

© 2010 Christopher John Quinn

ESTIMATING DIRECTED INFORMATION TO INFER CAUSAL
RELATIONSHIPS BETWEEN NEURAL SPIKE TRAINS AND
APPROXIMATING DISCRETE PROBABILITY DISTRIBUTIONS
WITH CAUSAL DEPENDENCE TREES

BY

CHRISTOPHER JOHN QUINN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Adviser:

Assistant Professor Todd P. Coleman

ABSTRACT

This work examines an information theoretic quantity known as directed information, which measures statistically causal influences between processes. It is shown to be a general quantity, applicable to arbitrary probability distributions. It is interpreted in terms of prediction, communication with feedback, source coding with feed forward, control over noisy channels, and other settings. It is also shown to be consistent with Granger’s philosophical definition. The concepts of direct and indirect causation in a network of processes are formalized. Next, two applications of directed information are investigated.

Neuroscience researchers have been attempting to identify causal relationships between neural spike trains in electrode recordings, but have been doing so with correlation measures and measures based on Granger causality. We discuss why these methods are not robust, and do not have statistical guarantees. We use a point process GLM model and MDL (as a model order selection tool) for consistent estimation of directed information between neural spike trains. We have successfully applied this methodology to a network of simulated neurons and electrode array recordings.

This work then develops a procedure, similar to Chow and Liu’s, for finding the “best” approximation (in terms of KL divergence) of a full, joint distribution over a set of random processes, using a causal dependence tree distribution. Chow and Liu’s procedure had been shown to be equivalent to maximizing a sum of mutual informations, and the procedure presented here is shown to be equivalent to maximizing a sum of directed informations. An algorithm is presented for efficiently finding the optimal causal tree, similar to that in Chow and Liu’s work.

To my parents, for their love and support

ACKNOWLEDGMENTS

I would especially like to thank my wonderful advisers, Todd Coleman and Negar Kiyavash, for all of their help and patience. I would also like to thank my family, friends, and everyone else who has helped me.

TABLE OF CONTENTS

CHAPTER 1	INTRODUCTION	1
CHAPTER 2	DEFINITIONS	3
CHAPTER 3	DIRECTED INFORMATION AS A ROBUST MEASURE OF STATISTICAL CAUSALITY	8
3.1	Background - Granger Causality	8
3.2	Definition of Directed Information	9
3.3	Example of Measuring Causal Influences	10
3.4	Interpretations	12
CHAPTER 4	CAUSAL RELATIONSHIPS IN A NETWORK OF PROCESSES	17
4.1	Causal Conditioning and Direct, Causal Conditioning	18
4.2	Graphical Depiction and Indirect Influences	19
4.3	Causal Conditioning and Directed Information	21
CHAPTER 5	CONSISTENT ESTIMATION OF DIRECTED INFORMATION BETWEEN NEURAL SPIKE TRAINS	24
5.1	Previous Approaches to Identify Causal Relationships in Neural Data	24
5.2	Estimation	27
CHAPTER 6	RESULTS	41
6.1	Simulated Data	41
6.2	Experimental Data	49
CHAPTER 7	APPROXIMATING DISCRETE PROBABILITY DISTRIBUTIONS WITH CAUSAL DEPENDENCE TREES	56
7.1	Introduction	56
7.2	Background: Dependence Tree Approximations	57
7.3	Main Result: Causal Dependence Tree Approximations	59
CHAPTER 8	CONCLUSION	63
REFERENCES	64

CHAPTER 1

INTRODUCTION

In recent decades, there has been much interest in investigating causal influences between stochastic processes in neuroscience, economics, communications, social sciences, and other disciplines. The first formal definition of causality was proposed by Granger [1]. 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” [1]. However, his statistical definition, which uses linear models, is not robust enough to model networks of arbitrary random processes. Although there have been many variations of Granger causality, they have all required some restrictive modelling for consistent results. In the past two decades, there has been work in information theory to quantify causality in a more general and meaningful way.

Directed information, the information theoretic quantification of causality, was formally introduced by Massey [2]. It was motivated by Marko’s work [3]. Related work was independently done by Rissanen and Wax [4]. It has since been investigated in a number of research settings, and shown to play a fundamental role in communication with feedback [3, 5, 6, 7, 2, 8], prediction with causal side info [4], gambling with causal side information [9, 10], control over noisy channels [11, 12, 13, 14, 6], and source coding with feed forward [15, 10]. Conceptually, mutual information and directed information are related. However, while mutual information quantifies correlation (in the colloquial sense of statistical interdependence), directed information quantifies *causation*.

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 neu-

rons. 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. The causal measures used so far have had limited success, due to tacitly strong assumptions placed on the data and lack of robustness. This paper presents a novel, provably-robust technique which is philosophically consistent with the widely used Granger causality and based on recent advances in information theory. In particular, this procedure models neural spike trains with point process generalized linear models, performs parameter and model order selection using maximum likelihood estimates and minimum description length, and calculates estimates for the directed information. The procedure is tested on a simulated network of neurons, for which it correctly identifies all of the relationships (whether there is an influence or not) and yields a quantity which can be interpreted as the *strength* of each influence. This procedure is also used to analyze ensemble spike train recordings in the primary motor cortex of an awake monkey performing target reaching tasks. The procedure identified strong structure which is consistent with predictions made from the wave propagation of simultaneously recorded local field potentials.

Chow and Liu developed an approach for approximating discrete joint probability distributions with dependence tree distributions [16]. For joint distributions of multiple random processes, their method would find the “best” dependence tree approximation (in terms of KL divergence) of the joint distribution, but would intermix the processes and ignore timing information. In this work, a method is presented which approximates a joint distribution of multiple random processes with a causal dependence tree approximation, in which the processes are not intermixed and the timing information is maintained. Chow and Liu’s method involves maximizing a sum of mutual informations, and this method involves maximizing a sum of directed informations. An algorithm is presented for efficiently finding the optimal causal tree, similar to that in Chow and Liu’s work.

CHAPTER 2

DEFINITIONS

This section presents probabilistic notations and information-theoretic definitions and identities that will be used throughout the remainder of the manuscript. Unless otherwise noted, the definitions and identities come from Thomas and Cover [17].

- For integers $i \leq j$, define $x_i^j \triangleq (x_i, \dots, x_j)$. For brevity, define $x^n \triangleq x_1^n = (x_1, \dots, x_n)$.
- Throughout this paper, \mathcal{X} corresponds to a measurable space that a random variable, denoted with upper-case letters (X), takes values in. Lower-case values $x \in \mathcal{X}$ correspond to specific realizations.
- Define the probability mass function (PMF) of a discrete random variable by

$$P_X(x) \triangleq P(X = x).$$

- For a length n , discrete random vector, denoted as $X^n \triangleq (X_1, \dots, X_n)$, the joint PMF is defined as

$$P_{X^n}(x^n) \triangleq P(X^n = x^n).$$

Let $P_{X^n}(\cdot)$ denote $P_{X^n}(x^n)$ (when the argument is implied by context).

- For two random vectors X^n and Y^m , the conditional probability $P(X^n|Y^m)$ is defined as

$$P(X^n|Y^m) \triangleq \frac{P(X^n, Y^m)}{P(Y^m)}.$$

- The chain rule for joint probabilities is

$$P_{X^n|Y^n}(x^n|y^n) = \prod_{i=1}^n P_{X_i|X^{i-1}, Y^n}(x_i|x^{i-1}, y^n).$$

- The entropy of a discrete random variable is given by

$$H(X) \triangleq \sum_{x \in \mathcal{X}} -P_X(x) \log P_X(x) = E_{P_X} [-\log P_X(X)]. \quad (2.1)$$

- The conditional entropy is given by

$$H(Y|X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} -P_{X,Y}(x, y) \log P_{Y|X}(y|x). \quad (2.2)$$

- The chain rule for entropy is given by

$$H(X^n) = \sum_{i=1}^n H(X_i|X^{i-1}). \quad (2.3)$$

- For two probability distributions P and Q on \mathcal{X} , the Kullback-Leibler divergence is given by

$$D(P\|Q) \triangleq E_P \left[\log \frac{P(X)}{Q(X)} \right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \geq 0. \quad (2.4)$$

