left[\log \frac{P_{Y_i|Y^{i-1},X^i}\left(Y_i|Y^{i-1}, X^i\right)}{P_{Y_i|,Y^{i-1}}\left(Y_i|, Y^{i-1}\right)}\right] \tag{31}$$

$$= I\left(X^n \to Y^n\right) \tag{32}$$

where Eq. (29) follows from entropy chain rule of conditional probability; Eq. (30) follows because X^i is a deterministic function of W and Y^{i-1}; Eq. (31) because in complete generality, the statistical nature of the channel output Y_i is linked to W only through the inputs causal X^i and previous outputs Y^i (Cover and Thomas 2006); Eq. (32) follows from Eq. (24). So when the encoder has feedback, $I(W; Y^n)$ $\mathrm{I}(X^n \to Y^n)$ and so the directed information $I(X^n = Y^n)$ replaces the mutual information $I(X^n; Y^n)$ as the fundamental limit of communication over noisy channels with feedback (Kramer 1998; Tatikonda and Mitter 2009; Permuter et al. 2009b).

4.2.2 Other interpretations—Permuter et al. considered directed information in the context of gambling and investment, and showed that directed information can be interpreted as the difference of capital growth rates due to available, causal side information (Permuter et al. 2008, 2009a). Permuter et al. have also investigated the role of directed information in data compression with causal side information and hypothesis testing of whether one sequence statistically causally influences another (Permuter et al. 2009a). Venkataramanan and Pradhan consider the setting of sequential lossy compression (quantization) where the decoder has causal side information about the source and demonstrated that the fundamental limit (the rate-distortion function) is given in terms of the directed information (Venkataramanan and Pradhan 2007). Recently, Kim et al. have demonstrated how the directed information can be interpreted from an optimal causal estimation viewpoint (Kim et al. 2009). Fundamental limits of control when the controller has noisy information about the state of the system have been specified in terms of directed information (Tatikonda 2000; Elia 2004; Martins and Dahleh 2008; Gorantla and Coleman 2010).

5 Estimation

5.1 Previous estimation approaches for information theoretic quantities

For many neuroscientific scenarios of interest pertaining to ensemble-recorded neural signals $\mathbf{X}$ and $\mathbf{Y}$, the underlying joint probability distributions $P_{\mathbf{X},\mathbf{Y}}$ is a priori unknown. Consequently, the normalized information-theoretic quantity (i.e. entropy rate, mutual information rate, etc) cannot be directly computed must be estimated. There are two principled ways of estimating information theoretic quantities (which are functionals of the underlying $P_{\mathbf{X},\mathbf{Y}}$). One approach is to estimate the underlying joint probability distribution $P_{\mathbf{X},\mathbf{Y}}$, and then plug this estimate into the formula—for example, the normalized directed information $\mathrm{I}_n(X \to Y) \triangleq \frac{1}{n}\mathrm{I}(X^n; Y^n)$. Note from Eq. (22) that I_n is a functional on the joint PMF of X^n and Y^n:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$I_n(X \to Y) = g_n\left(\mathrm{P}_{X^n,Y^n}(\cdot,\cdot)\right) = \sum_{i=1}^{n} \mathrm{E}_{\mathrm{P}_{X^n,Y^n}}\left[\log \frac{\mathrm{P}_{Y_i|Y^{i-1},X^i}(Y_i|Y^{i-1},X^i)}{\mathrm{P}_{Y_i|Y^{i-1}}(Y_i|Y^{i-1})}\right].$$

Similar expressions, in terms of functional on PMFs, can be described for entropy, conditional entropy, divergence, and mutual information.

A *plug-in* estimator, first attempts to estimate the density $\mathrm{P}_{X^n,Y^n}(\cdot,\cdot)$. We denote the estimate of the density by $\widehat{P}_{X^n,Y^n}(\cdot,\cdot)$. In general, $\widehat{P}_{X^n,Y^n}(\cdot,\cdot)$ will not be a consistent estimate of $\mathrm{P}_{X^n,Y^n}(\cdot,\cdot)$, as only a single realization of (X^n, Y^n) is observed, and there are $|\mathcal{X} \times \mathcal{Y}|^n$ possible realizations, and a probability estimate needs to be made for each. Consequently, the normalized directed information estimate

$$\widehat{\mathrm{I}}_n(X \to Y) = g_n\left(\widehat{P}_{X^n,Y^n}(\cdot,\cdot)\right)$$

will not be consistent. Note that for i.i.d. processes, there are consistent density estimators, but there are none (known) for general processes (Cesa-Bianchi and Lugosi 2006).

We note that making an *i.i.d.* sample assumption is not sensical within the context of developing measures to understand causal dynamics in random processes. This is because with i.i.d. processes, there is no causation through *time*. Thus, any estimation procedure that relies on i.i.d. assumptions is not applicable to the estimation of the directed information.

More recently, non-parametric procedures have been developed. These procedures attempt to directly estimate the functional on the joint distribution of interest. For information theoretic quantities such as entropy and KL divergence, there are successful universal estimators, including Lempel-Ziv '77 (Ziv and Lempel 1977), the Burroughs-Wheeler Transform (BWT) estimator (Cai et al. 2004), and context weighting tree methods (Cai et al. 2006). Additionally, there has been work extending the context weighting tree method to estimating directed information (Zhao et al. 2010). Unfortunately, these methods are often computationally expensive and have slow convergence rates. There has also been some recent work by Perez-Cruz (2008) for estimating numerous information theoretic quantities with better convergence rates and more moderate computational expense, but these procedures depend on i.i.d. assumptions.

5.2 A consistent direct estimator for the directed information rate

In this section, we propose a consistent estimator for the directed information rate, under some appropriate assumptions that have physical meaning for questions of causality, and are analogous to the canonical i.i.d.-like assumptions for other information-theoretic like quantities.

– **Assumption 1:** $P_{\mathbf{X},Y} \in \mathbf{SE}$ $(\mathcal{X} \times \mathcal{Y})$.

Here, we assume that the random processes X and Y are stationary and ergodic. Under this assumption, as will be seen below, this means that the entropy rate $\mathcal{H}(\mathrm{Y})$, the causal entropy rate $\mathcal{H}(\mathrm{Y}\|\mathrm{X})$, and the directed information rate $\mathcal{I}(\mathrm{X} \to \mathrm{Y})$ all exist. Thus, an estimation procedure can be developed which *separately* estimates the entropy rate and the causal entropy rate, then takes the difference between the two (see Eq. (24)).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Lemma 1 *Let Assumptions* 1. *Let* $P_{\mathbf{X},Y} \in \mathbf{SE} \quad (\mathcal{X} \times \mathcal{Y})$. *Then* $\mathcal{H}(Y)$, $\mathcal{H}(Y\|X)$, and $\mathcal{I}(X \to Y)$ all exist.

The proof is in Appendix A.

– **Assumption 2:** $P_{\mathbf{X},Y} \in \mathcal{M} \quad (\mathcal{X} \times \mathcal{Y})$.

This assumption is the complete analog to the standard *i.i.d.* sample assumption that is used in the simplest of statistical estimation paradigms. Note that by assuming a Markov model, we are incorporating a dynamic coupling, through time, on the processes **X** and **Y** which is physically important for any causal estimation paradigm. The Markov model enables, amongst other things, the strong law of large numbers (SLNN) for Markov chains to hold (Meyn and Tweedie 2009). Many Granger causality, DTF, and other previously discussed estimation procedures assume Markov-like assumptions (Granger 1969; Kaminski and Blinowska 1991; Schreiber 2000), in addition to other constraints.

Lemma 2 *Let Assumptions* 1 *and* 2 *hold, and let* $P_{\mathbf{X},Y} \in \mathcal{M}_{J,K} \quad (\mathcal{X} \times \mathcal{Y})$. *Then for all n*,

$$\frac{1}{n} H(Y^n \| X^n) = E\left[g_{JK}\left(Y_{i-J}^{l}, X_{l-(K-1)}^{l} \right) \right] \tag{33}$$

for the function $g_{JK}\left(a^{J+1}, b^{K}\right) = -\log P_{Y_i | Y_{i-J}^{i-1} X_{i-(J-1)}^{i}}\left(a_{J+1} a_1^J, b_1^K\right)$, where the expectation is taken with respect to the stationary distribution for the Markov chain.

The proof is Appendix B. Since the right hand side of Eq. (33) has no dependence on *n*, taking the limit of the above as $n \to \infty$ results in

$$\begin{aligned} \mathcal{H}(\mathrm{Y}\|\mathrm{X}) &\triangleq \lim_{n\to\infty} \frac{1}{n} \mathrm{H}(Y^n \| X^n) \\ &= \mathrm{E}\left[g_{JK}\left(Y_{i-J}^{l}, X_{l-(K-1)}^{l} \right) \right]. \end{aligned}$$

By exploiting how sample averages converge to ensemble averages with our Markov assumption, we have:

Theorem 1 *Let Assumptions* 1 *and* 2 *hold, and let J satisfy* $P_{\mathbf{X},Y} \in \mathcal{M}_{J,K} \quad (\mathcal{X} \times \mathcal{Y})$. *Then*

$$\frac{1}{n} \sum_{i=1}^{n} g_{JK}\left(Y_{i-J}^{i}, X_{i-(K-1)}^{i} \right) \xrightarrow{a.s.} \mathcal{H}(Y\|X)$$

Proof Taking the limit on both sides of Eq. (33) as $n \to \infty$,

$$\mathcal{H}(\mathrm{Y}\|\mathrm{X}) = \mathrm{E}\left[g_{JK}\left(Y_{i-J}^{l}, X_{l-(K-1)}^{l} \right) \right].$$

Using the SLLN for Markov chains (Meyn and Tweedie 2009), for a fixed function $g(\cdot)$ over the states of the Markov chain, as $n \to \infty$, the sample mean will converge almost surely to the expected value:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\frac{1}{n}\sum_{i=1}^{n} g_{JK}\left(Y_{i-J}^{i}, X_{i-(K-1)}^{i}\right) \xrightarrow{a.s.} \mathrm{E}\left[g_{JK}\left(Y_{i-J}^{i}, X_{i-(K-1)}^{i}\right)\right] = \mathcal{H}(\mathrm{Y}\|\mathrm{X}).$$

With these results, if a consistent estimate $\widehat{g}_{JK}(\cdot)$ for the function $g_{JK}(\cdot)$ can be found, then the sample mean of this function will converge almost surely to the causal entropy rate $\mathcal{H}(\mathrm{Y})$, and thus directed information rate can be estimated with almost sure convergence. Note that if Y alone forms a discrete-time, finite state, stationary, and ergodic Markov chain, then this result can be used to estimate $\mathcal{H}(\mathrm{Y})$ by taking X to be a known, deterministic process.

– **Assumption 3: For point processes $X \in \mathcal{Y}_T$ and $Y \in \mathcal{Y}_T$ and a pre-specified set of functions $\{h_k : k \geq 0\}$, $\lambda(i \,\|\, \mathcal{F}_i) \in$ GLM** (h).

The recorded neural spiking activity—in millisecond time resolution—is known to be well-modeled using point process theory (Truccolo et al. 2005). Because of the duration of a neural spike and its refractory period, we will partition continuous time into $\Delta = 1$ millisecond time bins, and denote $dy_i = 1$ if a neural spike occurs within it, and 0 otherwise. Generalized linear models (GLM) for point processes (Truccolo et al. 2005) are a flexible class of parametric point process neural spiking models that allows for dependencies on a neuron's own past spiking, the spiking of other neurons, and ex trinsic covariates. GLM models have the following conditional intensity:

$$\log \lambda(i \,\|\, \mathcal{F}_i) = \alpha_0 + \sum_{j=1}^{J} \alpha_j dy_{i-j} + \sum_{k=1}^{K} \beta_k h_k\left(x_{i-(k-1)}\right) \tag{34}$$

where $h_k(\cdot)$ is some function of the extrinsic covariate, and

$$\theta = \{\alpha_0, \alpha_1, \cdots, \alpha_J, \beta_1, \cdots, \beta_K\}$$

is the parameter vector. Note that with such a GLM model, from Theorem 1, we have:

$$-\frac{1}{n}\log f_{Y\|X}\left(Y_1^n \| X_1^n; \theta\right) = \frac{1}{n}\sum_{i=1}^{n} -\left(\log\left(\lambda_\theta(i|\mathcal{H}_i)\right) dy_i - \lambda_\theta(i|\mathcal{H}_i)\,\Delta\right) \tag{35}$$

$$= \frac{1}{n}\sum_{i=1}^{n} g_\theta\left(Y_{i-J}^{i}, X_{i-(K-1)}^{i}\right) \xrightarrow{a.s.} \mathrm{E}\left[g_\theta\left(Y_{i-J}^{i}, X_{i-(K-1)}^{i}\right)\right] = \mathcal{H}(\mathrm{Y}\|\mathrm{X}) \tag{36}$$

where Eq. (36) shows that the estimate is a sample mean of a *fixed* function (independent of i) of the data. Note that any probabilistic model (parametric or nonparametric) could be used to estimated the directed information, not just GLM.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

5.3 Parameterized estimation and MDL

Define $\Omega(J, K)$ to be vector space of possible parameters $\theta = \{\alpha_0, \alpha_1, \cdots, \alpha_J, \beta_1, \cdots, \beta_K\}$. If it is known a priori that $\lambda\ (i \parallel \mathcal{F}_i) \in \quad \mathrm{GLM}_{J,K} \quad (h)$, then θ_0 can be consistently estimated using Assumptions 1–3 and a maximum likelihood estimate (MLE) (Casella et al. 2002):

$$\begin{aligned}\widehat{\theta}(J,K) &= \arg\min_{\theta\in\Omega(J,K)} -\tfrac{1}{n}\log\ f_{Y\|X}\left(Y_1^n\|X_1^n;\theta\right)\\ &= \arg\min_{\theta\in\Omega(J,K)} \tfrac{1}{n}\sum_{i=1}^{n} g_\theta\left(Y_{i-J}^i, X_{i-(K-1)}^i\right).\end{aligned}$$

In practice, J_0 and K_0 are unknown. A model order selection procedure can be used to find estimates $\hat{J}$, $\hat{K}$, and subsequently $\widehat{\theta} \in \Omega\left(\widehat{J}, \widehat{K}\right)$ by penalizing "more complex" models, that is —those with larger $J + K$ values. The minimum description length (MDL) (Grünwald and Rissanen 2007) is a model order selection procedure, which is known to be have strong consistency guarantees (Barron and Cover 1991). In particular, under the assumption that $\lambda\ (i \parallel \mathcal{F}_i) \in \quad \mathrm{GLM} \quad (h)$—which means that $\theta_0 \in \Omega(J_0, K_0)$, for some J_0 and K_0, then it can be shown that an appropriately designed estimate $\widehat{\theta} \to \theta_0\ a.s.$. Specifically, MDL selects the $(\hat{J}, \hat{K})$ and $\widehat{\theta} \in \Omega\left(\widehat{J}, \widehat{K}\right)$ according to

$$\begin{aligned}\left(\widehat{J},\widehat{K}\right) &= \arg\min_{(J,K)} \min_{\theta\in\Omega(J,K)} -\tfrac{1}{n}\log\ f_{Y\|X}\left(Y_1^n\|X_1^n;\theta\right) + \tfrac{J+K}{2n}\log n\\ &= \arg\min_{(J,K)} \min_{\theta\in\Omega(J,K)} \tfrac{1}{n}\sum_{i=1}^{n} g_\theta\left(Y_{i-J}^i, X_{i-(K-1)}^i\right) + \tfrac{J+K}{2n}\log n\\ \widehat{\theta} &= \widehat{\theta}\left(\widehat{J},\widehat{K}\right)\end{aligned} \tag{37}$$

As K is the number of extrinsic parameters, if $\hat{K} = 0$, then *we say that no causal influence was detected*, since $\widehat{H}(\mathrm{Y}\|\mathrm{X}) = \widehat{H}(\mathrm{Y})$ which implies that $\widehat{I}(\mathbf{X} \to \boldsymbol{Y}) = 0$. Thus, to determine whether there is a detected causal influence or not does not require computation of the directed information; only the $\hat{K}$ from the best-fitting model is necessary. If $\hat{K} = 0$, there is no detected influence $\left(\widehat{I}(\mathrm{X} \to \mathrm{Y}) = 0\right)$. If $\hat{K} > 0$, there is a detected influence $\left(\widehat{I}(\mathrm{X} \to \mathrm{Y}) > 0\right)$.