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

$$I(X; Y) \triangleq D(P_{XY}(\cdot, \cdot) \| P_X(\cdot) P_Y(\cdot)) \quad (2.5a)$$

$$= E_{P_{XY}} \left[\log \frac{P_{Y|X}(Y|X)}{P_Y(Y)} \right] \quad (2.5b)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y) \log \frac{P_{Y|X}(y|x)}{P_Y(y)} \quad (2.5c)$$

$$= H(Y) - H(Y|X). \quad (2.5d)$$

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

- The chain rule for mutual information is given by

$$I(X^n; Y^n) = \sum_{i=1}^n I(Y_i; X^n | Y^{i-1}) \quad (2.6)$$

$$= \sum_{i=1}^n D(P_{Y_i|X^n, Y^{i-1}} \| P_{Y_i|Y^{i-1}}), \quad (2.7)$$

where the conditional mutual information is given by

$$I(X; Y|Z) = \mathbb{E}_{P_{XYZ}} \left[\log \frac{P_{Y|X,Z}(Y|X, Z)}{P_{Y|Z}(Y|Z)} \right]. \quad (2.8)$$

- We denote the set of k th order Markov chains as

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

with $X_j \triangleq \emptyset$ 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 $\text{SE}(\mathcal{X})$.
- The entropy rate and mutual information rate, assuming they exist, are given as follows:

$$\mathcal{H}(Y) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} H(Y^n) \quad (2.9)$$

$$\mathcal{I}(X; Y) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} I(X^n; Y^n). \quad (2.10)$$

- 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 $y : (0, T] \rightarrow \mathbb{Z}_+$ 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:

$$\mathcal{F}_t = \sigma(X_\tau : \tau \in [0, t], Y_\tau : \tau \in [0, t)) \quad (2.11a)$$

$$\mathcal{F}'_t = \sigma(Y_\tau : \tau \in [0, t)). \quad (2.11b)$$

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 [18]. The CIF is defined as [19]

$$\lambda(t|\mathcal{F}_t) \triangleq \lim_{\Delta \rightarrow 0} \frac{P(Y_{t+\Delta} - Y_t = 1 | \mathcal{F}_t)}{\Delta}, \quad (2.12a)$$

$$\lambda(t|\mathcal{F}'_t) \triangleq \lim_{\Delta \rightarrow 0} \frac{P(Y_{t+\Delta} - Y_t = 1 | \mathcal{F}'_t)}{\Delta}. \quad (2.12b)$$

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 [20] implicitly use this *causal* assumption in the definition of \mathcal{F}_t given by (2.11). 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 $\lambda(t|\mathcal{F}_t)$ and $\lambda(t|\mathcal{F}'_t)$, the likelihood or density of Y at y given x is given by [18]

$$f_{Y\|X}(y\|x; \lambda) = \exp \left\{ \int_0^T \log \lambda(t|\mathcal{F}_t) dy_t - \lambda(t|\mathcal{F}_t) dt \right\}, \quad (2.13)$$

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

$$f_Y(y; \lambda) = \exp \left\{ \int_0^T \log \lambda(t|\mathcal{F}'_t) dy_t - \lambda(t|\mathcal{F}'_t) dt \right\}. \quad (2.14)$$

We use the $\|$ notation to explicitly speak to how these conditional probabilities in (2.12) are taken with respect to causal histories, specified in (2.11). By discretizing $(0, T]$ into $n = T/\Delta$ intervals of length $\Delta \ll 1$ so that $dy = (dy_1, \dots, dy_n)$ with $dy_i \triangleq y_{(i+1)\Delta} - y_{i\Delta} \in \{0, 1\}$, we can

approximate (2.13) and (2.14) by

$$-\log f_{Y\|X}(y\|x; \lambda) \simeq \sum_{i=1}^n -\log \lambda(i\|\mathcal{F}_i) dy_i + \lambda(i\|\mathcal{F}_i) \Delta \quad (2.15)$$

$$-\log f_Y(y; \lambda) \simeq \sum_{i=1}^n -\log \lambda(i\|\mathcal{F}'_i) dy_i + \lambda(i\|\mathcal{F}'_i) \Delta, \quad (2.16)$$

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 as

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

The function (h_1, \dots, h_K) operates on the extrinsic covariate X in the recent past. We subsequently define $\text{GLM}(h)$ as

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

CHAPTER 3

DIRECTED INFORMATION AS A ROBUST MEASURE OF STATISTICAL CAUSALITY

3.1 Background - Granger Causality

One of the best established methods for identifying causation in a network of processes is 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” [1]. 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 [1]. While these assumptions might be acceptable for econometrics, for which Granger’s work was originally designed, they are not for point processes (binary sequences) such as neural spike trains, where variances of correction terms to linear models are not as meaningful. Thus, this statistical measure of causality is not general, and without strong assumptions is not consistent (if data not coming from a linear model).

Two decades after Granger’s work, Rissanen and Massey, both Shannon award winners, independently introduced a different functional definition of causality [4, 2]. Massey, whose work is based on earlier work by Marko [3], 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, however, directed information is not tied to any particular modeling class.

3.2 Definition of Directed Information

The directed information from a process \mathbf{X} to a process \mathbf{Y} , both of length n , is defined by

$$I(X^n \rightarrow Y^n) \triangleq \sum_{i=1}^n I(Y_i; X^i | Y^{i-1}) \quad (3.1a)$$

$$= \sum_{i=1}^n \mathbb{E}_{P_{X^i, Y^i}} \left[\log \frac{P_{Y_i | X^i, Y^{i-1}}(Y_i | X^i, Y^{i-1})}{P_{Y_i | Y^{i-1}}(Y_i | Y^{i-1})} \right] \quad (3.1b)$$

$$= \sum_{i=1}^n D(P_{Y_i | X^i, Y^{i-1}}(\cdot) \| P_{Y_i | Y^{i-1}}(\cdot)) \quad (3.1c)$$

$$= \mathbb{E}_{P_{X^n, Y^n}} \left[\log \frac{\prod_{i=1}^n P_{Y_i | X^i, Y^{i-1}}(Y_i | X^i, Y^{i-1})}{\prod_{i=1}^n P_{Y_i | Y^{i-1}}(Y_i | Y^{i-1})} \right] \quad (3.1d)$$

$$= \mathbb{E}_{P_{X^n, Y^n}} \left[\log \frac{P_{Y^n | X^n}(Y^n | X^n)}{P_{Y^n}(Y^n)} \right] \quad (3.1e)$$

$$= D(P_{Y^n | X^n}(\cdot) \| P_{Y^n}(\cdot)) \quad (3.1f)$$

$$= \mathbb{E}_{P_{X^n, Y^n}} [-\log P_{Y^n}(Y^n)] - \mathbb{E}_{P_{X^n, Y^n}} [-\log P_{Y^n | X^n}(Y^n | X^n)] \quad (3.1g)$$

$$= H(Y^n) - H(Y^n | X^n), \quad (3.1h)$$

where the *causally conditioned entropy*, $H(Y^n | X^n)$, is defined as

$$H(Y^n | X^n) \triangleq \sum_{i=1}^n H(Y_i | Y^{i-1}, X^i). \quad (3.2)$$

Causal conditioning was introduced by Kramer [5]. The difference between mutual information (2.6) and directed information (3.1a) is the change of X^n to X^i . Conceptually, this looks at how the present and the past X^i and present Y_i are correlated (conditioned upon Y^{i-1}), but ignores how the future X_{i+1}^n and present Y_i are correlated (conditioned upon Y^{i-1}) at each time step. Thus, it only takes into the account the *causal* influence of process \mathbf{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 and thus is well-defined for arbitrary joint probability distributions (for example, of point processes [21, 22]). As one can determine a “degree of correlation” (statisti-

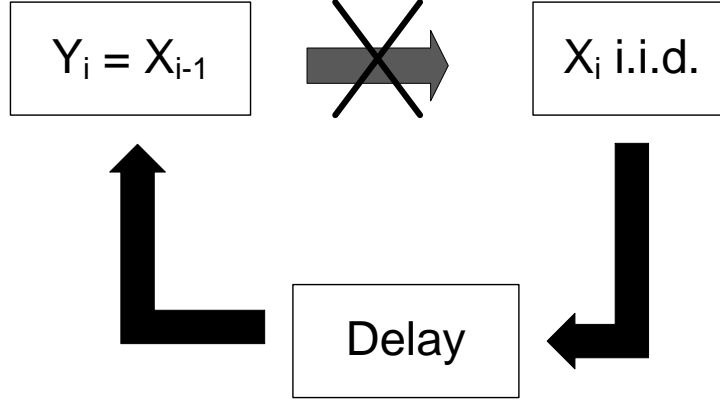


Figure 3.1: Diagram of the processes and their causal relationship. \mathbf{X} is drawn i.i.d. equi-probably to be 0 or 1, and $Y_i = X_{i-1}$. Clearly \mathbf{X} is causally influencing \mathbf{Y} . Moreover, \mathbf{Y} is *not* causally influencing \mathbf{X} .

cal 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* \mathbf{Y} is statistically causally influenced by \mathbf{X} . It is not clear that the values of Granger causality and related measures have similar meaning.