Although one can identify whether there is a detected causal influence without computing the directed information, the *extent* of an influence cannot be determined by the GLM model alone. Directed information considers both the model and the data to determine the influence. An example which illustrates this point is as follows. Let A and B be two neurons, such that whenever B spikes, A will spike with probability 1 within each of the next 12 ms except when A has just fired (refractory period). Let A have a large average spiking rate, such as one spike per 10 ms, and let B have a very low average spiking rate, such as one spike per second (see Fig. 3).

The best fitting GLM model (provided the data recording is sufficiently long) of neuron A using neuron B as B as an extrinsic process will have $\hat{K} \approx 12$ and $\{\beta_1, \cdots, \beta_{\hat{K}}\}$ large and Thus, it would seem, from the GLM model alone, that B strongly influences A. However, since there are few instances where B spikes, few of A's spikes are caused by B's, and so B will have a small, causal influence influence on A. If B has a much larger firing rate, however, then many more of A's spikes could statistically be explained by B's spikes (if the β parameters remain the same), and thus B would have a larger, causal influence. Changes in the data, with a fixed model, can result in changes in the extent of the influence. Thus, directed information, which considers both, is able to measure the extent of the influence, which the model alone cannot.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

5.4 The proposed estimation procedure

Under the Assumptions 1–3, we provide the following consistent estimation procedure:

1. Find $\hat{J}$, $\hat{K}$, and $\widehat{\theta}$ according to the MDL procedure Eq. (37).
2. Calculate $\widehat{\mathcal{H}}(\mathrm{Y} \parallel \mathrm{X})$ according to Eq. (36) using the estimated parameter values $\widehat{\theta} \in \Omega\left(\widehat{J}, \widehat{K}\right)$.
3. Compute an estimate for the unconditional entropy rate $\widehat{H}(\mathrm{Y})$ using a well-established entropy estimator (such as Lempel-Ziv '77 (Ziv and Lempel 1977) or the BWT based estimator (Cai et al. 2004)).
4. Calculate the directed information rate estimate

$$\widehat{\mathcal{I}}(\mathrm{X} \to \mathrm{Y}) \triangleq \widehat{\mathcal{H}}(\mathrm{Y}) - \widehat{\mathcal{H}}(\mathrm{Y} \| \mathrm{X})$$

Theorem 2 *If Assumptions* 1, 2, *and* 3 *hold, then*

$$\widehat{\mathcal{I}}(X \to Y) \xrightarrow{a.s.} \mathcal{I}(X \to Y) \tag{38}$$

Proof

1. If Assumptions 1–3 hold, then the MDL procedure will identify the "true" parameter values $\theta \in \Omega(J, K)$ (Barron and Cover 1991): $\hat{J} \to J$ *a.s.*, $\hat{K} \to K$ *a.s.*, and $\widehat{\theta} \to \theta$ *a.s.*.
2. Note that since $\widehat{\theta} \to \theta$ *a.s.*, from the continuity of $g_\theta \widehat{\mathcal{H}}(\mathrm{Y} \parallel \mathrm{X})$ specified above satisfies $\widehat{\mathcal{H}}(\mathrm{Y} \parallel \mathrm{X}) \to \mathcal{H}(\mathrm{Y} \| \mathrm{X})$ *a.s.* by virtue of Theorem 1.
3. Universal estimators such as Lempel Ziv '77 and the BWT based estimator converge almost surely to the unconditional entropy rate $\mathcal{H}(\mathrm{Y})$ for stationary and ergodic finite-order Markov processes (Lastras 2002; Cai et al. 2004).
4. Combining these results,

$$\widehat{\mathcal{I}}(\mathrm{X} \to \mathrm{Y}) \triangleq \widehat{\mathcal{H}}(\mathrm{Y}) - \widehat{\mathcal{H}}(\mathrm{Y} \| \mathrm{X}) \xrightarrow{a.s.} \mathcal{H}(\mathrm{Y}) - \mathcal{H}(\mathrm{Y} \| \mathrm{X}) = \mathcal{I}(\mathrm{X} \to \mathrm{Y})$$

5.5 Implementation details

To perform the MDL search procedure, we examine values of $J, K \in \{0, 1, \cdots, M\}$, where M is a user-specified maximum value. M should be chosen to be sufficiently large that any causal influences of interest in the data occur within the timescale of $M\Delta$. However, if the best-fitting models have and/or $\hat{K}$ near M, then M can be increased adaptively to search for larger parameter orders (thus, it is not a *hard* limit). M is chosen a priori to save computation for when the procedure settles on small values for $\hat{J}$ and $\hat{K}$. More precisely, when local communications within a small brain region is of interest, picking a relatively small M is sufficient (Truccolo et al. 2005). If we anticipate that an upper bound for the maximum time scale for a spike from one neuron to influence another neuron (including time to propagate) is around 25 ms (Vogels and Abbott 2005), then it would be appropriate to pick an $M \approx 25$. However, in the case of motor feedback, e.g. hand movement, the longer delays for the signal to propagate should be taken into account, and a larger M, such as $M \approx 150$ should be selected (Paninski et al. 2004).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

For each (*J, K*), the MLE parameter vector $\widehat{\theta}(J,K)$ can be computed using the built-in Matlab function *glmfit*(·), called with a Poisson link parameter. Then Eq. (37) is computed to determine $\widehat{\theta}$. The estimate for the causal entropy rate is taken to be the sample mean:

$$\widehat{\mathcal{H}}(\mathrm{Y}\|\mathrm{X}) = \frac{1}{n}\sum_{i=1}^{n} g_{\widehat{\theta}}\left(Y^i_{i-\vec{J}}, X^i_{i-(\vec{K}-1)}\right).$$

To compute an estimate of the entropy rate, $\widehat{\mathcal{H}}(\mathrm{Y})$, a universal estimator such as the BWT based estimator could be used (which has a faster convergence rate than LZ '77) (Cai et al. 2004). Alternatively, the above procedure could be used with $K = 0$ fixed. Through trials with large neural data binary time series (on the order of 100,000 bins), the values were quite close, and obtained quicker than with the universal estimator. The difference between the two estimates, $\widehat{\mathcal{H}}(\mathrm{Y}) - \widehat{\mathcal{H}}(\mathrm{Y}\|\mathrm{X})$, then becomes the directed information estimate, $\widehat{\mathcal{I}}(\mathrm{X} \to \mathrm{Y})$.

In some cases, the relative influence of a process X on a process Y is of interest. The *normalized directed information rate* can be computed by normalizing the directed information by the entropy rate of the process Y:

$$\frac{\widehat{\mathcal{I}}(\mathrm{X} \to \mathrm{Y})}{\widehat{\mathcal{H}}(\mathrm{Y})} = \frac{\widehat{\mathcal{H}}(\mathrm{Y}) - \widehat{\mathcal{H}}(\mathrm{Y}\|\mathrm{X})}{\widehat{\mathcal{H}}(\mathrm{Y})} = 1 - \frac{\widehat{\mathcal{H}}(\mathrm{Y}\|\mathrm{X})}{\widehat{\mathcal{H}}(\mathrm{Y})}. \quad (39)$$

For values of this quantity close to 1, X can be interpreted as having a strong causal influence on Y, and for values close to 0, X can be interpreted as having a weak causal influence on Y.

In addition to the bound on the model order search space, *M*, there is another design choice to be made before running the procedure, that of the time resolution Δ. The GLM framework which is used for modeling depends on having binary time series, such that the data can be modeled as a point process (Truccolo et al. 2005). It has been found that Δ = 1 ms is a sufficiently small time window, such that using this resolution will result in binary data (no more than one spike in that time window) (Truccolo et al. 2005). However, such resolution is not necessary for the point-process-GLM framework; all that is necessary for the modeling is that the temporal resolution is small enough that the data is binary (Truccolo et al. 2005).

There are both potential benefits and harms to choosing Δ > 1 ms. One benefit of choosing Δ > is that there is a reduction of the length of data (number of bins), which can increase the speed of the procedure. Another potential benefit is that the fits could be better. The procedure finds the best fitting α and β parameters. Choosing a larger Δ will cause the data to become less zeros). There could then be more instances of multiple spikes within the $M\Delta$ time window to fit the window to fit the α's and β's. Also, for a fixed upper bound on the time scale over which take place, increasing Δ will decrease the corresponding *M* to ensure that the maximum time scale searched over, $M\Delta$, is large enough. The smaller search space would increase the speed of the procedure.

One possible problem of choosing Δ to be larger, such as 2, 5, or 10 ms, is that there is a potential loss of timing information which can effect detection of causal influences. For example, consider two neurons, A and B, such that whenever A fires, B fires 3 ms afterwards with very high probability. Also, let A and B have very low average firing rates,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

so with 1ms time resolution, there are many more 0's than 1's. While using $\Delta = 10$ ms might result in A and B having binary spike trains (so the framework can still be applied), it is possible that the spikes from A and the corresponding spikes from B will be grouped in the same time bin. They will then appear to have occurred simultaneously, instead of B firing with a firing with a slight delay. The loss of relevant timing information such as this could effect how well the detect the underlying influences.

Issues such as the aforementioned problem could potentially be screened beforehand, to determine if both the time differences between spikes of the different neurons and the time differences between spikes of the neuron are smaller than the proposed Δ. However, the authors are not aware of any study which has compared how different choices of sufficiently small values of Δ (sufficient so that the data is binary) correspond to differences in how well the best-fitting models (for a given Δ) compare with others. There is no known general procedure for deciding the best value of δ.

Computation time might be a factor in deciding the maximum model order M, the time resolution Δ, and the amount of data to use. Our simulations were designed and tested in the Matlab environment, using built-in Matlab functions (the code is available upon request). The primary computational bottleneck is finding the α and β parameters for a given J and K, which were computed with *glmfit*(·). In our tests (see Section 7), we used datasets on the order of 100,000 elements. Trials ran on computers with a 2.6 GHz processor (each estimate of $\widehat{I}(\mathrm{X} \to \mathrm{Y})$ ran on a single computer). A single *glmfit*(·) operation with $1 \leq J, K \leq 5$ took a few seconds. A single *glmfit*(·) operation with $20 \leq J, K \leq 25$ took upwards of 2 min. The proposed procedure involves searching over a larger (J, K) parameter space for each ordered pair of processes. If $M = 25$, then there are $M^2 = 625$ calls to *glmfit*(·). For each ordered pair, this search took approximately 2 h. With six processes total, there are 6*5=30 ordered pairs. The total procedure took about 2 and a half days for each directed information estimate. For causally conditioned directed information estimates (going beyond pairwise estimates; see Section 6), the search space increased. For causally conditioning on two elements, the search space involves $M^4 \approx 400{,}000$ calls to *glmfit*(·). Since the runtime for *glmfit*(·) changes with model order, the total time does not scale multiplicatively. The computations for different (J, K) orders can be done in parallel. It is possible that other implementations of the GLM model fitting (a convex optimization procedure) could be faster than the Matlab implementation, thus reducing computation times.

5.6 Confidence intervals

To obtain confidence intervals on the directed information estimates, sensitivity analysis using the Fisher Information is used. Once the observed data is fixed (a given spike train $y \in \mathcal{Y}_T$, possibly with an extrinsic spike train $x \in \mathcal{Y}_T$), the directed information estimate is a function only of the estimated parameters $\widehat{\theta} \in \Omega\left(\widehat{J}, \widehat{K}\right)$. We here perform a sensitivity analysis to characterize how much the directed information estimate changes as a function of the parameter values used, in the neighborhood of the original parameters $\widehat{\theta}$. The variation in estimate values is then taken into account by specifying a confidence interval.

For this particular estimation problem, since for fixed ($\hat{J}, \hat{K}$), the search for the best fitting model is a MLE problem, and, in particular, since the probability class being considered (point process GLMs) are convex in the parameters, the MLE will be the global maximum of the probability function

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$f_{Y\|X}(y\|x;\theta) = \exp\left(\sum_{i=1}^{T/\Delta} \log \lambda \quad (i \,\|\, \mathcal{F}_i))\, dy_i - \lambda \quad (i \,\|\, \mathcal{F}_i)\, \Delta\right)$$

over the space of parameter values $\Omega(\hat{J}, \hat{K})$ (Casella et al. 2002). Under appropriate aforementioned assumptions that guarantee consistency, the global maximum converges to the true model almost surely. With a finite amount of data, we use the curvature of the likelihood function in the neighborhood—*observed Fisher information*—to estimate a 95% confidence interval on the directed information. The observed Fisher information matrix, denoted as $I(z, \theta)$, where z denotes the data and θ the parameter values, is defined as the second derivative (or Hessian) of the negative log likelihood, with respect to the parameter values. Analogous to approximating a continuous function using a Taylor series approximation, one can approximate the probability density function near the global maximum with a gaussian distribution, with a mean value at the global maximum, and a covariance matrix $I(z, \theta)^{-1}$ (Casella et al. 2002):

$$f_Z(z) \approx \mathcal{N}\left(\theta_{MLE}, I(z,\theta)^{-1}\right)$$

in the neighborhood of θ_{MLE}. Using this approximation, an approximate 95% confidence interval for the picked θ_{MLE} (interpreted as an interval about θ_{MLE} that with 95% probability contains the true parameter θ_0) (Casella et al. 2002):

$$\theta_{MLE} \pm \frac{1.96}{\sqrt{I(z,\theta)}}. \tag{40}$$

For the purposes of this problem, since the parameters of interest are those corresponding to whether or not there is statistically causal influence, the β_is, assume that only the β_is from the best fitting model might vary from those of the true model. Assume that $\hat{J}$, $\hat{K}$, and $(\widehat{\alpha}_j : 1 \le j \le J)$ are correct. To find a confidence interval on any particular parameter β_k, consider second order partial derivatives of the form $\frac{\partial^2}{\partial \beta_k^2}$. For each $l \in \left\{1, \cdots, \widehat{K}\right\}$, compute the $(l, l)^{th}$ entry of the observed Fisher information matrix:

$$I_{\text{Fisher}}\left(dy^n, dx^n; \widehat{\theta}\right)_{l,l} = -\frac{\partial^2}{\partial \beta_l^2}\left[\sum_{i=1}^{n} \log\left(\lambda\left(i|\mathcal{H}_i\right)\right) dy_i - \lambda\ (i \,\|\, \mathcal{F}_i)\, \Delta\right]\Bigg|_{\widehat{\theta}} = \sum_{i=1}^{n} \left(dx_{i-(l-1)}\right)^2 e^{\left(\widehat{\alpha}_0 + \sum_{j=1}^{\widehat{J}} \widehat{\alpha}_j dy_{i-j} + \sum_{k=1}^{\widehat{K}} \widehat{\beta}_k dx_{i-(k-1)}\right)} \Delta.$$

With this value, the 95% confidence interval for this parameter $\widehat{\beta}_l$ can be calculated using Eq. (40). When the confidence intervals for all the $\widehat{\beta}_l$ parameters are determined, then the region of parameter values where $\widehat{\alpha}_j$s are the same as the best fitting model, and the $\widehat{\beta}_l$s are within the Once the maximum variations from the original directed information estimate values are identified, they can be considered to be the corresponding bounds of the 95% confidence interval for the directed information estimate.

6 Causal relationships in a network of processes

Although there might be situations where researchers are primarily interested in whether one process "causes" another, there are many situations in neuroscience as well as communications, economics, social sciences, and other fields were researchers want to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

identify the causal relationships in a *network* of processes. For example, an electrode array recording of a brain section might detect the spike trains of 50 neurons, and the researcher might be interested in which of the neurons causally influence other neurons. In particular, the researcher might be interested in identifying the *direct*, causal influences (as opposed to indirect influences through other recorded neurons).

Researchers have already begun investigating the problem of identifying causal relationships in neuronal networks (Eguiluz et al. 2005; Goebel et al. 2003; Hesse et al. 2003; Okatan et al. 2005; Uddin et al. 2009; Kramer et al. 2009; Ramnani et al. 2004; Smith et al. 2006; Seth and Edelman 2007; Stevenson et al. 2009). As the framework presented in this paper for measuring causal inferences is principally different than previous, known research, the previous approaches are not directly applicable. There recently has been research on using directed information to infer causality graphs for neuroscience (Amblard and Michel 2010), but they do not propose an estimation scheme and their conditions for estimating whether there is a direct, causal influence or not is different than the definitions in the paper. More relevant to this paper is research on Bayesian networks (Pearl 2009). Bayesian networks, or "belief networks," define causality between random variables by using properties of the joint distribution (Pearl 2009). There is also a corresponding graphical depiction of the network using a directed, acyclic graph. Note, however, that causality as defined by Bayesian networks is not philosophically consistent with Granger's definition. The elements of Bayesian networks are random variables, so there is no sense of time or prediction. This work is concerned with the causal relationships between random *processes*, where there is a sense of time. Thus, the methods and definitions developed for Bayesian networks cannot be directly applied. However, some of the underlying ideas are related to the related to the methods and definitions for networks of random processes presented here. This section will define causal influences in the context of networks of random processes and introduce graphical structures to represent these influences.

6.1 Causal conditioning and direct, causal conditioning

Define the *causal conditioning* of a length n random process B on the marginal distribution of another length n random process A to be

$$\mathrm{P}_{\mathrm{A}\|\mathrm{B}}(\cdot)=\mathrm{P}_{A_1^n\|B_1^n}(\cdot)\triangleq\prod_{i=1}^{n}\mathrm{P}_{A_i|A^{i-1},B^i}(\cdot) \tag{41}$$

Define causal influences as follows. Let V be a set of $m + 1$ random processes, $\mathrm{V} = \{\mathrm{X}_1, \cdots, X_m, \mathrm{Y}\}$, where each process is a length n vector, $\forall \quad \mathrm{Z} \in \mathrm{V}, \mathrm{Z}=(Z_i)_{i=1}^n$.

Definition 1 The random process X_i is said to *causally influence* the random process Y if:

$$\mathrm{P}_{\mathrm{Y}\|\mathrm{X}_i}(\cdot)\neq\mathrm{P}_{\mathrm{Y}}(\cdot) \tag{42}$$

Note that this definition only identifies if there is influence through *some* path, possibly This form of influence will also be denoted as "pairwise" influence, since it is from one process to another. In many circumstances, causal influences can be fully explained by paths of causal influence through other processes, without any "direct" influence.

Definition 2 The random process X_i is said to *directly*, causally influence the random process Y with respect to V if:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\forall \quad W \subseteq V = \{X_i, Y\} \quad P_{Y\|W,X_i}(\cdot) \neq P_{Y\|W}(\cdot) \tag{43}$$

Thus, even with causal knowledge of any of the other processes in the network, there is still some influence influence from X_i to Y. Here, the "directness" of an influence is only with respect to the known processes V. For example, in an electrode array recording of neurons, there could be many undetected neurons which greatly influence the recorded ones. It might even be the case that none of the recorded ones have direct, physical connections, but instead all go through other, unrecorded neurons. Thus, the meaning of "direct" in this context is statistical, and if no subset of the other, known processes (recorded neurons) can explain statistically the influence of one process X_i on another Y, then it is said that X_i has a direct influence on Y. These conditions are related to the conditions of "d-separation" in Bayesian networks (Pearl 2009). Let V_Y denote the set of all the Y. Let V'_Y denote the set of all the processes that causally influence Y. By the above definitions,

$$V_Y \subseteq V'_Y.$$

The set of direct, causal influences amongst processes in set V is a subset of the causal influences amongst processes V.

6.2 Graphical depiction and indirect influences

Bayesian networks and other approaches to identifying causal relationships in networks often use directed graphs to depict the relationships (Pearl 2009). They can be used here as well. Let each of the processes in V be represented as a node. Let there be a solid arrow from process X_i to process X_j ($i \neq j$) iff X_i directly, causally influences X_j. Otherwise, let there be no arrow. An example is shown below for processes A, B, C, D, and E is Fig. 4.

A similar representation for causal influences in a network (that is, not just those which are direct) will be used. Let there be a long-dashed arrow from process X_i to process X_j ($i \neq j$) iff X_i causally influences X_j. Otherwise, let there be no arrow. An example is shown below for processes A, B, C, D, and E, which is consistent with the above graph for the direct influences is Fig. 5. It is consistent because all of the direct, causal influences are present, and the extra arrows could be due to indirect influences, which are discussed below.

Two types of indirect influences which result in more causal influences than direct, causal influences will be denoted as "proxy" and "cascading" influences. In a proxy influence, process X_1 influences process X_2 which in turn influences X_3, but with no direct influence from X_1 to X_3. In some cases, there will be a causal influence from X_1 to X_3 *through* X_2 (Fig. 6), and causal knowledge of X_2 renders X_1 and X_3 statistically independent. Thus, proxy effects can be considered analogous to the analogous to the Markovicity property. Note that if there is a loop of direct, causal influence between a set of processes (such as $X_1 \rightarrow X_2$, $X_2 \rightarrow X_3$, $X_3 \rightarrow X_4$, and $X_4 \rightarrow X_1$), then the causal influences from every process to all the others, due to proxy effects.

Another form of indirect influence is "cascading" influence. Here two processes X_2 and X_3 have a common influencing process X_1. Knowledge of X_1 renders X_2 and X_3 statistically independent, but there is causal influence between the two possibly accounted for by residual self dependence in X_1 (Fig. 6).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

6.3 Causal conditioning and directed information

The definitions of causal influences and direct, causal influences can be used to establish related conditions using causally conditioned directed information.

Theorem 1 *The process* X_i causally influences *the process Y if and only f* $I(X_i \to Y) > 0$.

Proof That causal influence implies positive directed information is proven as follows. $P_{Y\|X_i}(\cdot) \neq P_Y(\cdot)$ by definition. Recall that the KL distance is 0 if and only if $P_{Y\|X_i}(\cdot) \neq P_Y(\cdot)$. Secondly, note from Eq. (23) that

$$\mathrm{I}(\mathrm{X}_i \to \mathrm{Y}) = \sum_{j=1}^{n} \mathrm{D}\left(P_{Y_j|Y^{j-1},X_{i,1},\ldots,X_{i,j}} \,\|\, P_{Y_j|Y^{j-1}}\right).$$

The definition of direct, causal influences can also be extended to conditions of directed information. The conditions will require causally conditioning on extrinsic processes. Kramer introduced *causally conditioned directed information* for a process X_i, process Y, and set of processes W as (Kramer 1998):

$$\mathrm{I}(\mathrm{X}_i \to \mathrm{Y}\|\mathrm{W}) \triangleq \mathrm{H}(\mathrm{Y}\|\mathrm{W}) - \mathrm{H}(\mathrm{Y}\|\mathrm{X}_i, \mathrm{W}) \tag{44}$$

Lemma 2 $P_{Y\|W,X_i}(\cdot) \neq P_{Y\|W}(\cdot)$ *if and only if* $I(X_i \to Y\|W) > 0$.

The proof is identical to the proof of the above theorem but causally conditioning on the set of processes W.

Theorem 3 *The random process* X_i directly, causally influences the random process Y with respect to V iff:

$$\forall \quad W \subseteq V = \{X_i, Y\} \quad I(X_i \to Y\|W) > 0 \tag{45}$$

The proof here follows from Lemma 2 and the definition of direct, causal influences.

6.4 Identifying the direct, causal influences in a network of processes

Identification of all of the causal influences in a network of processes V is straightforward by the definition. For each ordered pair of distinct processes $(\mathrm{X}_i, \mathrm{X}_j)$, compute $\mathrm{I}(\mathrm{X}_i \to \mathrm{X}_j)$. If the value is positive, then there is causal influence from X_i to X_j, or $\mathrm{X}_i \to \mathrm{X}_j$. Otherwise, there is no causal influence.

Identificaton of all the direct, causal influences in a network of processes is more complicated, as there are more conditions to check than for causal influences. Since every direct, causal influence is causal influence, one could first identify all of the causal influences, and then determine which of those of those were also direct, causal influences. Consider two processes in V, X_i and Y, such that $\mathrm{I}(\mathrm{X}_i \to \mathrm{Y}) > 0$. Thus, X_i causally influences Y. To determine if X_i directly, causally influences Y, one could check that for each $\mathrm{W} \subseteq \mathrm{V} = \{\mathrm{X}_i, \mathrm{Y}\}$, $\mathrm{I}(\mathrm{X}_i \to \mathrm{Y}\|\mathrm{W}) > 0$. If so, then X_i directly, causally influences Y, else if there is even one such W for which $\mathrm{I}(\mathrm{X}_i \to \mathrm{Y}\|\mathrm{W}) = 0$, then the influence is not direct. Since some processes are statistically independent of X_i and/or Y, it can be helpful to focus on the subsets W which contain those X_j's such that each X_j causally influences Y and causally influences or is influenced by there is a causal subgraph that could contain a proxy or cascading influence, to check those first.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

7 Results

7.1 Simulated data

To test the effectiveness of this estimation procedure, it was applied to simulated data. A small network of six binary processes, modeled as neuronal spike trains, was simulated. Each process will be referred to as a "neuron," and is labeled with a letter between "A" and "F." Twenty independent samples of the network were randomly generated using the same values and procedure. Point process GLM models were used to generate the spike trains. For fixed values of the model orders J and K, the conditional intensity functions were selected according to $\lambda\ (i \parallel \mathcal{F}_i) \in \mathrm{GLM}_{J,K} \quad (h)$, where $(h_k : 1 \le k \le K)$ were all the identity function. The values of the parameters were selected to be within the range of parameters (J, K, and α, β values) previously identified in point process GLM model fits to spike trains from electrode array recording data of goldfish retinal ganglia (Iyengar and Liao 1997) and primate primary motor cortex (Wu and Hatsopoulos 2006). In particular, $3 \le J \le 20$, $0 \le K \le 20$, $-10 \le \alpha_i, \beta_j \le 10$. The time width $\Delta = 1$ ms was used, and 160,000 ms of data were generated. Once the data and experimental design parameters were determined, the time series for each neuron was obtained by generating a sequence of i.i.d. unit rate exponentials and inverting the time-rescaling theorem (Brown et al. 2002).

The designed influence structure, or the "functional topology," is shown in Fig. 7. An arrow from neuron X to neuron Y depicts that during the generation of Y's spike train, the spike train of X was used as an extrinsic covariate (thus X directly, causally influences Y). The β_is were either positive, corresponding to an excitatory influence, or negative, corresponding to an inhibitory influence. An arrow from neuron Y to neuron Y depicts autoregressive influence, such that at time step i, the recent past of Y's spike train (beyond a 2–3 ms refractory period) influenced the present. The absence of an arrow from neuron X to neuron Y depicts that the spike train of neuron X was not used as an extrinsic covariate in generating the spike train of Y. Note that some of the neurons, in particular E and F, both have two arrows from two other neurons. For these, two sets of extrinsic covariates were used when calculating the conditional intensity function:

$$\log\left(\lambda\left(i|\mathcal{H}_i\right)\right) = \alpha_0 + \sum_{j=1}^{J} \alpha_j dy_{i-j} + \sum_{k_1=1}^{K_1} \beta_{k_1} dx_{1;i-(k_1-1)} + \sum_{k_2=1}^{K_2} \beta_{k_2} dx_{2;i-(k_{2-1})},$$

where $dx^1_{i-(k_1-1)}$ corresponds to the $i - (k_1 - 1)^{th}$ value of the first extrinsic spike train.

As an example of the selected parameters, neuron F, which was influenced by C and D (inhibitory and excitatory respectively), was set to have constant firing rate $\alpha_0 = 1.8$, $J = 3$, $K_C = 5$, $K_D = 7$,

$$\begin{aligned} \{\alpha_1, \alpha_2, \alpha_3\} &= \{-7.8, -5.5, -3.4\} \\ \left\{\beta_1^C, \cdots, \beta_5^C\right\} &= \{-8.1, -5.8, -4.4, -4.1, -2.1\} \\ \left\{\beta_1^D, \cdots, \beta_7^D\right\} &= \{0.15, 0.9, 3.8, 5.1, 4.7, 2.7, 1.1\} \end{aligned}$$

A sample of the time series for neurons C, F, and D respectively are shown in Fig. 8.