3.3 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 [8]. 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 3.1 depicts the relationship between the processes. Calculating the normalized mutual

information between \mathbf{X} and \mathbf{Y} ,

$$\frac{1}{n}I(X^n; Y^n) = \frac{1}{n} \sum_{i=1}^n I(Y_i; X^n | Y^{i-1}) \quad (3.3)$$

$$= \frac{1}{n} \left[I(Y_1; X^n) + \sum_{i=2}^n I(Y_i; X^n | Y^{i-1}) \right]$$

$$= \frac{1}{n} \left[I(X_0; X^n) + \sum_{i=2}^n I(X_{i-1}; X^n | X^{i-2}) \right] \quad (3.4)$$

$$= \frac{1}{n} \left[\sum_{i=2}^n H(X_{i-1} | X^{i-2}) - H(X_{i-1} | X^n) \right] \quad (3.5)$$

$$= \frac{1}{n} \sum_{i=2}^n (1 - 0) = \frac{1}{n} (n - 1), \quad (3.6)$$

where (3.3) follows from (2.6), (3.4) follows from substituting $Y_i = X_{i-1}$, (3.5) follows from (2.8), and (3.6) is due to Y_i s being i.i.d. Bernoulli (1/2). Taking the limit, $\lim_{n \rightarrow \infty} \frac{1}{n}I(X^n; Y^n) = 1$. The mutual information detects a strong relationship, but offers no 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 \mathbf{Y} to \mathbf{X} is

$$\frac{1}{n}I(Y^n \rightarrow X^n) = \frac{1}{n} \sum_{i=1}^n I(Y^i; X_i | X^{i-1})$$

$$= \frac{1}{n} \left[I(Y_1; X_1) + \sum_{i=2}^n I(Y^i; X_i | X^{i-1}) \right]$$

$$= \frac{1}{n} \left[0 + \sum_{i=2}^n I(X^{i-1}; X_i | X^{i-1}) \right] \quad (3.7)$$

$$= 0, \quad (3.8)$$

where in (7.6), X^{i-1} is substituted for Y^i , and (7.7) follows because the X_i s are i.i.d. The normalized directed information in the reverse direction is

$$\begin{aligned} \frac{1}{n}I(X^n \rightarrow Y^n) &= \frac{1}{n} \sum_{i=1}^n I(X^i; Y_i | Y^{i-1}) \\ &= \frac{1}{n} \left[I(X_1; Y_1) + \sum_{i=2}^n I(X^i; Y_i | Y^{i-1}) \right] \\ &= \frac{1}{n} \left[0 + \sum_{i=2}^n I(X^i; X_{i-1} | X^{i-2}) \right] \end{aligned} \quad (3.9)$$

$$= \frac{1}{n} \left[\sum_{i=2}^n H(X_{i-1} | X^{i-2}) - H(X_{i-1} | X^i) \right] \quad (3.10)$$

$$= \frac{1}{n} \sum_{i=2}^n (1 - 0) = \frac{1}{n}(n - 1), \quad (3.11)$$

where (3.9) follows with X_{i-1} substituted for Y_i , (3.10) follows from (2.8), and (3.11) is due to the X_i s being i.i.d. Therefore, $\lim_{n \rightarrow \infty} \frac{1}{n}I(X^n \rightarrow Y^n) = 1$. This example demonstrates the merit of directed information in causal inference as it correctly characterizes the direction of information flow while mutual information fails to do so.

3.4 Interpretations

3.4.1 Directed Information and Prediction

Directed information has an important “information gain” interpretation of the divergence with respect to prediction and source coding. Consider Shannon code lengths. Shannon codelengths are the “ideal” codelengths (description lengths) of a random sequence in the sense that the lengths of all code-words to describe all possible realizations of any sequence from a source are within one bit of their theoretical ideal limit. Shannon codes are a function of the random sequence’s probability distribution [17]:

$$l_{Shannon}(x) \triangleq -\log P_X(x),$$

where the random variable X is generated from distribution $P_X(\cdot)$ (for clarity ignore integer constraint). The expected codelength using the Shannon code is the entropy $H(X)$ [17]. Note that if the wrong distribution is used, $Q(\cdot)$, there is a penalty in the expected excess codelength

$$\mathbb{E}_{P_X} \left[\log \frac{1}{Q(X)} - \log \frac{1}{P_X(X)} \right] = D(P_X \| Q),$$

which is measured by the KL divergence.

Now consider mutual information. For random variables X and Y with joint distribution $P_{X,Y}$, the mutual information (2.5) is a function of the log likelihood ratio of the joint distribution to the product of the marginals and measures how the random variables X and Y are statistically related. The mutual information is nonzero if and only if the two random variables are statistically independent. The symmetric structure of the mutual information, $I(X; Y) = I(Y; X)$, implies that mutual information only measures the correlation - in the colloquial sense of statistical interdependence. 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 (2.5a). Alternatively, by dividing through by P_X in (2.5a) (equivalent to adding $\log \frac{P_X}{P_X} = 0$),

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

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 (2.7),

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

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

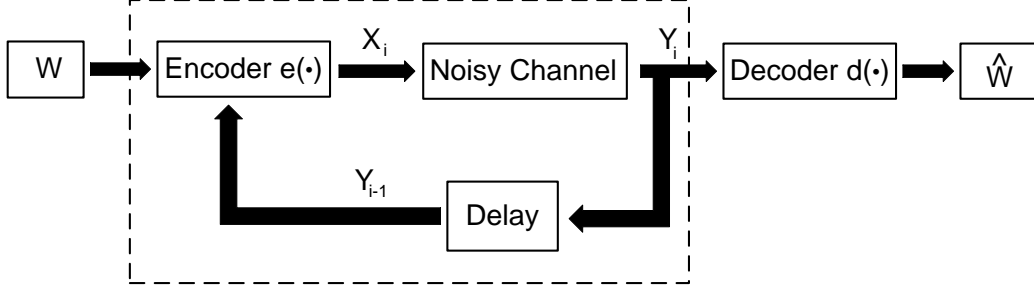


Figure 3.2: Diagram of a noisy channel. The capacity of the noisy channel without feedback is a function of $I(X^n; Y^n)$. With feedback, the capacity of the noisy channel changes. The capacity of the whole channel (inside the dotted line), which includes both the noisy channel and the feedback, is always a function of $I(W; Y^n) = I(X^n \rightarrow Y^n)$.

(3.1c):

$$I(X^n \rightarrow Y^n) = \sum_{i=1}^n D(P_{Y_i|X^i, Y^{i-1}}(\cdot) \| P_{Y_i|Y^{i-1}}(\cdot)).$$

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 viewpoint, which was motivated by earlier work by Wiener [1]. However, it operationally differs from Granger's measure in that it is well accepted that Shannon codelengths capture the uncertainty (the difficulty in prediction) of random variables, while there exists no such belief about the variance of error terms for linear models.

3.4.2 Communication with Feedback

Consider communication of a message W across the stochastic channel using n channel uses, without feedback. An encoder denoted by e converts the message to a suitable format for transmission of each symbol X_i over the noisy channel. A decoder denoted by $d(\cdot)$ tries to recover the message after observing all n outputs of the channel, Y^n . At time step i , the encoder takes the message W and transmits $X_i = e_i(W)$ across the channel. The decoder takes the channel outputs Y^n and forms an estimate of the original message $\hat{W} = d(Y^n)$. It is assumed that the communication designer only has control over the encoding and decoding strategies, but not the message or the channel. Since the message W will be assumed to be a random variable,

and the channel is noisy, \widehat{W} will also be a random variable. To communicate W reliably, it can be shown that the “essence” of this problem is to design $e(\cdot)$ and subsequently $d(\cdot)$ to maximize the mutual information $I(W; Y^n)$. In the absence of feedback, it can be shown that maximizing $I(W; Y^n)$ is equivalent to maximizing $I(X^n; Y^n)$ [17].

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 Figure 3.2. In the presence of feedback, $I(W; Y^n)$ can be re-written as:

$$\begin{aligned} I(W; Y^n) &= H(Y^n) - H(Y^n|W) \\ &= H(Y^n) - \sum_{i=1}^n H(Y_i|Y^{i-1}, W) \end{aligned} \quad (3.12)$$

$$= H(Y^n) - \sum_{i=1}^n H(Y_i|Y^{i-1}, W, X^i) \quad (3.13)$$

$$= H(Y^n) - \sum_{i=1}^n H(Y_i|Y^{i-1}, X^i) \quad (3.14)$$

$$= H(Y^n) - H(Y^n||X^n) \quad (3.15)$$

$$= I(X^n \rightarrow Y^n), \quad (3.16)$$

where (3.12) follows from entropy chain rule (2.3), (3.13) holds because X^i is a deterministic function of W and Y^{i-1} , (3.14) is due to W , X_i , and Y_i being a Markov chain, in the sense that the statistical nature of the channel is linked to W only through the inputs X , (3.15) follows from definition of causal entropy in (3.2), and (3.16) follows from (3.1h). Therefore, maximizing $I(W; Y^n)$ is equivalent to maximizing $I(X^n \rightarrow Y^n)$. Only in the case of no feedback are $I(W; Y^n)$, $I(X^n \rightarrow Y^n)$, and $I(X^n; Y^n)$ equivalent. Without feedback, knowledge of W is statistically sufficient to have knowledge of X^n , so in (3.13), X^i would get changed to X^n . Therefore, directed information is the function that (when maximized) characterizes the capacity of the *whole* channel (the original noisy channel and effect of feedback) between the original message W and output Y^n .

Several researchers have investigated the role of directed information in this context [5, 6, 7, 2]. Massey showed that the mutual information between the input X^n and the output Y^n is the sum of the causal influence of the input on the output (equivalently, the mutual information without any feedback)

plus the effect of the feedback [8]:

$$\frac{1}{n}I(X^n; Y^n) = \frac{1}{n}I(X^n \rightarrow Y^n) + \frac{1}{n}I(0 * Y^{n-1} \rightarrow X^n),$$

where $0 * Y^{n-1}$ denotes any constant 0 concatenated with Y^{n-1} . It is this factor $I(0 * Y^{n-1} \rightarrow X^n)$, the effect of the feedback, which the mutual information $I(X^n; Y^n)$ “overcounts” that causes it to not characterize the mutual information $I(W; Y^n)$ of the whole channel (in the presence of feedback).

3.4.3 Other Interpretations

In addition to the prediction and channel communication perspectives, there are also other ways of examining directed information. 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 [9, 10]. 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 [10]. Venkataramanan and Pradhan examined directed information in the setting of source coding with feed forward, and they showed that the rate distortion function is characterized by it [15]. There has also been work investigating the role of directed information in characterizing control over noisy channels with causal feedback [11, 14, 6, 13, 12]. Shalizi has examined the role of directed information in the framework of self-organization [23]. Lastly, there has been some work identifying relationships between directed information and other causal influence measures [24].

The work in this paper assumes a discrete time setting. Kim et al. have begun extending the directed information framework into the continuous time, and examine its meaning for continuous time communication channels [25]. However, that work does not propose an estimation scheme.

CHAPTER 4

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 where researchers want to 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 stochastic problems. Bayesian networks, or “belief networks,” define causality between random variables by using properties of the joint distribution [26]. 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 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.

4.1 Causal Conditioning and Direct, Causal Conditioning

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

$$P_{\mathbf{A}||\mathbf{B}}(\cdot) = P_{A_1^n||B_1^n}(\cdot) \triangleq \prod_{i=1}^n P_{A_i|A^{i-1}, B^i}(\cdot). \quad (4.1)$$

Define causal influences as follows. Let V be a set of $m+1$ random processes, $V = \{\mathbf{X}_1, \dots, \mathbf{X}_m, \mathbf{Y}\}$, where each process is a length n vector, $\forall \mathbf{Z} \in V$, $\mathbf{Z} = (Z_i)_{i=1}^n$. The random process \mathbf{X}_i is said to *causally influence* the random process \mathbf{Y} iff

$$P_{\mathbf{Y}||\mathbf{X}_i}(\cdot) \neq P_{\mathbf{Y}}(\cdot). \quad (4.2)$$

Note that this definition only identifies if there is influence through *some* path, possibly directly. This form of influence will also be denoted as “pair-wise” 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. The random process \mathbf{X}_i is said to *directly*, causally influence the random process \mathbf{Y} with respect to V iff

$$\forall W \subseteq V \setminus \{\mathbf{X}_i, \mathbf{Y}\} \quad P_{\mathbf{Y}||W, \mathbf{X}_i}(\cdot) \neq P_{\mathbf{Y}||W}(\cdot). \quad (4.3)$$

Thus, even with causal knowledge of any of the other processes in the network, there is still some influence from \mathbf{X}_i to \mathbf{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 \mathbf{X}_i on another \mathbf{Y} , then it is said that \mathbf{X}_i has a direct influence on \mathbf{Y} . These conditions are related to the conditions of “d-separation” in Bayesian networks

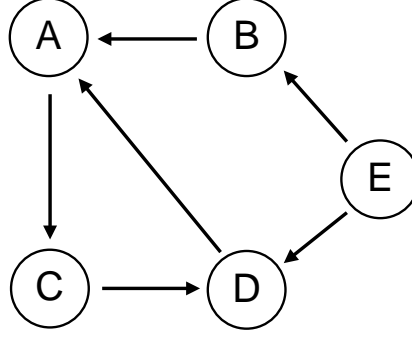


Figure 4.1: A graphical depiction of the direct, causal influences in a network of processes.

[26]. Let $V_{\mathbf{Y}}$ denote the set of all the processes that directly, causally influence \mathbf{Y} . Let $V'_{\mathbf{Y}}$ denote the set of all the processes that causally influence \mathbf{Y} . By the above definitions,

$$V_{\mathbf{Y}} \subseteq V'_{\mathbf{Y}}.$$

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

4.2 Graphical Depiction and Indirect Influences

Bayesian networks and other approaches to identifying causal relationships in networks often use directed graphs to depict the relationships [26]. 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 \mathbf{X}_i to process \mathbf{X}_j ($i \neq j$) iff \mathbf{X}_i directly, causally influences \mathbf{X}_j . Otherwise, let there be no arrow. An example is shown for processes A, B, C, D, and E in Figure 4.1.

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 \mathbf{X}_i to process \mathbf{X}_j ($i \neq j$) iff \mathbf{X}_i causally influences \mathbf{X}_j . Otherwise, let there be no arrow. An example is shown in Figure 4.2 for processes A, B, C, D, and E, which is consistent with the above graph for the direct influences. It is consistent because all of the direct, causal influences are present, and the extra arrows could be due to indirect influences, such as proxy effects ($A \rightarrow D$, $D \rightarrow C$, $C \rightarrow A$, and $B \rightarrow C$) and cascading effects ($B \rightarrow D$ and $D \rightarrow B$), which are discussed below.

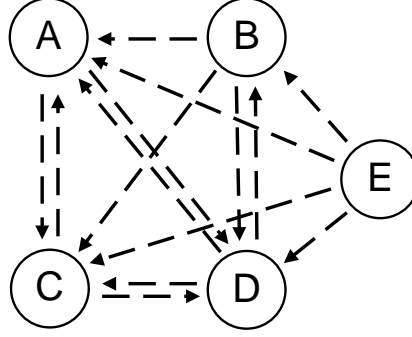


Figure 4.2: A graphical depiction of the causal influences in a network of processes.

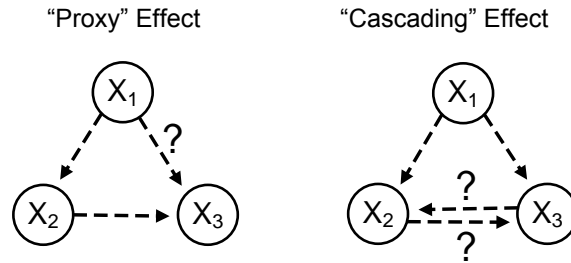


Figure 4.3: A graphical depiction of two types of indirect influences. Each arrow depicts a causal influence. The arrows with a question mark are the indirect influences.