After the data was generated for each of the 20 samples, the estimation algorithm described in the previous section was used for each sample, using Matlab (code available upon request). No knowledge of the parameters for generating the data was used in the estimation procedure. First all of the pairwise directed information rates, $\widehat{I}\,(\mathrm{X} \to \mathrm{Y})$, were computed.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

All $0 \leq, J, K \leq 15$ were examined. None of the design parameter values were more than 10, and none of the estimated $\hat{J}$ or $\hat{K}$ were larger than 12, so increasing the range would not have effected the procedure. If any of the $\hat{J}$ or $\hat{K}$ were near 15, then the range for J and K examined would have been increased. The pairwise directed information estimates were then normalized with the respective unconditional entropy estimates $\widehat{H}(\mathrm{Y})$, which were found using the same procedure with $K = 0$. The same ordered pairs (X, Y) were estimated as having nonzero directed information rates across all 20 samples, and all of the other ordered pairs were estimated as having zero directed information rate in all the samples (thus, the same structures were found for each sample). Figure 9 shows the averaged normalized estimates (see Eq. (39)) for all of the nonzero values with averaged normalized 95% confidence intervals. The averages were taken over the 20 samples. The empirical standard deviations for the estimated rate values (across the samples) were between 0.001 and 0.007 for each of the nonzero estimated rates. The empirical standard deviations for the confidence intervals (across the samples) were between 0.001 and 0.031.

An arrow in Fig. 9 indicates that causal influence was detected ($\widehat{K}_{>0}$), and the corresponding normalized estimate is adjacent to it. Absence of an arrow indicates that $\widehat{K}_{=0}$, so no statistically causal influence was detected. The procedure identified all of the planned causal relationships, which are depicted with thick arrows (see Fig. 7). Note that no *invalid* causal influences were detected, such as from A → B and D → A. There were 18 of the possible 30 influences which would have been invalid, and all of these had pairwise directed information estimates of 0. Also, no planned causal influences were undetected (6 of the possible 30 influences). It also identified some indirect influences, which are depicted with thin arrows, such as "cascading" influences (see Fig. 10) (C→E, E→C, D→E, E→D), "proxy" influences (see Fig. 11) (A→F and B→F), and higher order influences (E→F, F→E).

After the pairwise estimates were computed, causally conditioned directed information rates were computed and the spurious influences were removed (see Fig. 12). Neurons A and B did not have any detected influencing neurons, so they were not examined. There were no neurons with only one input; if there had been, the input would have been accepted. Neurons C and D both had two influencing neurons, and for both there were connections amongst the influencing neurons. For neuron C, A and E were found to be influences. $\widehat{\mathcal{I}}(\mathrm{A} \to \mathrm{C}||\mathrm{E})$ and $\widehat{\mathcal{I}}(\mathrm{E} \to \mathrm{C}||\mathrm{A})$ were computed by first computing $\widehat{\mathcal{H}}(\mathrm{C}||\mathrm{E}, \mathrm{A})$ and then comparing with $\widehat{\mathcal{H}}(\mathrm{C}||\mathrm{E})$ and $\widehat{\mathcal{H}}(\mathrm{C}||\mathrm{A})$ respectively. For all samples, it was estimated that $\widehat{\mathcal{I}}(\mathrm{A} \to \mathrm{C}||\mathrm{E}) > 0$ and $\widehat{\mathcal{I}}(\mathrm{E} \to \mathrm{C}||\mathrm{A}) = 0$, so A→C was kept and E→C was rejected. The same procedure was performed for neuron D with influences B and E, and it was found that $\widehat{\mathcal{I}}(\mathrm{B} \to \mathrm{D}||\mathrm{E}) > 0$ and $\widehat{\mathcal{I}}(\mathrm{E} \to \mathrm{D}||\mathrm{B}) = 0$, so B→D was kept and E→D was rejected.

Neurons E and F both had five influences, but those influences were not all connected. For example, the subsets {A, C} and {B, D} each were estimated as having influences on both E and F, but not with the other subset. Thus, they could be considered separately (for example, the hypothesis that A influences F through B did not need to be tested). First, the influences for neuron E were examined. $\widehat{\mathcal{I}}(\mathrm{C} \to \mathrm{E}||\mathrm{A})$ was found to be 0 as was $\widehat{\mathcal{I}}(\mathrm{D} \to \mathrm{E}||\mathrm{B})$, for all of the samples, so C→E and D→E were rejected. Since A and B did not have any detected influences between them, A→E and B→E were kept. The same tests were done for F instead of E, but the estimates were nonzero in most cases, so they were inconclusive (since they were nonzero, but the other inputs to F were not also causally conditioned upon). To resolve this, $\widehat{\mathcal{I}}(\mathrm{A} \to \mathrm{F}||\mathrm{C}, \mathrm{D})$ and $\widehat{\mathcal{I}}(\mathrm{B} \to \mathrm{F}||\mathrm{C}, \mathrm{D})$ were both computed and found to be 0 for all the samples, and consequently A→F and B→F were rejected.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

A, B, C, D were now considered unambiguous in terms of influences on them. E and F were still ambiguous. E had A, B, and F as possible direct influences, and F had C, D, and E as possible direct influences. To resolve the ambiguity with E, $\widehat{I}(\mathrm{F} \to \mathrm{E}||\mathrm{A}, \mathrm{B})$ was computed. For 15 of the 20 samples, it was 0, and thus F→E was rejected, with A→E and B E kept. E's influences were now unambiguous. For the five samples where the estimated rate was greater than $\widehat{I}(\mathrm{A} \to \mathrm{E}||\mathrm{F}, \mathrm{B})$ and $\widehat{I}(\mathrm{B} \to \mathrm{E}||\mathrm{A}, \mathrm{F})$ were both computed and found to be nonzero, so F→E, A→E, and A→E, and B→E were all kept, and E's influences were now unambiguous. A similar procedure was done 12 of the 20 samples, E→F was rejected, leaving C→F and D→F; for the rest E→F was also kept. was also kept. Two of the samples kept both E→F and F→E. All of the influences for each of the neurons was thus resolved. The remaining influences were taken to be the direct, causal influences between the neurons (see Fig. 13).

Figure 13 depicts the averaged non-zero normalized causally conditioned estimated directed information rates for the simulated data set (with averaged 95% confidence intervals). For each of the samples, all of the planned direct, causal influences (see Fig. 7) were detected, and these all had “reliable” estimated rates (the rates much larger than the confidence interval). These are depicted with solid arrows. For some of the samples, the procedure only selected the direct, causal influences and no spurious (indirect) ones. Only two spurious influences, E to F and F to E, were detected amongst any of the samples, and their estimated rates were small and found to be unreliable (rates much smaller than confidence interval). The rates and confidence intervals for these two influences were calculated only using those samples which had detected them. Of the 20 samples, the procedure picked E→F for only eight samples, and F→E for only five samples (two of these had both). Enforcing the criterion that only reliable estimated rates would be accepted would result in only the planned, direct causal influences being accepted (that is, there would be no errors). The values in the graph (Fig. 13) are: $\widehat{I}(\mathrm{A} \to \mathrm{C})$, $\widehat{I}(\mathrm{B} \to \mathrm{D})$, $\widehat{I}(\mathrm{A} \to \mathrm{E}||\mathrm{B})$, $\widehat{I}(\mathrm{B} \to \mathrm{E}||\mathrm{A})$, $\widehat{I}(\mathrm{C} \to \mathrm{F}||\mathrm{D})$, $\widehat{I}(\mathrm{D} \to \mathrm{F}||\mathrm{C})$, $\widehat{I}(\mathrm{F} \to \mathrm{E}||\mathrm{A}, \mathrm{B})$ and $\widehat{I}(\mathrm{E} \to \mathrm{F}||\mathrm{C}, \mathrm{D})$.

It is difficult to determine how accurate the directed information estimates for the synthetic data set are. Calculating the joint statistics of the neurons using the design parameters (which were choices of *J, K, K*, $\{\alpha_i\}_{i=1}^{J}$, and $\{\beta_i\}_{i=1}^{K}$ for each neuron) is difficult and was not done for data set. However, none of the values obtained for the normalized directed information rates were substantially larger or smaller than what was anticipated given the design parameters. For two neurons X and Y, where Y is designed to causally depend on X's past spiking, if the α values of Y are fixed, then the extent of X's influence can be changed by varying X's spiking rate and the β values used in generating Y. In Eq. (34) of the conditional intensity function, which for neurons uses $h_k(x_{i-(k-1)}) = x_{i-(k-1)}$, a 1 if X had a spike at time index $i - (k - 1)$ and 0 otherwise, larger (positive or negative) values of β will generally cause the sum $\sum_{k=1}^{K} \beta_k x_{i-(k-1)}$ to have a larger magnitude (in particular when the β's have the same sign), thus having more of an effect on the conditional intensity and thus Y's spiking rate. Also, if X has a larger spiking rate, there will, in general, be more non-zero values in the sum $\sum_{k=1}^{K} \beta_k x_{i-(k-1)}$, also effecting the conditional intensity more and thus Y's spiking rate. For example, $\widehat{I}(A \to C) \approx 0.4$. C was designed to depend on A with parameters $J = 6$, $K = 5$,

$$\begin{aligned} \{\alpha_1, \cdots, \alpha_6\} &= \{-9.03, -7.02, 0.15, 1.02, 3.8, 1.3\} \\ \{\beta_1, \cdots, \beta_5\} &= \{0.1, 0.9, 4.5, 4.8, 4.1\} \end{aligned}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

A had approximately 18,000 spikes total, and C had 15,000. In contrast, $\widehat{I}(B \to D) \approx 0.9$. D was designed to depend on B with parameters $J = 8$, $K = 5$,

$$\begin{aligned}\{\alpha_1, \cdots, \alpha_8\} &= \{-9.03, -7.02, -2.5, -0.15, 0.02, 1.3, 4.8, 0.3\} \\ \{\beta_1, \cdots, \beta_5\} &= \{0.1, 7.8, 5.4, 3.1, 1.1\}\end{aligned}$$

B and D both had approximately 25,000 spikes. The α's were comparable, but the larger β values for D's dependence on B and B's larger spiking rate resulted in a larger influence for B→D as compared to A→C.

7.2 Experimental data

7.2.1 Data source—In addition to simulated data trials, experimental data from Wu and Hatsopoulos (2006) was analyzed. The data consisted of electrode array recordings from the arm area of the primary motor cortex (MI) in a juvenile male macaque monkey. The monkey was performing a series of trials involving contralateral arm movement tasks. One of the monkey's arms was attached to a robotic arm system, which constrained the arm (the shoulder joint was abducted 90°) such that shoulder and elbow movements were restricted to the horizontal plane. In each trial, a series of seven targets appeared in a workspace on the horizontal plane. The monkey moved its arm, which correspondingly moved a cursor, to hit the current target. Each target was presented for a maximum of 2 s, and if the monkey did not hit the target within that time period, the target would disappear and the next target was presented. The targets were randomly positioned, with a bias towards the exterior of the workspace, to ensure full movement of the arm. The monkey had been operantly trained to perform this task. When the monkey successfully hit the seven targets presented in a trial, the monkey was rewarded with a drop of water or juice at the end of the trial.

The recordings were obtained with a silicon micro-electrode array, which consisted of 100 platinized tip electrodes, 1.0 mm in length and with 400 μm separation (Cyberkinetics Inc, Salt Lake City, UT, USA) (Wu and Hatsopoulos 2006). The arrays were implanted in the arm area of the monkey's primary motor cortex (MI). The signals were filtered, amplified (gain 5,000), and digitally recorded (14-bit) at 30 kHz per channel (Cerebus acquisition system; Cyberkinetics, Inc.). After the experiment, the waveforms (1.6 ms in duration) with a peak voltage that passed a set threshold were stored. These selected waveforms were then spike-sorted (Offline Sorter; Plexon Inc., Dallas, TX, USA). For the sorting process, the Contours and Templates methods were used to manually extract single units. After sorting, only the single units with signal-to-noise ratio greater than 3 were kept (Wu and Hatsopoulos 2006).

7.2.2 Data analysis—For the purposes of testing the proposed directed information estimation procedure, a single data set (recordings from a single monkey in one session, with several hundred trials) was used. The data set contained spike train data (spike times) for 115 neurons for a duration of an hour. The data for each neuron was converted to a binary times series with 1 ms time resolution. 7 s samples of the data selected for neurons 3 and 1 are shown in Fig. 14. Due to the computational cost of analyzing the complete data set, only a subset of the data was used. Spike train data for only the 37 neurons with the highest total spike count (over the whole session) were kept, and only data from the first 500 s (from the beginning of the first trial) were used. Due to the sparsity of the data (the largest total spike count in the first 500 s for selected neurons was about spikes, or approximately one spike every 62 ms on average), $\Delta = 5$ ms was used. Although the resulting data was not strictly binary, there were very few instances with more than one spike in the same 5 ms time window. Directed information estimates for all ordered pairs of neurons were computed. Figure 15 is a graph of the pairwise results. Each box with a label of $i \in \{1, 2, \cdot\cdot$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

· , 37} corresponds to a different neuron, but the labeling is arbitrary (the numbers do not correspond to any sorting of the data). The position of a neuron in the graph corresponds to the position of the electrode on the array that detected that neuron. Note that adjacent boxes, such as {2, 3, 4} and {5, 6} correspond to multiple neurons detected on the same electrode, although for visual purposes the boxes are only partially overlapping.

A directed arrow is graphed for each ordered pair (X,Y) of neurons for which the estimation procedure detected a statistically causal influence ($\widehat{K}_{>0}$). Absence of an arrow between an ordered pair (X,Y) depicts that the estimation procedure detected that there was no statistically causal influence ($\widehat{K}_{=0}$). The normalized directed information estimates are not included in the graph for clarity purposes. Most of the normalized directed information estimates were on the order of 10^{-2} to 10^{-3}. Note that the causal influences detected in this data set were not as large as those detected in the simulated data set. The simulated data set was constructed to have large statistically causal influences, whereas neurons recorded from in brain tissue could have many neighboring neurons exciting or inhibiting it (thus the influence from any one neuron could be small). It is also possible that the neurons which were detected to have a statistically causal relationship do not directly communicate with each other, but only do so through other neurons that might not be present in the data set.