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 \mathbf{X}_1 influences process \mathbf{X}_2 which in turn influences \mathbf{X}_3 , but with no direct influence from \mathbf{X}_1 to \mathbf{X}_3 . In some cases, there will be a causal influence from \mathbf{X}_1 to \mathbf{X}_3 *through* \mathbf{X}_2 (Figure 4.3), and causal knowledge of \mathbf{X}_2 renders \mathbf{X}_1 and \mathbf{X}_3 statistically independent. Thus, proxy effects can be considered analogous to the Markovicity property. Note that if there is a loop of direct, causal influence between a set of processes (such as $\mathbf{X}_1 \rightarrow \mathbf{X}_2$, $\mathbf{X}_2 \rightarrow \mathbf{X}_3$, $\mathbf{X}_3 \rightarrow \mathbf{X}_4$, and $\mathbf{X}_4 \rightarrow \mathbf{X}_1$), then the set could have causal influences from every process to all the others, due to proxy effects.

Another form of indirect influence is “cascading” influence. Here two processes \mathbf{X}_2 and \mathbf{X}_3 have a common influencing process \mathbf{X}_1 . Knowledge of \mathbf{X}_1 renders \mathbf{X}_2 and \mathbf{X}_3 statistically independent, but there is causal influence between the two, possibly accounted for by residual self dependence in \mathbf{X}_1 (Figure 4.3). Another, related type of indirect influence is an “inverted-

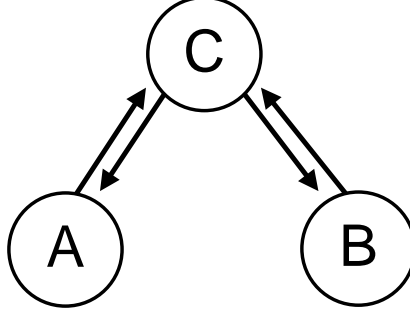


Figure 4.4: A graphical depiction of “inverted-cascade” influences.

cascading” influence (Figure 4.4). In this case, two processes \mathbf{X}_1 and \mathbf{X}_2 are statistically independent, and both have causal influences to and from third process, \mathbf{X}_3 . Additionally, causal knowledge of \mathbf{X}_3 induces causal influences between A and B. By the definitions of causal influence and direct, causal influence, the causal influences are all direct, causal influences, as pairwise influences between \mathbf{X}_1 and \mathbf{X}_2 are 0. An example of this is if the third process \mathbf{X}_3 is the product of two statistically independent processes ($\mathbf{X}_{3,i} = \mathbf{X}_{1,i} * \mathbf{X}_{2,i}$).

4.3 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 \mathbf{X}_i causally influences the process \mathbf{Y} iff*

$$I(\mathbf{X}_i \rightarrow \mathbf{Y}) > 0.$$

Proof. That causal influence implies positive directed information is proven as follows. $P_{\mathbf{Y}|\mathbf{X}_i}(\cdot) \neq P_{\mathbf{Y}}(\cdot)$ by definition. Recall that the KL distance is always positive for non-identical distributions. Thus $D(P_{\mathbf{Y}|\mathbf{X}_i}||P_{\mathbf{Y}}) > 0$.

Using this,

$$I(\mathbf{X}_i \rightarrow \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}_i) \quad (4.4)$$

$$= H(\mathbf{Y}) - E_{P_{\mathbf{Y}, \mathbf{X}_i}} [-\log P_{\mathbf{Y}|\mathbf{X}_i}(\mathbf{Y}|\mathbf{X}_i)] \quad (4.5)$$

$$= E_{P_{\mathbf{Y}, \mathbf{X}_i}} [-\log P_{\mathbf{Y}}(\mathbf{Y})] \\ - E_{P_{\mathbf{Y}, \mathbf{X}_i}} [-\log P_{\mathbf{Y}|\mathbf{X}_i}(\mathbf{Y}|\mathbf{X}_i)] \quad (4.6)$$

$$= E_{P_{\mathbf{Y}, \mathbf{X}_i}} \left[\log \frac{P_{\mathbf{Y}|\mathbf{X}_i}(\mathbf{Y}|\mathbf{X}_i)}{P_{\mathbf{Y}}(\mathbf{Y})} \right] \quad (4.7)$$

$$= D(P_{\mathbf{Y}|\mathbf{X}_i} \| P_{\mathbf{Y}}) \quad (4.8)$$

$$> 0. \quad (4.9)$$

Equation (4.4) follows from identity of directed information, (4.5) follows from definition of causally conditioned entropy, (4.6) follows from definition of entropy, (4.7) uses linearity of expectation and property of logarithm, (4.8) uses definition of KL distance, and (4.9) follows from beginning of proof. That positive directed information implies causal influence is proven using the reverse of above steps and applying assumption $I(\mathbf{X}_i \rightarrow \mathbf{Y}) > 0$. \square

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 \mathbf{X}_i , process \mathbf{Y} , and set of processes W as [5]

$$I(\mathbf{X}_i \rightarrow \mathbf{Y} | W) \triangleq H(\mathbf{Y} | W) - H(\mathbf{Y} | \mathbf{X}_i, W). \quad (4.10)$$

Lemma 2. $P_{\mathbf{Y} | W, \mathbf{X}_i}(\cdot) \neq P_{\mathbf{Y} | W}(\cdot)$ iff $I(\mathbf{X}_i \rightarrow \mathbf{Y} | W) > 0$.

Proof. The proof is identical to above, but causally conditioning on the set of processes W . \square

Theorem 3. *The random process \mathbf{X}_i directly, causally influences the random process \mathbf{Y} with respect to V iff*

$$\forall W \subseteq V \setminus \{\mathbf{X}_i, \mathbf{Y}\} \quad I(\mathbf{X}_i \rightarrow \mathbf{Y} | W) > 0. \quad (4.11)$$

Proof. Follows from lemma 2 and the definition of direct, causal influences. \square

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 $(\mathbf{X}_i, \mathbf{X}_j)$, compute $I(\mathbf{X}_i \rightarrow \mathbf{X}_j)$. If the value is positive, then there is causal influence from \mathbf{X}_i to \mathbf{X}_j , or $\mathbf{X}_i \rightarrow \mathbf{X}_j$. 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 also a causal influence, one could first identify all of the causal influences, and then determine which of those were also direct, causal influences. Consider two processes in V , \mathbf{X}_i and \mathbf{Y} , such that $I(\mathbf{X}_i \rightarrow \mathbf{Y}) > 0$. Thus, \mathbf{X}_i causally influences \mathbf{Y} . To determine if \mathbf{X}_i directly, causally influences \mathbf{Y} , one could check that for each $W \subseteq V \setminus \{\mathbf{X}_i, \mathbf{Y}\}$, $I(\mathbf{X}_i \rightarrow \mathbf{Y} || W) > 0$. If so, then \mathbf{X}_i directly, causally influences \mathbf{Y} , else if there is even one such W for which $I(\mathbf{X}_i \rightarrow \mathbf{Y} || W) = 0$, then the influence is not direct. Since some processes are statistically independent of \mathbf{X}_i and/or \mathbf{Y} , it can be helpful to focus on the subsets W which contain those \mathbf{X}_j 's such that each \mathbf{X}_j causally influences \mathbf{Y} and causally influences or is influenced by \mathbf{X}_i . For example, if there is a causal subgraph that could contain a proxy or cascading influence, then those indirect influences could be checked first.

CHAPTER 5

CONSISTENT ESTIMATION OF DIRECTED INFORMATION BETWEEN NEURAL SPIKE TRAINS

5.1 Previous Approaches to Identify Causal Relationships in Neural Data

5.1.1 Granger Causality and DTF

Granger causality [1] has been perhaps the most widely-established means of identifying causal relations between two time series [27]. 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 \mathbf{X} causally influences \mathbf{Y} , \mathbf{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 \mathbf{Y} is modeled again, but this time using the \mathbf{X} series as causal side information:

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

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 [28, 29]. The Granger causality is defined as

$$G_{\mathbf{X} \rightarrow \mathbf{Y}} \triangleq \log \frac{\text{var}(V)}{\text{var}(\tilde{V})}. \quad (5.1)$$

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

The directed transfer function (DTF) [30] is related to Granger causality, with the principle difference being that it transforms the autoregressive model into the spectral domain [31]. 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 (for experiments with several hundred, closely positioned electrodes recording in brain tissue, values obtained using Granger causality can be misleading [31]).

These and derivative techniques have been used extensively [27, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. These approaches can be attractive. They are generally fast to compute, with the main computational difficulty being the parameter and order estimation. Additionally, due to the simple model, they are easy to interpret. In some cases, they can identify statistically causal structures in the data. However, because of the sample-variance calculations, which are not necessarily statistically informative for point processes inference, these techniques are not necessarily reliable for analyzing neural spike trains. Autocorrelations and spectral transforms on binary time series data often do not work well and do not have meaningful, conceptual interpretations in this context. Moreover, these approaches do not have strong statistical guarantees of correctly identifying causal relations. They are only consistent under strong assumptions on the data, which might be acceptable in econometrics and finance, for which Granger causality was originally developed, but are generally not reasonable for neural data. 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, $G_{X \rightarrow Y}$, have a physical meaning beyond comparison with the opposite direction (e.g. $G_{X \rightarrow Y}$ v.s. $G_{Y \rightarrow X}$).

5.1.2 Transfer Entropy

Transfer entropy was developed by Schreiber [41]. It assumes two stochastic processes $\mathbf{X} = (X_i : i \geq 1)$ and $\mathbf{Y} = (Y_i : i \geq 1)$ satisfy a Markov property:

$$P_{Y_{n+1}|Y^n, X^n}(y_{n+1}|y^n, x^n) = P_{Y_{n+1}|Y_{n-J+1}^n, X_{n-K+1}^n}(y_{n+1}|y_{n-J+1}^n, x_{n-K+1}^n)$$

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

$$T_{X \rightarrow Y}(i) = I(Y_{i+1}; X_{i-K+1}^i | Y_{i-J+1}^i).$$

This term is part of the sum of terms (3.1a) that is equivalent to the directed information (with a Markov assumption applied). Some studies have employed this measure [42, 43, 44]. This has not been as widely employed as Granger causality and related measures, principally due to the lack of convergence properties [45]. As no model for the underlying distribution is suggested, the straightforward approach to estimate the transfer entropy is to use plug-in estimates, which are not consistent estimators for joint distributions (only for marginal distributions, under certain assumptions). There are no other, known estimation schemes for transfer entropy.

5.1.3 Dynamic Causal Modeling

Dynamic causal modeling (DCM) [46] 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 [46]. 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 [47, 48, 49, 50, 51], and EEG and MEG studies [52]. While it has been applied with some success to certain brain imaging studies, 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 proper-

ties of EM estimation [46], the model as a whole does not have statistically guaranteed convergence properties.

5.2 Estimation

These previous approaches are not provably good and do not have strong interpretations of measuring causal influences like directed information does. Now we will develop a procedure of consistently estimating the directed information between neural spike trains.

5.2.1 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 distribution $P_{\mathbf{X}, \mathbf{Y}}$ 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 principal 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 $I_n(\mathbf{X} \rightarrow \mathbf{Y}) \triangleq \frac{1}{n} I(X^n \rightarrow Y^n)$. Note from (3.1b) that I_n is a functional on the joint PMF of X^n and Y^n :

$$\begin{aligned} I_n(X \rightarrow Y) &= g_n(P_{X^n, Y^n}(\cdot, \cdot)) \\ &= \sum_{i=1}^n \mathbb{E}_{P_{X^n, Y^n}} \left[\log \frac{P_{Y_i | Y^{i-1}, X^i}(y_i | y^{i-1}, x^i)}{P_{Y_i | Y^{i-1}}(y_i | y^{i-1})} \right]. \end{aligned}$$

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 $P_{X^n, Y^n}(\cdot, \cdot)$. We denote the estimate of the density by $\hat{P}_{X^n, Y^n}(\cdot, \cdot)$. In general, $\hat{P}_{X^n, Y^n}(\cdot, \cdot)$ will not be a consistent estimate of $P_{X^n, Y^n}(\cdot, \cdot)$, 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

$$\hat{\mathcal{I}}_n(X \rightarrow Y) = g_n \left(\hat{P}_{X^n, Y^n}(\cdot, \cdot) \right)$$

will not be consistent. Note that for i.i.d. processes, there are consistent density estimators, but there are none (known) for general processes [53].

We note that making an *i.i.d.* sample assumption is not sensible 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.

There are procedures that attempt to directly estimate the functional on the joint distribution of interest. For information theoretic quantities such as entropy and Kullback-Leibler divergence, there are successful universal estimators, including Lempel-Ziv '77 [54], the Burroughs-Wheeler Transform (BWT) estimator [55], and context weighting tree methods [56]. Additionally, there has been work extending the context weighting tree method to estimating directed information [57]. Unfortunately, these methods are often computationally expensive and have slow convergence rates. There has also been some recent work by Perez-Cruz [58] for estimating numerous information theoretic quantities with better convergence rates and more moderate computational expense, but these procedures depend on i.i.d. assumptions.

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

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 $\mathcal{H}(\mathbf{Y})$, the causal entropy rate $\mathcal{H}(\mathbf{Y}||\mathbf{X})$, and the directed information rate $\mathcal{I}(\mathbf{X} \rightarrow \mathbf{Y})$ 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 equation (3.1h)).

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

Proof. First, prove that $\mathcal{H}(\mathbf{Y}|\mathbf{X})$. This proof closely follows the proof for the unconditional entropy rate in [17]. An important theorem used for the proof is the Cesaro mean theorem [17]: For sequences of real numbers (a_1, \dots, a_n) and (b_1, \dots, b_n) , if $\lim_{n \rightarrow \infty} a_n = a$, and $b_n = \frac{1}{n} \sum_{i=1}^n a_n$, then $\lim_{n \rightarrow \infty} b_n = a$.

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

$$0 \leq H(Y_i|Y^{i-1}, X^i) \leq H(Y_1) \quad \forall i.$$

Observe that

$$H(Y_i|Y^{i-1}, X^i) \leq H(Y_i|Y_2^{i-1}, X_2^i) \quad (5.2)$$

$$= H(Y_{i-1}|Y^{i-2}, X^{i-1}), \quad (5.3)$$

where (5.2) uses the property that conditioning reduces entropy (in reverse) and (5.3) uses stationarity. This sequence of real numbers, $a_i \triangleq H(Y_i|Y^{i-1}, X^i)$, is nonincreasing and bounded below by 0. Therefore, limit of a_n as $n \rightarrow \infty$ exists, and thus, by employing Cesaro mean theorem, $H(\mathcal{Y}|\mathcal{X}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} H(Y^n|X^n)$ exists.

Next, taking X^n to be a deterministic sequence, and following the above, $H(\mathcal{Y}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} H(Y^n)$ exists. Taking the limit in equation (3.1h), $\mathcal{I}(\mathbf{X} \rightarrow \mathbf{Y}) \triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \mathcal{I}(X^n \rightarrow Y^n)$ also exists. \square

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

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 \mathbf{X} and \mathbf{Y} which is physically important for any causal estimation paradigm.

The Markov model enables, among other things, the strong law of large numbers (SLNN) for Markov chains to hold [59]. Many Granger causality, DTF, and other previously discussed estimation procedures assume Markov-like assumptions [1, 30, 41], in addition to other constraints.

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

$$\frac{1}{n} H(Y^n || X^n) = E[g_{J,K}(Y_{i-J}^i, X_{i-(K-1)}^i)] \quad (5.4)$$

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

Proof. The normalized causal entropy can be rewritten as

$$\frac{1}{n} H(Y^n || X^n) = \frac{1}{n} \sum_{i=1}^n H(Y_i | X^i, Y^{i-1}) \quad (5.5)$$

$$= \frac{1}{n} \sum_{i=1}^n E[-\log_2 p_{Y_i|Y^{i-1}, X^i}(Y_i | Y^{i-1}, X^i)] \quad (5.6)$$

$$= \frac{1}{n} \sum_{i=1}^n E[-\log_2 p_{Y_i|Y_{i-J}^{i-1}, X_{i-(K-1)}^i}(Y_i | Y_{i-J}^{i-1}, X_{i-(K-1)}^i)] \quad (5.7)$$

$$= \frac{1}{n} \sum_{i=1}^n E[-\log_2 p_{Y_i|Y_{i-J}^{i-1}, X_{i-(K-1)}^i}(Y_i | Y_{i-J}^{i-1}, X_{i-(K-1)}^i)] \quad (5.8)$$

$$= \frac{1}{n} \sum_{i=1}^n E[-\log_2 p_{Y_i|Y_{i-J}^{i-1}, X_{i-(K-1)}^i}(Y_i | Y_{i-J}^{i-1}, X_{i-(K-1)}^i)] \quad (5.9)$$

$$= E[-\log_2 p_{Y_l|Y_{l-J}^{l-1}, X_{l-(K-1)}^l}(Y_l | Y_{l-J}^{l-1}, X_{l-(K-1)}^l)] \quad (5.10)$$

$$= E[g(Y_l | Y_{l-J}^{l-1}, X_{l-(K-1)}^l)], \quad (5.11)$$

where (5.5) is by definition of LHS, (5.6) is by definition of entropy, (5.7) uses the Markov assumption, (5.8) uses stationarity so the probability distribution is fixed, (5.9) uses stationarity within the expectation so the random variables have same expectation with time shifts, (5.10) is taking the sample mean of a constant, (5.11) renames the function to show it is a fixed function. \square