After the pairwise directed information estimates were computed, a small number of nodes were selected which had few pairwise influences and whose influences were ambiguous. These nodes and their respective influences were then examined using causally conditioned directed information, to determine which of the influences were direct. The subsets examined include {1, 4, 9}, {3, 10, 13}, {5, 13, 35}, {8, 10, 27}, {13, 18, 25}, and {32, 33, 36}. For each of the subsets {1, 4, 9}, {3, 10, 13}, and {13, 18, 25}, one of the causally conditioned directed information estimates were 0, and thus one of the estimated of the estimated pairwise influences was removed from each. See Figs. 16, 17, 18, 19, 20 and 21. For the other subsets, all of the causally conditioned directed information estimates were greater than 0, and so they were kept.

A strong structure can be seen in the graph (Fig. 15). Some neurons have many incoming and outgoing connections, such as 1, 8, and 12. Some have more incoming than outgoing, such as 8, and 18. Some have very few, if any, incoming or outgoing connections. Note that this is only suggestive of the *functional* connectivity of the neurons, and only amongst those used in the analysis. It is unclear what the underlying physical connectivity structure of the region of recorded brain tissue is. That a statistically causal influence from a neuron X to a neuron Y is detected in this data set is only suggestive that there might be *some* physical pathway between the two neurons, such that the spiking activity of activity of X could influence the spiking activity of Y. Many of the neurons present in the section of brain tissue recorded from are not present in this analysis (Wu and Hatsopoulos 2006). Similar to the analysis of the simulated data set, even amongst the recorded neurons, it is unclear what influences are “direct,” which might be accounted for by “proxy” or “cascading” effects (see Fig. 6).

In addition to the number of detected influence relationships between the neurons, there is also a visibly dominant orientation of the connections (see Fig. 15). While the procedure detected relationships in many directions, there are a large number of connections along the bottom left to upper right diagonal (oriented with respect to the recording electrode array). Neurons 1, 5, 12, 13, and 31 all have several arrows (incoming and outgoing) along this diagonal. This result is promising, because it might correspond to propagating waves of high frequency oscillations in the beta range (10–45 Hz) in the motor cortex (Rubino et al. 2006). These oscillation waves observed in local field potentials (LFPs) in the motor cortex have been found to encode information about visual targets in reaching tasks, and are thought to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

facilitate information transfer between intra- and inter-cortical regions during movement preparation and execution (Rubino et al. 2006). Other studies have found that in the turtle visual cortex, these waves were present during the introduction of visual stimuli (Prechtl et al. 1997) and have been shown to encode information related to target position (Du et al. 2005). Similar wave-like spatiotemporal activity has been observed in other areas of the nervous systems of a variety of animals and are thought to have an important role in the communication between different areas of the brain (Ermentrout and Kleinfeld 2001). Physically, beta oscillations are believed to correspond to the summed effects of multiple, synchronous postsynaptic potentials from neurons close to the recording electrode (Rubino et al. 2006). There is little is known about the precise mechanisms through which the propagation of these waves occur (Rubino et al. 2006). The proposed estimation procedure could provide insight into these mechanisms. The procedure could potentially both identify the local propagation pathways (by detecting structure as in Fig. 15) as well as the specific relationship dynamics between the recorded neurons (by identifying the coefficients of the conditional intensity function, the α_is and β_js).

8 Future work

The current estimation procedure has been proven theoretically and shown through simulated data trials to correctly identify statistically causal influences. It has also identified strong structure in an electrode recording data set. There are several improvements that could furthermore enhance both the theoretical and practical aspects of this procedure. To improve computation time, stochastic optimization procedures, such as cross-entropy (De Boer et al. 2005) and model reference adaptive search (MRAS) (Hu et al. 2007), will be tested. In the current, deterministic, estimation procedure, the MLEs for a large number of (J, K) values are separately computed and then compared with corresponding complexity penalty terms. This can be computationally expensive and possibly redundant (as the MLE calculation for a large (J, K) searches over the spaces examined for smaller values of (J, K)). Stochastic optimization procedures, however, directly optimize over the objective function (Eq. (37)), potentially accelerating the optimization process. Additionally, the development of efficient algorithms to identify the direct, causal structure of a network could benefit the employment of this procedure to large data sets. In the simulated and electrode array recordings data sets analyzed in the results section, all of the pairwise directed information estimates were analyzed first, and then indirect influences examined. This might not be efficient for all networks of interest, in particular when there is a priori knowledge of underlying structure.

9 Conclusion

The information theoretic quantity directed information was introduced as a measure of statistical causality. The directed information has been shown to characterize statistically causal influences between arbitrary processes, including binary time series of neuronal spiking activity, in a more robust and meaningful way than previously used methods such as Granger causality. It was also noted that while it is quantitatively different than Granger causality, their philosophical underpinnings are identical. Using an established, statistical model class for neuronal spiking activity, a parametric estimation procedure for the directed information was developed and consistency properties were proven.

The procedure was tested on simulated data and applied to experimental data, both with promising results. Further tests on more complicated simulated data, and further analyses of real data will be performed to further test the effectivenss of this procedure. Also, several theoretical and practical improvements will be made to enhance this methodology. This technique could become a practical, provably-good, and philosophically well-grounded

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

means of identifying the statistically causal, complex relationships between neurons in large data sets of simultaneous, multiple electrode recordings.

Appendix A: Proof of Lemma 1

Proof First, prove that $\mathcal{H}(\mathrm{Y}\|\mathrm{X})$. This proof closely follows the proof for the unconditional entropy rate in Cover and Thomas (2006). An important theorem used for the proof is the Cesaro mean theorem (Cover and Thomas 2006): For sequences of real numbers $(a_1, \cdots a_n)$ and $(b_1, \cdots b_n)$, if $\lim_{n\to\infty} a_n = a$, and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_n$, then $\lim_{n\to\infty} b_n = a$.

By definition, $\mathrm{H}(Y^n\|X^n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}\left(Y_i|Y^{i-1}, X^i\right)$. Since conditioning reduces entropy, entropy is nonnegative, and the processes are jointly stationary, we have

$$0 \le \mathrm{H}\left(Y_i|Y^{i-1}, X^i\right) \le \mathrm{H}(Y_1) \quad \forall_i.$$

Observe that

$$\mathrm{H}\left(Y_i|Y^{i-1}, X^i\right) \le \mathrm{H}\left(Y_i|Y_2^{i-1}, X_2^i\right) \tag{46}$$

$$= \mathrm{H}\left(Y_{i-1}|Y^{i-2}, X^{i-1}\right), \tag{47}$$

where Eq. (46) uses the property that conditioning reduces entropy (in reverse) and Eq. (47) uses stationarity. This sequence of real numbers (once the process is defined, that is, the underlying probability distribution is specified), the entropies are deterministic numbers) $a_i \triangleq \mathrm{H}\left(Y_i|Y^{i-1}, X^i\right)$ are nonincreasing and bounded below by 0. Therefore, limit of a_n as $n \to \infty$ exists, and thus, by employing Cesaro mean theorem, $\mathrm{H}(\mathcal{Y}\|\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{1}{n}\mathrm{H}(Y^n\|X^n)$ exists.

Next, taking X^n to be a deterministic sequence, and following the above, $\mathrm{H}(\mathcal{Y}) \triangleq \lim_{n\to\infty} \frac{1}{n}\mathrm{H}(Y^n)$ exists. Taking the limit in Eq. (24), $\mathcal{I}(\mathrm{X} \to \mathrm{Y}) \triangleq \lim_{n\to\infty} \frac{1}{n}\mathrm{I}(X^n \to Y^n)$ also exists.

Appendix B: Proof of Lemma 2

Proof The normalized causal entropy can be rewritten as

$$\frac{1}{n}\mathrm{H}(Y^n\|X^n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{H}\left(Y_i|X^i, Y^{i-1}\right) \tag{48}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[-\log \quad P_{Y_i|Y^{i-1}X^i}\left(Y_i|Y^{i-1}, X^i\right)\right] \tag{49}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[-\log \quad P_{Y_i|Y_{i-J}^{i-1},X_{i-(K-1)}^{i}}\left(Y_i|Y_{i-J}^{i-1}, X_{i-(K-1)}^{i}\right)\right] \tag{50}$$

$$\begin{aligned} &= \tfrac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[g_{JK}\left(Y_{i-J}^{i}, X_{i-(K-1)}^{i}\right)\right] \\ &= \mathrm{E}\left[g_{JK}\left(Y_{l-J}^{l}, X_{l-(K-1)}^{l}\right)\right] \end{aligned} \tag{51}$$

where Eq. (48) follows by the definition of causally conditioned entropy, Eq. (49) follows by chain rule for entropy, Eq. (50) follows from the Markov assumption, and Eq. (51) follows from the stationarity assumption.

References

Abler B, Roebroeck A, Goebel R, Höse A, Schönfeldt-Lecuona C, Hole G, et al. Investigating directed influences between activated brain areas in a motor-response task using fMRI. Magnetic Resonance Imaging. 2006; 24(2):181–185. [PubMed: 16455407]

Akaike H. An information criterion (AIC). Mathematical Scientist. 1976; 14(153):5–9.

Al-khassaweneh M, Aviyente S. The relationship between two directed information measures. IEEE Signal Processing Letters. 2008; 15:801–804.

Amblard P, Michel O. On directed information theory and Granger causality graphs. Arxiv preprint. 2010 arXiv:1002.1446.

Barron A, Cover T. Minimum complexity density estimation. IEEE Transactions on Information Theory. 1991; 37(4):1034–1054.

Bitan T, Booth J, Choy J, Burman D, Gitelman D, Mesulam M. Shifts of effective connectivity within a language network during rhyming and spelling. Journal of Neuroscience. 2005; 25(22):5397. [PubMed: 15930389]

Bremaud, P. Point processes and queues: martingale dynamics. Springer; New York: 1981.

Brovelli A, Ding M, Ledberg A, Chen Y, Nakamura R, Bressler S. 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. 2004; 101(26):9849. [PubMed: 15210971]

Brown, E.; Barbieri, R.; Eden, U.; Frank, L. Likelihood methods for neural spike train data analysis.. Computational neuroscience: A comprehensive approach. 2003.

Brown E, Barbieri R, Ventura V, Kass R, Frank L. The time-rescaling theorem and its application to neural spike train data analysis. Neural Computation. 2002; 14(2):325–346. [PubMed: 11802915]

Cai H, Kulkarni S, Verdú S. Universal entropy estimation via block sorting. IEEE Transactions on Information Theory. 2004; 50(7):1551–1561.

Cai H, Kulkarni S, Verdu S. An algorithm for universal lossless compression with side information. IEEE Transactions on Information Theory. 2006; 52(9):4008–4016.

Casella, G.; Berger, R.; Berger, R. Statistical inference. Duxbury; Pacific Grove: 2002.

Cesa-Bianchi, N.; Lugosi, G. Prediction, learning, and games. Cambridge University Press; Cambridge: 2006.

Chávez M, Martinerie J, Le Van Quyen M. Statistical assessment of nonlinear causality: Application to epileptic EEG signals. Journal of Neuroscience Methods. 2003; 124(2):113–128. [PubMed: 12706841]

Cover, T.; Thomas, J. Elements of information theory. Wiley-Interscience; New York: 2006.

Daley, D.; Vere-Jones, D. An introduction to the theory of point processes. Springer; New York: 1988.

David O, Kiebel S, Harrison L, Mattout J, Kilner J, Friston K. Dynamic causal modeling of evoked responses in EEG and MEG. NeuroImage. 2006; 30(4):1255–1272. [PubMed: 16473023]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

De Boer P, Kroese D, Mannor S, Rubinstein R. A tutorial on the cross-entropy method. Annals of Operations Research. 2005; 134(1):19–67.

Dhamala M, Rangarajan G, Ding M. Analyzing information flow in brain networks with nonparametric Granger causality. NeuroImage. 2008; 41(2):354–362. [PubMed: 18394927]

Diekman CO, Sastry P, Unnikrishnan K. Statistical significance of sequential firing patterns in multi-neuronal spike trains. Journal of Neuroscience Methods. 2009; 182(2):279–284. [PubMed: 19559053]

Du X, Ghosh B, Ulinski P. Encoding and decoding target locations with waves in the turtle visual cortex. IEEE Transactions on Biomedical Engineering. 2005; 52(4):566–577. [PubMed: 15825858]

Eguiluz V, Chialvo D, Cecchi G, Baliki M, Apkarian A. Scale-free brain functional networks. Physical Review Letters. 2005; 94(1):018102. [PubMed: 15698136]

Elia N. When bode meets Shannon: Control-oriented feedback communication schemes. IEEE Transactions on Automatic Control. 2004; 49(9):1477–1488.

Ermentrout G, Kleinfeld D. Traveling electrical waves in cortex insights from phase dynamics and speculation on a computational role. Neuron. 2001; 29(1):33–44. [PubMed: 11182079]

Friston K, Harrison L, Penny W. Dynamic causal modelling. NeuroImage. 2003; 19(4):1273–1302. [PubMed: 12948688]

Goebel R, Roebroeck A, Kim D, Formisano E. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. Magnetic Resonance Imaging. 2003; 21(10):1251–1261. [PubMed: 14725933]

Gorantla, S.; Coleman, T. IEEE international symposium on information theory (ISIT). Austin, TX: 2010. On reversible Markov chains and maximization of directed information.. (in press)

Gourevitch B, Eggermont J. Evaluating information transfer between auditory cortical neurons. Journal of Neurophysiology. 2007; 97(3):2533. [PubMed: 17202243]

Granger C. Investigating causal relations by econometric models and cross-spectral methods. Econometrica. 1969; 37(3):424–438.

Grefkes C, Eickhoff S, Nowak D, Dafotakis M, Fink G. Dynamic intra-and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM. NeuroImage. 2008; 41(4):1382–1394. [PubMed: 18486490]

Grünwald, P.; Rissanen, J. The minimum description length principle. MIT; Cambridge: 2007.

Hamandi K, Powell H, Laufs H, Symms M, Barker G, Parker G, et al. Combined EEG-fMRI and tractography to visualise propagation of epileptic activity. British Medical Journal. 2008; 79(5): 594–597.

Hesse W, Möller E, Arnold M, Schack B. The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies. Journal of Neuroscience Methods. 2003; 124(1):27–44. [PubMed: 12648763]

Hu J, Fu M, Marcus S. A model reference adaptive search method for global optimization. Operations Research. 2007; 55(3):549–568.

Iyengar S, Liao Q. Modeling neural activity using the generalized inverse Gaussian distribution. Biological Cybernetics. 1997; 77(4):289–295. [PubMed: 9394447]

Kaminski M, Blinowska K. A new method of the description of the information flow in the brain structures. Biological Cybernetics. 1991; 65(3):203–210. [PubMed: 1912013]

Kamiński M, Ding M, Truccolo W, Bressler S. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. Biological Cybernetics. 2001; 85(2):145–157. [PubMed: 11508777]

Kim, Y.; Pennuter, H.; Weissman, T. Directed information and causal estimation in continuous time.. IEEE international symposium on information theory (ISIT). 2009.

Korzeniewska A, Mańczak M, Kamiński M, Blinowska K, Kasicki S. Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method. Journal of Neuroscience Methods. 2003; 125(1–2):195–207. [PubMed: 12763246]

Kramer, G. Ph.D. thesis. University of Manitoba; Canada: 1998. Directed information for channels with feedback..

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Kramer M, Eden U, Cash S, Kolaczyk E. Network inference with confidence from multivariate time series. Physical Review E. 2009; 79(6):61916.

Kraskov, A. Synchronization and interdependence measures and their application to the electroencephalogram of epilepsy patients and clustering of data. 2008. Report Nr.: NIC series; 24

Lastras, L. An almost sure convergence proof of the sliding-window Lempel-Ziv algorithm.. Proceedings 2002 IEEE international symposium on information theory. 2002.

Marko H. The bidirectional communication theory–A generalization of information theory. IEEE Transactions on Communications. 1973; 21(12):1345–1351.

Martins N, Dahleh M. Feedback control in the presence of noisy channels: "Bode-like" fundamental limitations of performance. IEEE Transactions on Automatic Control. 2008; 53(7):1604–1615.

Massey, J. Causality, feedback and directed information.; Proc. int. symp. information theory application (ISITA-90). 1990. p. 303-305.

Massey, J.; Massey, P. Conservation of mutual and directed information.; Proceedings international symposium on information theory, 2005. ISIT 2005. 2005. p. 157-158.

Mathai, P.; Martins, N.; Shapiro, B. On the detection of gene network interconnections using directed mutual information. ITA; San Deigo: 2007.

Meyn, S.; Tweedie, R. Markov chains and stochastic stability. Cambridge Mathematical Library; Cambridge: 2009. p. 622

Okatan M, Wilson M, Brown E. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. Neural Computation. 2005; 17(9):1927–1961. [PubMed: 15992486]

Paninski L. Estimation of entropy and mutual information. Neural Computation. 2003; 15(6):1191–1253.

Paninski L, Fellows M, Hatsopoulos N, Donoghue J. Spatiotemporal tuning of motor cortical neurons for hand position and velocity. Journal of Neurophysiology. 2004; 91(1):515. [PubMed: 13679402]

Pearl, J. Causality: Models, reasoning and inference. Cambridge University Press; New York: 2009.

Pereda E, Quiroga R, Bhattacharya J. Nonlinear multivariate analysis of neurophysiological signals. Progress in Neurobiology. 2005; 77(1–2):1–37. [PubMed: 16289760]

Perez-Cruz, F. Estimation of information theoretic measures for continuous random variables. NIPS; 2008.

Permuter H, Kim Y, Weissman T. On directed information and gambling. In IEEE international symposium on information theory, 2008. ISIT 2008. 2008:1403–1407.

Permuter H, Kim Y, Weissman T. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. Arxiv preprint. 2009a arXiv:0912.4872.

Permuter H, Weissman T, Goldsmith A. Finite state channels with time-invariant deterministic feedback. IEEE Transactions on Information Theory. 2009b; 55(2):644–662.

Prechtl J, Cohen L, Pesaran B, Mitra P, Kleinfeld D. Visual stimuli induce waves of electrical activity in turtle cortex. Proceedings of the National Academy of Sciences of the United States of America. 1997; 94(14):7621. [PubMed: 9207142]

Ramnani N, Behrens T, Penny W, Matthews P. New approaches for exploring anatomical and functional connectivity in the human brain. Biological Psychiatry. 2004; 56(9):613–619. [PubMed: 15522243]

Rao A, Hero III A, States D, Engel J. Inference of biologically relevant gene influence networks using the directed information criterion. Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP). 2006; 2:1028–1031.

Rao A, Hero III A, States DJ, Engel JD. Inferring time-varying network topologies from gene expression data. EURASIP Journal on Bioinformatics and System Biology-Special Issue on Gene Networks. 2007; 2007:51947.

Rao A, Hero III A, David J, Engel J. Using directed information to build biologically relevant influence networks. Journal of Bioinformatics and Computational Biology. 2008; 6(3):493–519. [PubMed: 18574860]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Rissanen J, Wax M. Measures of mutual and causal dependence between two time series (Corresp.). IEEE Transactions on Information Theory. 1987; 33(4):598–601.

Roebroeck A, Formisano E, Goebel R. Mapping directed influence over the brain using Granger causality and fMRI. NeuroImage. 2005; 25(1):230–242. [PubMed: 15734358]

Rogers B, Morgan V, Newton A, Gore J. Assessing functional connectivity in the human brain by fMRI. Magnetic Resonance Imaging. 2007; 25(10):1347–1357. [PubMed: 17499467]

Rubino D, Robbins K, Hatsopoulos N. Propagating waves mediate information transfer in the motor cortex. Nature Neuroscience. 2006; 9(12):1549–1557.

Salvador R, Suckling J, Schwarzbauer C, Bullmore E. Undirected graphs of frequency-dependent functional connectivity in whole brain networks. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005; 360(1457):937–946.

Schreiber T. Measuring information transfer. Physical Review Letters. 2000; 85(2):461–464. [PubMed: 10991308]

Schuyler B, Ollinger J, Oakes T, Johnstone T, Davidson R. Dynamic Causal Modeling applied to fMRI data shows high reliability. NeuroImage. 2009; 49:603–611. [PubMed: 19619665]

Seth A, Edelman G. Distinguishing causal interactions in neural populations. Neural Computation. 2007; 19(4):910–933. [PubMed: 17348767]

Smith V, Yu J, Smulders T, Hartemink A, Jarvis E. Computational inference of neural information flow networks. PLoS Computational Biology. 2006; 2(11):e161. [PubMed: 17121460]

Stephan K, Kasper L, Harrison L, Daunizeau J, den Ouden H, Breakspear M, et al. Nonlinear dynamic causal models for fMRI. NeuroImage. 2008; 42(2):649–662. [PubMed: 18565765]

Stevenson I, Rebesco J, Hatsopoulos N, Haga Z, Miller L, Körding K. Bayesian inference of functional connectivity and network structure from spikes. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2009; 17(3):203. [PubMed: 19273038]

Sundaresan R, Verdú S. Capacity of queues via point-process channels. IEEE Transactions on Information Theory. 2006; 52(6):2697–2709.

Tatikonda, S. Ph.D. thesis. Massachusetts Institute of Technology; 2000. Control under communication constraints..

Tatikonda S, Mitter S. The capacity of channels with feedback. IEEE Transactions on Information Theory. 2009; 55(1):323–349.

Truccolo W, Eden U, Fellows M, Donoghue J, Brown E. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. Journal of Neurophysiology. 2005; 93(2):1074–1089. [PubMed: 15356183]

Uddin L, Clare Kelly A, Biswal B, Xavier Castellanos F, Milham M. Functional connectivity of default mode network components: Correlation, anticorrelation, and causality. Human Brain Mapping. 2009; 30(2):625–637. [PubMed: 18219617]

Venkataramanan R, Pradhan S. Source coding with feed-forward: Rate-distortion theorems and error exponents for a general source. IEEE Transactions on Information Theory. 2007; 53(6):2154–2179.

Vogels T, Abbott L. Signal propagation and logic gating in networks of integrate-and-fire neurons. Journal of Neuroscience. 2005; 25(46):10786. [PubMed: 16291952]

Wang X, Chen Y, Bressler S, Ding M. Granger causality between multiple interdependent neurobiological time series: Blockwise versus pairwise methods. International Journal of Neural Systems. 2007; 17(2):71. [PubMed: 17565503]

Wu W, Hatsopoulos N. Evidence against a single coordinate system representation in the motor cortex. Experimental Brain Research. 2006; 175(2):197–210.

Zhao, L.; Permuter, H.; Kim, Y.; Weissman, T. IEEE international symposium on information theory (ISIT). Austin, TX: 2010. Universal estimation of directed information.. (in press)

Ziv J, Lempel A. A universal algorithm for sequential data compression. IEEE Transactions on Information Theory. 1977; 23(3):337–343.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

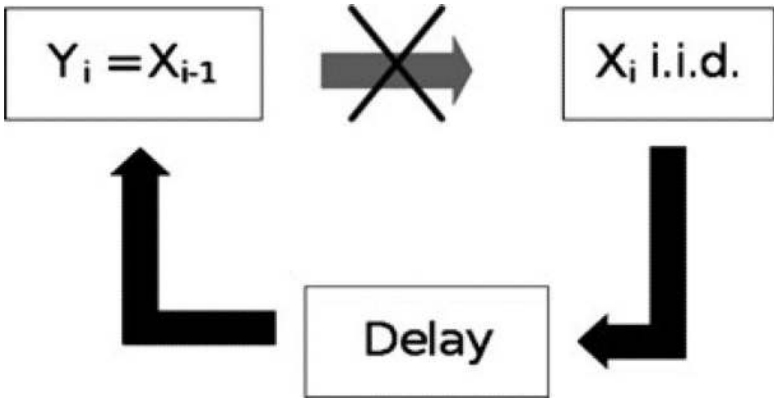

Fig. 1.

Diagram of the processes and their causal relationship. **X** is drawn i.i.d. equi-probably to be 0 or 1, and $Y_i = X_{i-1}$. Clearly **X** is causally influencing **Y**. Moreover, **Y** is *not* causally influencing *X*

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

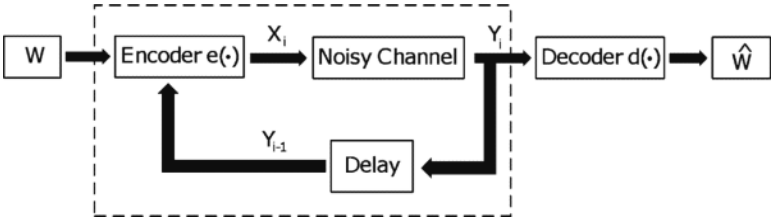

Fig. 2.

Diagram of a noisy channel. The capacity of the noisy channel without feedback is a function of $I(X^n; Y^n)$. With feedback, the capacity of the noisy channel changes. The capacity of the whole channel (*inside the dotted line*), which includes both the noisy channel and the feedback, is always a function of $I(W; Y^n) = I(X^n \rightarrow Y^n)$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

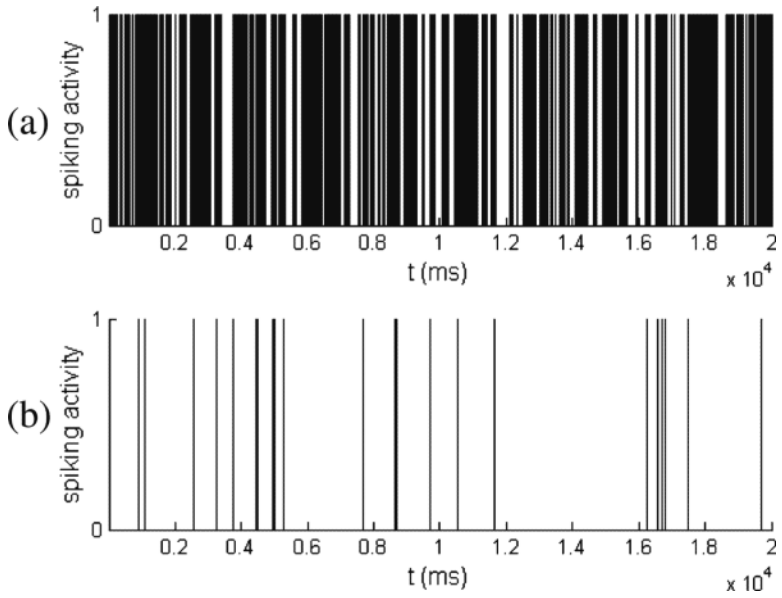

Fig. 3.
Spiking activity of neurons A (*top*) and B (*bottom*)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

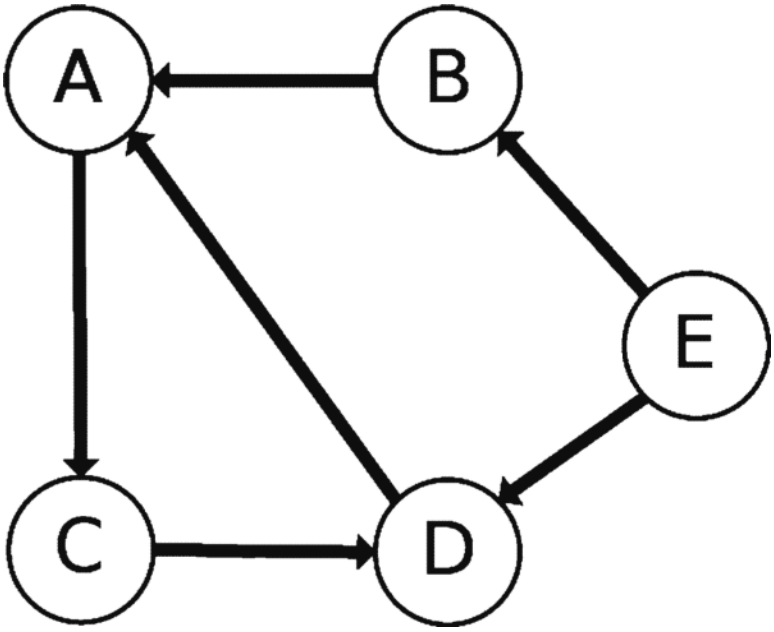

Fig. 4.
A graphical depiction of the direct, causal influences in a network of processes

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

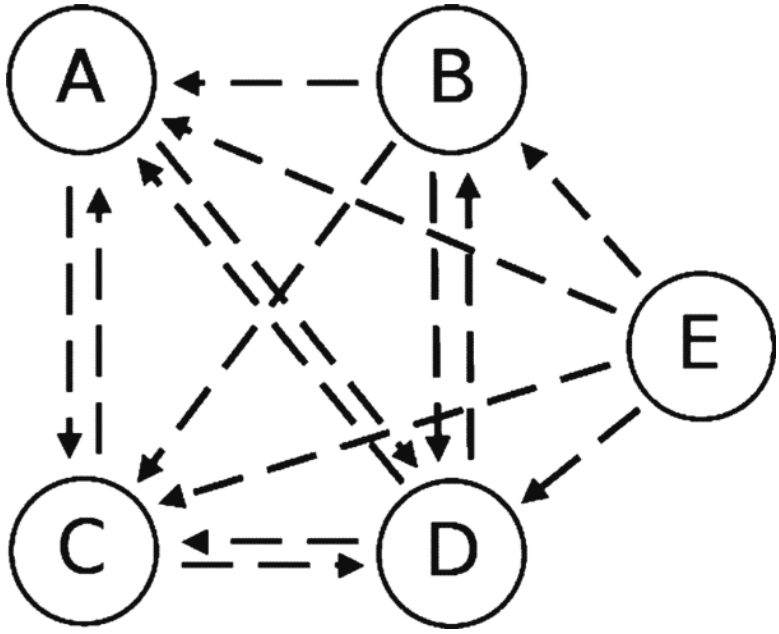

Fig. 5.
A graphical depiction of the causal influences in a network of processes

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

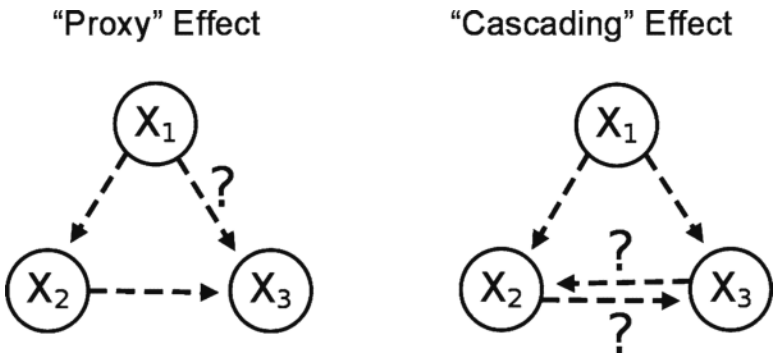

Fig. 6.