Since the right-hand side of equation (5.4) has no dependence on n , taking the limit of the above as $n \rightarrow \infty$ results in

$$\begin{aligned}\mathcal{H}(\mathbf{Y}||\mathbf{X}) &\triangleq \lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{H}(Y^n||X^n) \\ &= \mathbb{E} [g_{J,K}(Y_{i-J}^i, X_{i-(K-1)}^i)] .\end{aligned}$$

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 $P_{\mathbf{X},\mathbf{Y}} \in \mathcal{M}_{J,K}(\mathcal{X} \times \mathcal{Y})$. Then*

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

Proof. Taking the limit on both sides of equation (5.11) as $n \rightarrow \infty$,

$$\mathcal{H}(\mathbf{Y}||\mathbf{X}) = \mathbb{E} [g(Y_l|Y_{l-J}^{l-1}, X_{l-(K-1)}^l)] .$$

Using the SLLN for Markov chains [59], for a fixed function $g(\cdot)$ over the states of the Markov chain, as $n \rightarrow \infty$, the sample mean will converge almost surely to the expected value:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n g(Y_i|Y_{i-J}^{i-1}, X_{i-(K-1)}^i) &\xrightarrow{a.s.} \mathbb{E} [g(Y_l|Y_{l-J}^{l-1}, X_{l-(K-1)}^l)] \\ &= \mathcal{H}(\mathbf{Y}||\mathbf{X}).\end{aligned}$$

□

With these results, if a consistent estimate $\widehat{g}(\cdot)$ for the function $g(\cdot)$ can be found, then the sample mean of this function will converge almost surely to the causal entropy rate $\mathcal{H}(\mathbf{Y})$, and thus directed information rate can be estimated with almost sure convergence. Note that if \mathbf{Y} alone forms a discrete-time, finite state, stationary, and ergodic Markov chain, then this result can be used to estimate $\mathcal{H}(\mathbf{Y})$ by taking \mathbf{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 $\{h_k : k \geq 0\}$, $\lambda(i|\mathcal{F}_i) \in \mathbf{GLM}(h)$.

The recorded neural spiking activity - in millisecond time resolution - is known to be well-modeled using point process theory [60]. 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 [60] are a flexible class of parametric point process neural spiking models that allow for dependencies on a neuron's own past spiking, the spiking of other neurons, and extrinsic covariates. GLM models have the following conditional intensity:

$$\log \lambda(i|\mathcal{F}_i) = \alpha_0 + \sum_{j=1}^J \alpha_j dy_{i-j} + \sum_{k=1}^K \beta_k h_k(x_{i-(k-1)}), \quad (5.12)$$

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

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

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

$$\begin{aligned} -\frac{1}{n} \log f_{Y||X}(Y_1^n || X_1^n; \theta) \\ = \frac{1}{n} \sum_{i=1}^n -(\log(\lambda_\theta(i|\mathcal{H}_i)) dy_i - \lambda_\theta(i|\mathcal{H}_i) \Delta) \end{aligned} \quad (5.13)$$

$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^n g_\theta(Y_{i-J}^i, X_{i-(K-1)}^i) \quad (5.14) \\ &\xrightarrow{a.s.} \mathbb{E} [g_\theta(Y_{i-J}^i, X_{i-(K-1)}^i)] = \mathcal{H}(\mathbf{Y}||\mathbf{X}), \end{aligned}$$

where (5.14) 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 estimate the directed information, not just GLM.

5.2.3 Parameterized Estimation and MDL

Define $\Omega(J, K)$ to be vector space of possible parameters

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

If it is known a priori that $\lambda(i|\mathcal{F}_i) \in \text{GLM}_{J,K}(h)$, then θ can be consistently estimated using Assumptions 1–3 and a maximum likelihood estimate (MLE) [61]:

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

In practice, J and K are unknown. A model order selection procedure can be used to find estimates \hat{J}, \hat{K} , and subsequently $\hat{\theta} \in \Omega(\hat{J}, \hat{K})$ by penalizing “more complex” models, that is, those with larger $J + K$ values. The minimum description length (MDL) [62] is a model order selection procedure, which is known to have strong consistency guarantees [29]. In particular, under the assumption that $\lambda(i|\mathcal{F}_i) \in \text{GLM}(h)$, which means that $\theta \in \Omega(J, K)$, for some J and K , then it can be shown that an appropriately designed estimate $\hat{\theta} \rightarrow \theta_0$ *a.s.* Specifically, MDL selects the (\hat{J}, \hat{K}) and $\hat{\theta} \in \Omega(\hat{J}, \hat{K})$ according to

$$\begin{aligned} (\hat{J}, \hat{K}) &= \arg \min_{(J', K')} \min_{\theta' \in \Omega(J', K')} -\frac{1}{n} \log f_{Y||X}(Y_1^n || X_1^n; \theta') + \frac{J' + K'}{2n} \log n \\ &= \arg \min_{(J, K)} \min_{\theta' \in \Omega(J', K')} \frac{1}{n} \sum_{i=1}^n g_{\theta'}(Y_{i-J'}^i, X_{i-(K'-1)}^i) + \frac{J' + K'}{2n} \log n \\ \hat{\theta} &= \hat{\theta}(\hat{J}, \hat{K}). \end{aligned} \tag{5.15}$$

As K is the number of extrinsic parameters, if $\hat{K} = 0$, then *we say that no causal influence was detected*, since $\hat{\mathcal{H}}(\mathbf{Y}||\mathbf{X}) = \hat{\mathcal{H}}(\mathbf{Y})$ which implies that $\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y}) = 0$. 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 ($\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y}) = 0$). If $\hat{K} > 0$, there is a detected influence ($\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y}) > 0$).

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 1 spike per 10 ms, and let B have a very low average spiking rate, such as 1 spike per second (see Figure 5.1).

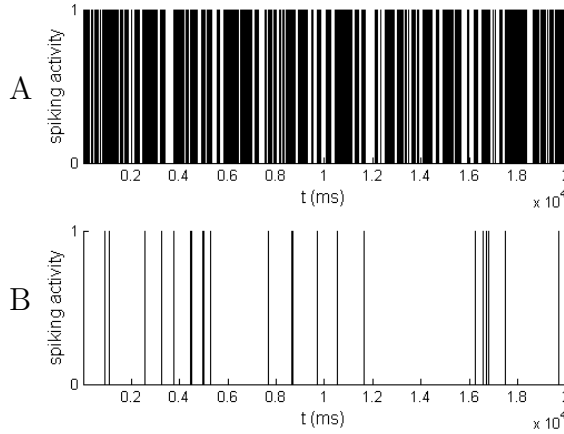


Figure 5.1: Spiking activity of neurons A (top) and B (bottom).

The best fitting GLM model (provided the data recording is sufficiently long) of neuron A using neuron B as an extrinsic process will have $\hat{K} \approx 12$ and $\{\beta_1, \dots, \beta_{\hat{K}}\}$ large and positive. 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 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.