A graphical depiction of two types of indirect influences. Each *arrow* depicts a causal influence. The *arrows with a question mark* are the indirect influencs influences

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

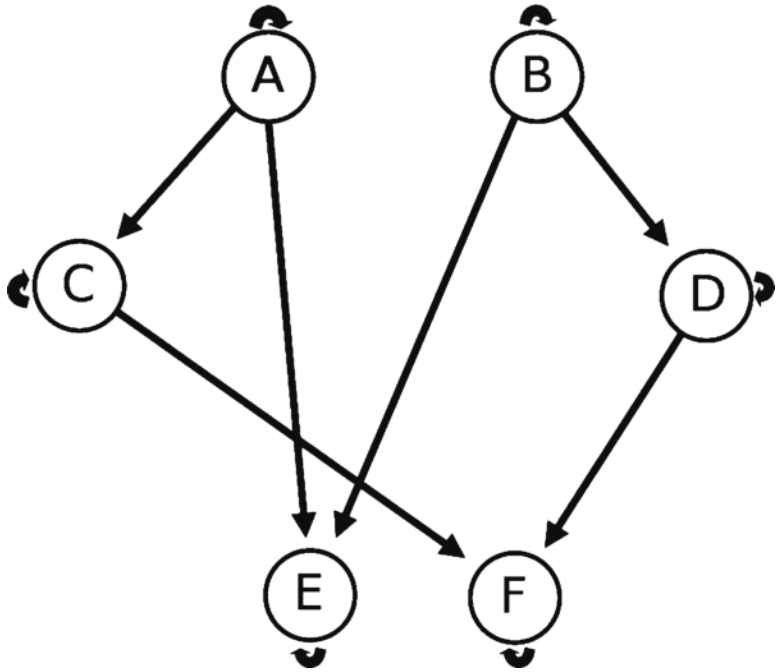

Fig. 7.

Diagram of the direct, causal influence structure that the simulated data set models. Note that an *arrow* from neuron M to neuron N (possibly with M = N) means that N was designed to be causally dependent on M's firing via N's conditional intensity function. Thus, the *arrows* represent direct, causal influences

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

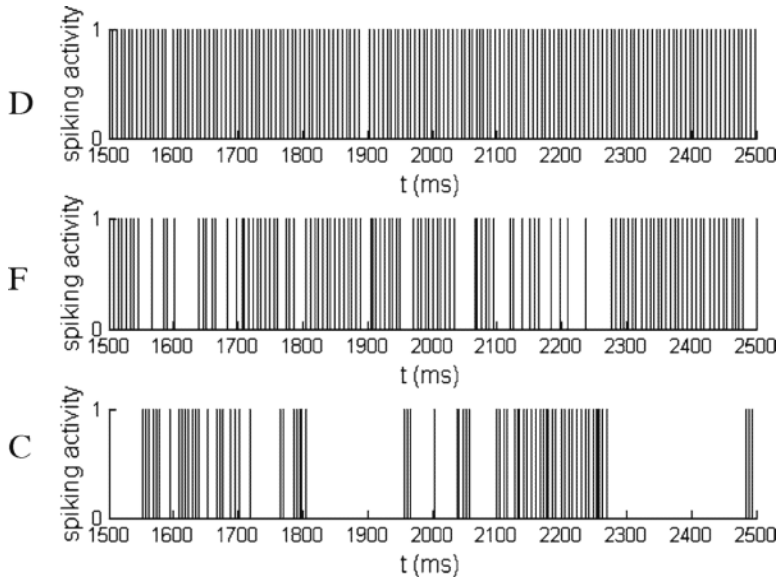

Fig. 8.

A one second sample of the spike trains generated for neurons D, F, and C. Neuron D was excitatory, whereas neuron C was inhibitory, in causally influencing neuron F

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

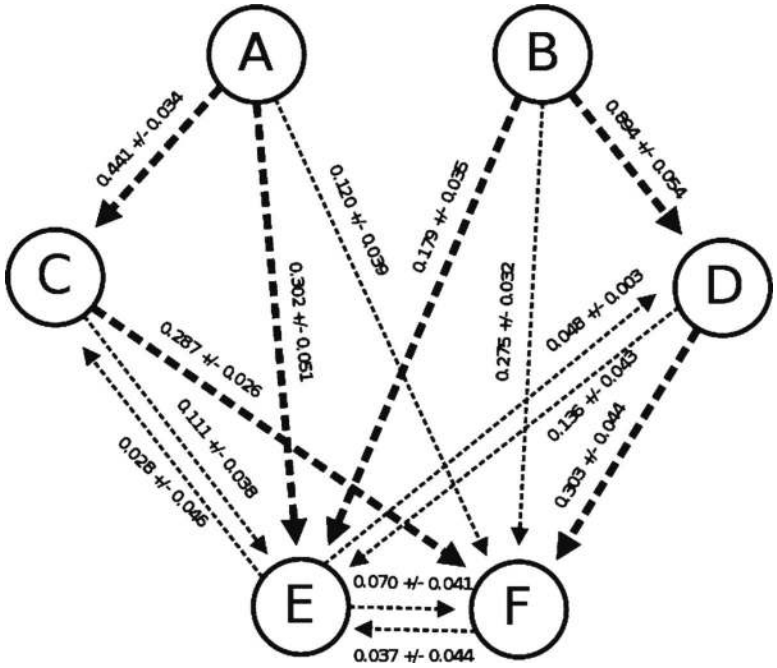

Fig. 9.

Diagram of the averaged non-zero, estimated normalized pairwise directed information rates (with averaged 95% confidence intervals) for the simulated data set, using 20 independently generated samples. The procedure selected the same structure for each sample. The procedure identified all the direct, causal relationships, which are depicted with *thick, dashed arrows* (see Fig. 7). No *invalid* (pairwise) causal influences were detected (18 of the possible 30 *arrows*), nor were any planned causal influences *undetected*. The procedure also identified some indirect influences, which are depicted with *thin, dashed arrows*, such as "proxy" influences (i.e. the groups B, D, and F) as well as some "cascading" influences (i.e. the groups B, D, and E)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

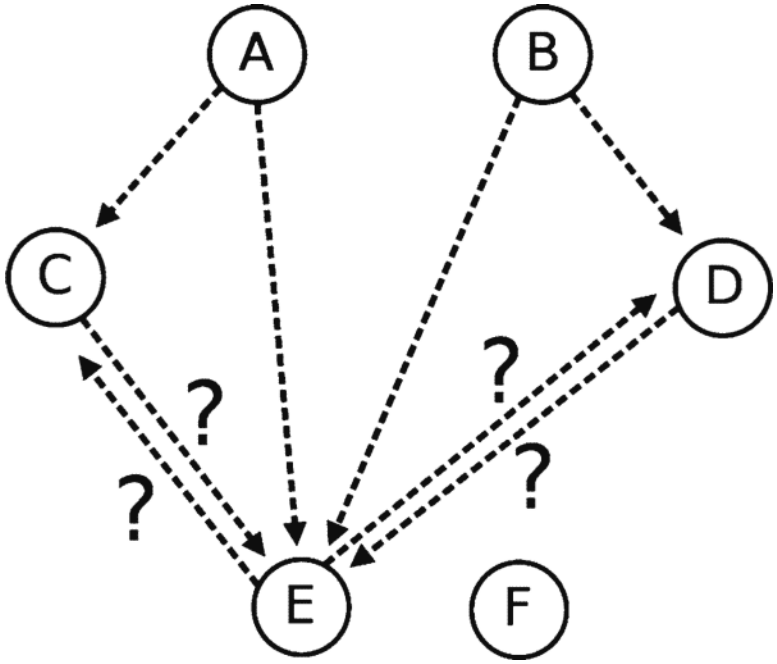

Fig. 10.
Diagram depicting a subgraph, in which cascading influences (denoted by *arrows* with adjacent"?") were detected by the pairwise directed information estimates

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

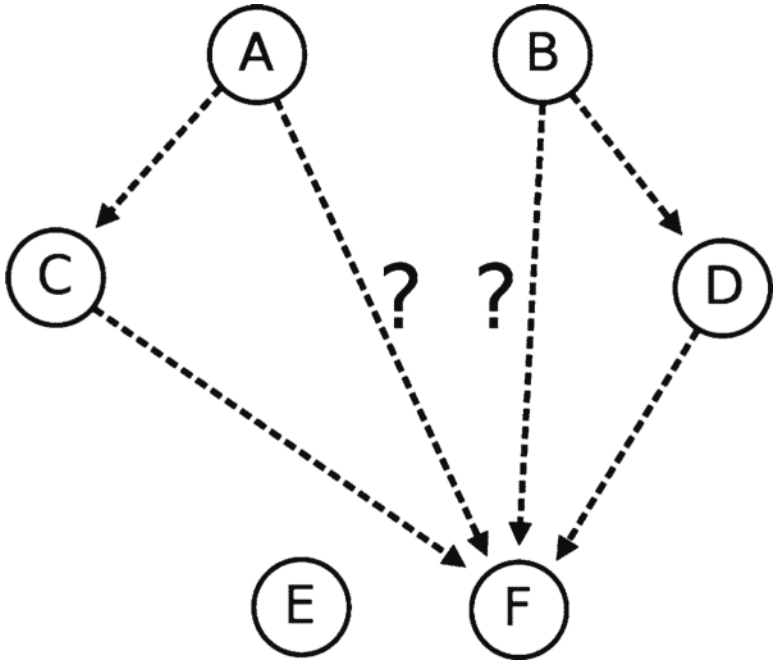

Fig. 11.
Diagram depicting a subgraph, in which proxy influences (denoted by *arrows* with adjacent"?") were detected by the pairwise directed information estimates

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

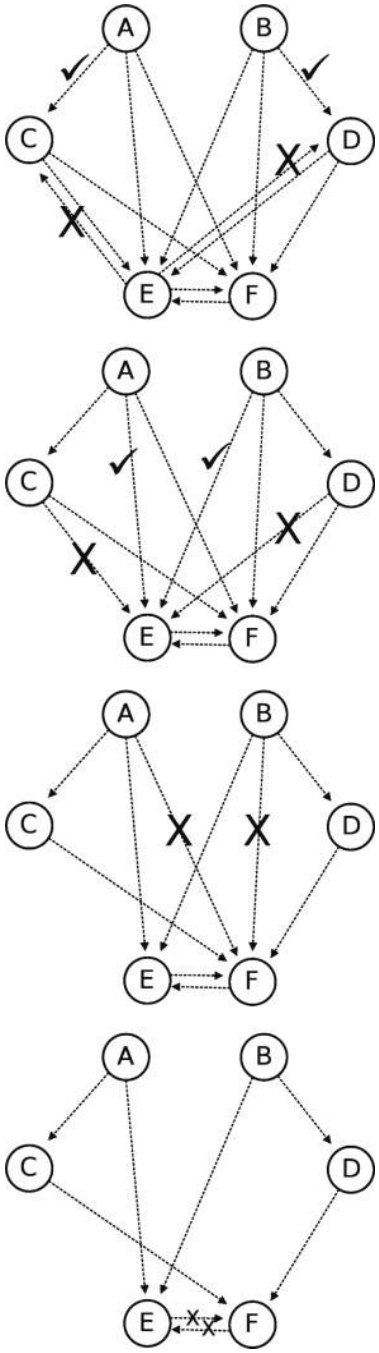

Fig. 12.

Steps of the algorithm to identify which of the detected (pairwise) causal influences are *direct* causal influences. A *check-mark* is placed next to influences that were tested and kept at that stage in the algorithm. An "X" is placed over the influences which were determined to not be direct, causal influences. The algorithm found the same results for the top three figures in each of the 20 sample sets. The *bottom figure* has smaller "X"s because the algorithm estimated that those influences were not direct in most, but not all, of the sample data sets

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

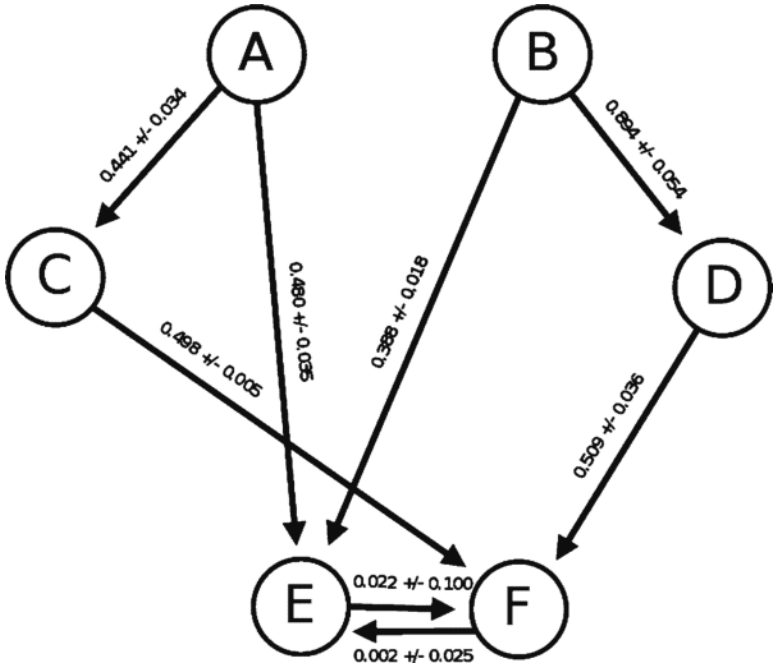

Fig. 13.

Diagram of the averaged non-zero normalized causally conditioned estimated directed information rates for the simulated data set (with averaged 95% confidence intervals). For each of the samples, all of the direct, causal influences (see Fig. 7) were detected. Only two spurious (indirect) influences, E to F and F to E, were detected amongst any of the samples. Nine of the 20 samples detected neither and nine of the 20 only detected one of them. Their estimated rates were small and found to be unreliable (rates much smaller than confidence interval)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

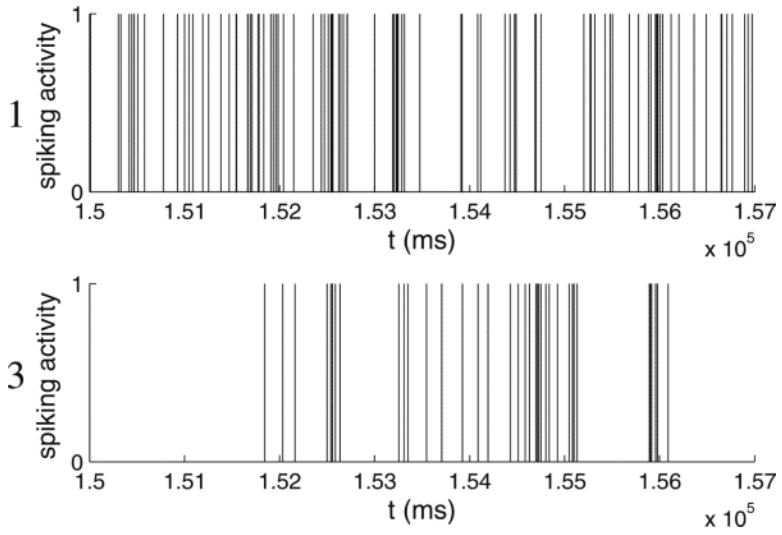

Fig. 14.

Seven second snapshot of spiking activity of neurons 1 and 3 in the data set from Wu and Hatsopoulos (2006) used for analysis. The procedure found that neuron 3 causally influences neuron 1, in an excitatory manner

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

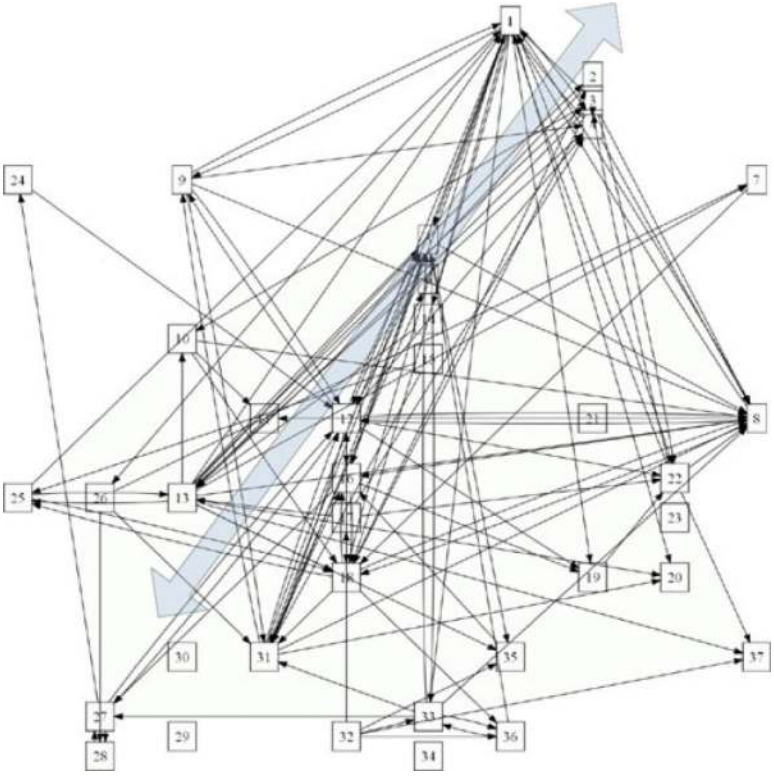

Fig. 15.

Diagram of statistically estimated causal relationships for the 37 neurons used from the subset of electrode recordings in the arm area of a monkey's primary motor cortex (MI) from Wu and Hatsopoulos (2006). Each *box* with a number indicates a different neuron. The relative positions of the neurons in the diagram correspond to the relative positions of the electrodes on the electrode array where the neurons were detected. An *arrow from a box labelled X to a box labelled Y* depicts that a statistically causal relationship was detected from X to Y (in particular, $\widehat{K}_{>0}$). Absence of an arrow from X to Y depicts that the procedure detected no statistically causal relationship from X to Y ($\widehat{K}_{=0}$). The *transparent diagonal arrow* represents a 'dominant' orientation of the detected causal influences. This might correspond to the direction of propagating local field potential waves discussed in Rubino et al. (2006)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

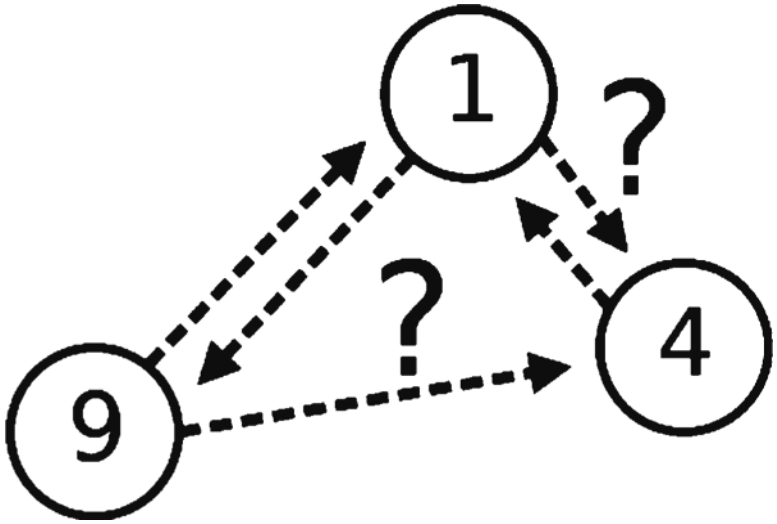

Fig. 16.

Diagram depicting the induced subgraph of neurons 1, 9, and 4. Both 1 and 9 have pairwise influences into 4, one of which might be due to an indirect influence. A *question mark* is drawn adjacent the *arrows* in question

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

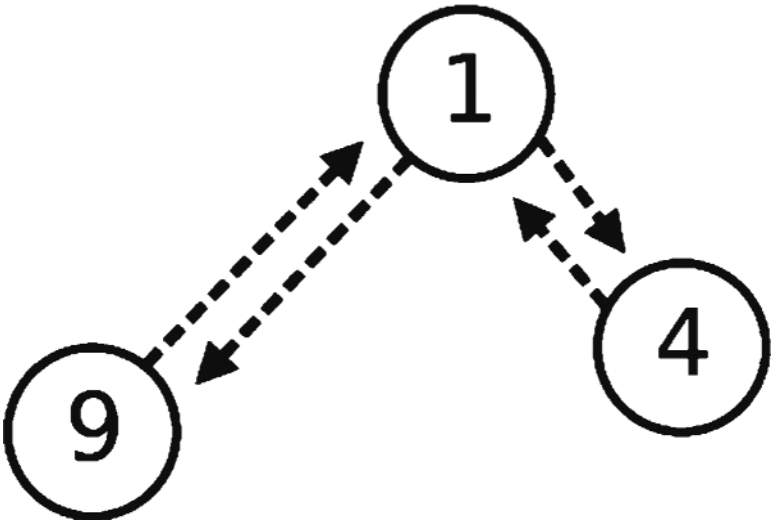

Fig. 17.

The resulting subgraph after computing causally conditioned directed information estimates. $\widehat{\mathcal{I}}(1 \to 4 \| 9) > 0$ and $\widehat{\mathcal{I}}(9 \to 4 \| 1) = 0$, so $9 \to 4$ was removed, and 1→4 kept

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

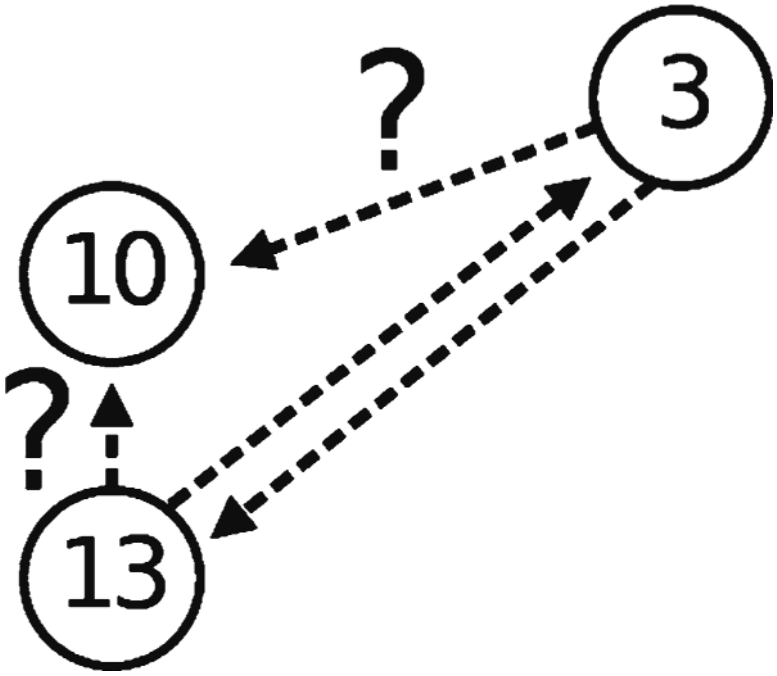

Fig. 18.

Diagram depicting the induced subgraph of neurons 3, 10, and 13. Both 3 and 13 have pairwise influences into 10, one of which might be due to an indirect influence. A *question mark* is drawn adjacent to the *arrows* in question

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

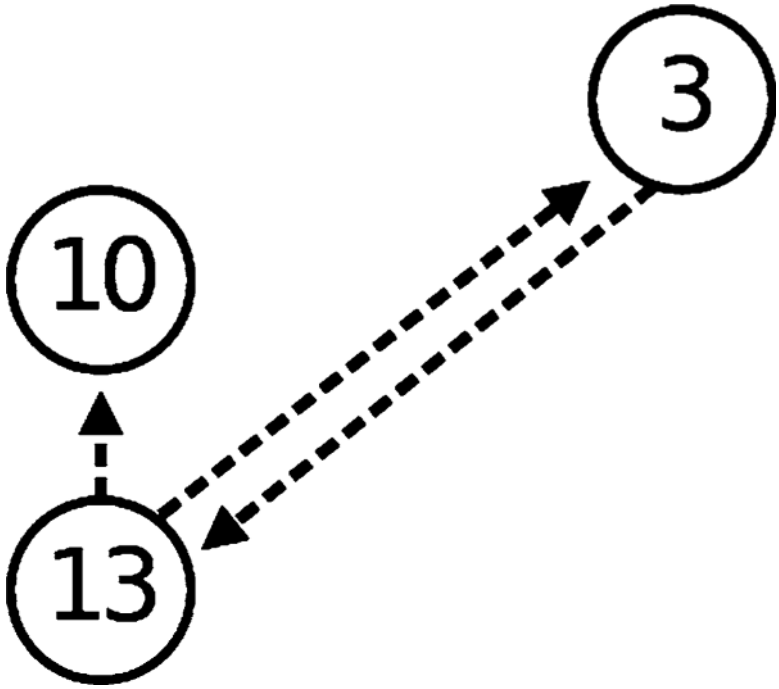

Fig. 19.

The resulting subgraph after computing causally conditioned directed information estimates. $\widehat{\mathcal{I}}(3 \to 10||13) = 0$ and $\widehat{\mathcal{I}}(13 \to 10||3) > 0$, so 3→10 was removed, and 13→10 was kept

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

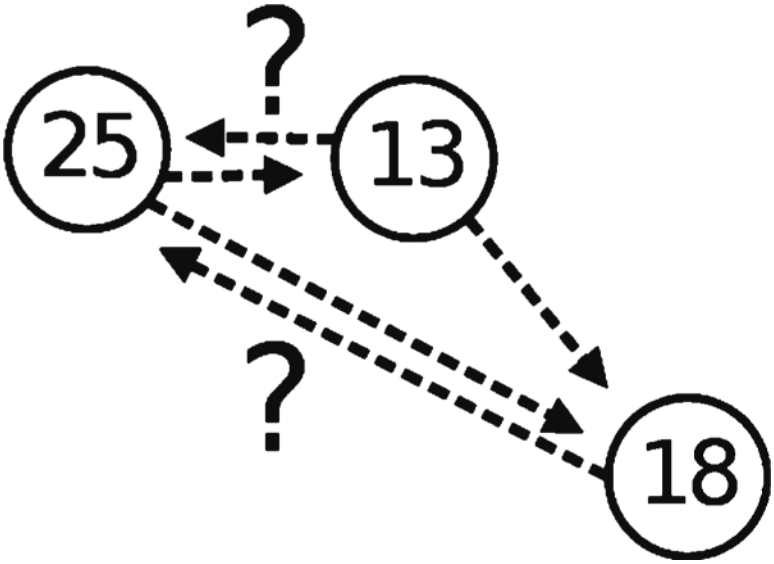

Fig. 20.

Diagram depicting the induced subgraph of neurons 13, 18, and 25. Both 13 and 18 have pairwise influences into 25, one of which might be due to an indirect influence. A *question mark* is drawn adjacent to the *arrows* in question

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

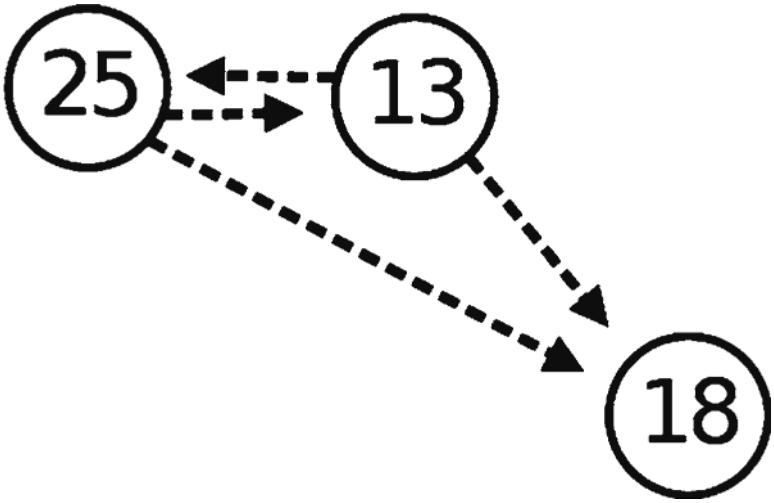

Fig. 21.

The resulting subgraph after computing causally conditioned directed information estimates. $\widehat{\mathcal{I}}(13 \to 25 \| 18) > 0$ and $\widehat{\mathcal{I}}(18 \to 25 \| 13) = 0$, so 18→25 was removed, and 13→25 was kept