5.2.4 The Proposed Estimation Procedure

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

1. Find \hat{J} , \hat{K} , and $\hat{\theta}$ according to the MDL procedure (5.15).
2. Calculate $\hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X})$ according to (5.14) using the estimated parameter values $\hat{\theta} \in \Omega(\hat{J}, \hat{K})$.
3. Compute an estimate for the unconditional entropy rate $\hat{\mathcal{H}}(\mathbf{Y})$ using a well-established entropy estimator (such as Lempel-Ziv '77 [54] or the BWT based estimator [55]).
4. Calculate the directed information rate estimate

$$\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y}) \triangleq \hat{\mathcal{H}}(\mathbf{Y}) - \hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X}).$$

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

$$\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y}) \xrightarrow{a.s.} \mathcal{I}(\mathbf{X} \rightarrow \mathbf{Y}). \quad (5.16)$$

- Proof.*
1. If Assumptions 1-3 hold, then the MDL procedure will identify the “true” parameter values $\theta \in \Omega(J, K)$ [29]: $\hat{J} \rightarrow J$ *a.s.*, $\hat{K} \rightarrow K$ *a.s.*, and $\hat{\theta} \rightarrow \theta$ *a.s.*
 2. Note that since $\hat{\theta} \rightarrow \theta$ *a.s.*, from the continuity of g_θ , $\hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X})$ specified above satisfies $\hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X}) \rightarrow \mathcal{H}(\mathbf{Y}|\mathbf{X})$ *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 $\mathcal{H}(\mathbf{Y})$ for stationary and ergodic finite-order Markov processes [63, 55].
 4. Combining these results,

$$\begin{aligned} \hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y}) &\triangleq \hat{\mathcal{H}}(\mathbf{Y}) - \hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X}) \\ &\xrightarrow{a.s.} \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{Y}|\mathbf{X}) \\ &= \mathcal{I}(\mathbf{X} \rightarrow \mathbf{Y}). \end{aligned}$$

□

5.2.5 Implementation Details

To perform the MDL search procedure, we select J' and K' values according to $J', K' \in \{0, 1, \dots, M\}$, 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 large \hat{J} and \hat{K} values, near M , then M can be increased adaptively to search for larger parameter orders (thus, it is not a *hard* limit). Choosing an M is done to save computation, in case the procedure settles on small values for \hat{J} and \hat{K} . If a researcher is only interested in local communications within a small brain region, then the researcher might pick a relatively small M [60]. For example, if the researcher anticipates that an upper bound for the maximum time scale for a spike from one neuron to influence another neuron in a recording (including time to propagate) is around 25 ms [64], then it would be appropriate to pick an $M \approx 25$. If a researcher is interested in motor feedback, such as with hand movement, then the longer delays for the signal to propagate should be taken into account, and so the researcher might pick a larger M , such as $M \approx 150$ [65].

For each (J', K') , the MLE parameter vector $\hat{\theta}(J', K')$ can be computed using the built-in Matlab function *glmfit*(\cdot), called with a Poisson link parameter. Then eq. (5.15) is computed to determine $\hat{\theta}$. The estimate for the causal entropy rate is taken to be the sample mean:

$$\hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n g_{\hat{\theta}}(Y_{i-\hat{J}}^i, X_{i-(\hat{K}-1)}^i).$$

To compute an estimate of the entropy rate, $\hat{\mathcal{H}}(\mathbf{Y})$, a universal estimator such as the BWT based estimator could be used (which has a faster convergence rate than LZ '77) [55] . 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, $\hat{\mathcal{H}}(\mathbf{Y}) - \hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X})$, then becomes the directed information estimate, $\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y})$.

In some cases, the relative influence of a process \mathbf{X} on a process \mathbf{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 :

$$\frac{\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y})}{\hat{\mathcal{H}}(\mathbf{Y})} = \frac{\hat{\mathcal{H}}(\mathbf{Y}) - \hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X})}{\hat{\mathcal{H}}(\mathbf{Y})} = 1 - \frac{\hat{\mathcal{H}}(\mathbf{Y}|\mathbf{X})}{\hat{\mathcal{H}}(\mathbf{Y})}. \quad (5.17)$$

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 [60]. It has been found that $\Delta = 1$ 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) [60]. 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 [60].

There are both potential benefits and dangers to choosing $\Delta > 1$ ms. One benefit of choosing $\Delta > 1$ ms 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 sparse (fewer zeros). There could then be more instances of multiple spikes within the $J' * \Delta$ or $K' * \Delta$ time windows to fit the α 's and β 's. Thus, the models might then have better fits. Also, for a fixed upper bound on the time scale over which causal influences will 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 ms, 5 ms, 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, so with 1 ms 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 slight delay. The loss of relevant timing information such as this could affect how well the procedure can 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 same neuron are smaller than the proposed Δ . However, there are no known studies that compare 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, guiding procedure for deciding when to choose $\Delta > 1$ ms and what Δ to choose in that case.

Computation time might be a factor in deciding the maximum model order M , the time resolution Δ , and the amount of data to use. This procedure was designed and tested in the Matlab environment, and used built-in Matlab functions (the code is available upon request). The computational complexity of this procedure is consequently unknown. The primary computational bottleneck is finding the α and β parameters for a given J and K . Using Matlab, this would be calls to *glmfit*(\cdot). In the tests (see Chapter 7), the author used data sets on the order of 100,000 elements. Data trials ran on computers with a 2.6 GHz processor (each estimate of $\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y})$ ran on a single computer). A single *glmfit*(\cdot) operation with $1 \leq J, K \leq 5$ took a few seconds. A single *glmfit*(\cdot) operation with $20 \leq J, K \leq 25$ took upwards of two minutes. The proposed procedure involves searching over a larger (J, K) parameter space for each ordered pair of processes. If $M = 25$, then there are $M * M = 625$ calls to *glmfit*(\cdot). For each ordered pair, this search took approximately 2 hours. With 6 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, the search space increased. For causal conditioning on two elements, the search space involves $M * M * (M * M) \approx 400,000$ calls to *glmfit*(\cdot). Since the runtime for that function changes with model order, the total time does not scale multiplicatively. The computations for different (J, K) orders can

be done in parallel. The author has not done any comprehensive time trials for different model orders and data lengths. Also, it is possible that other implementations of the GLM model fitting (which is a convex optimization procedure) could be substantially faster than the Matlab implementation, thus reducing computation times.

5.2.6 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 $\hat{\theta} \in \Omega(\hat{J}, \hat{K})$. 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 $\hat{\theta}$. 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) is convex in the parameters, the MLE will be the global maximum of the probability function

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

over the space of parameter values $\Omega(\hat{J}, \hat{K})$ [61]. 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 $I(z, \theta)^{-1}$ [61]:

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

in the neighborhood of θ_{MLE} . Using this, the approximate 95% confidence interval for the picked θ_{MLE} is [61]

$$\theta_{MLE} \pm \frac{1.96}{\sqrt{I(z, \theta)}}. \quad (5.18)$$

This can be interpreted as an interval about θ_{MLE} that with 95% probability contains the true parameter θ_0 .

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_i s, assume that only the β_i s from the best fitting model might vary from those of the true model. Assume that \hat{J} , \hat{K} , and $(\hat{\alpha}_j : 1 \leq j \leq J)$ 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 $l \in \{1, \dots, \hat{K}\}$, compute the $(l, l)^{th}$ entry of the observed Fisher information matrix:

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

With this value, the 95% confidence interval for this parameter $\hat{\beta}_l$ can be calculated using (5.18). 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.

CHAPTER 6

RESULTS

6.1 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 $\lambda(i|\mathcal{F}_i) \in \text{GLM}_{J,K}(h)$, where $(h_k : 1 \leq k \leq K)$ 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 [66] and primate primary motor cortex [67]. In particular, $3 \leq J \leq 20$, $0 \leq K \leq 20$, $-10 \leq \alpha_i, \beta_j \leq 10$. 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 [20].

The planned statistically causal influence structure, or the “functional topology,” is shown in Figure 6.1. 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. The β_i s 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

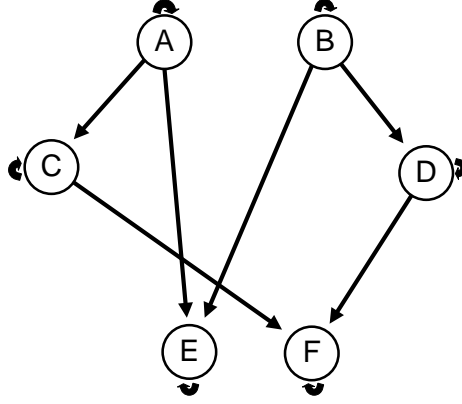


Figure 6.1: Diagram of planned causal influence structure for the simulated data set. Note that an arrow from neuron M to neuron N (possibly with M = N) means that M was designed to be causally dependent on N's firing.

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:

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

where $dx_{i-(k_1-1)}^1$ 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 $\alpha_0 = 1.8$, $J = 3$, $K_C = 5$, $K_D = 7$,

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

A sample of the time series for neurons C, F, and D respectively is shown in

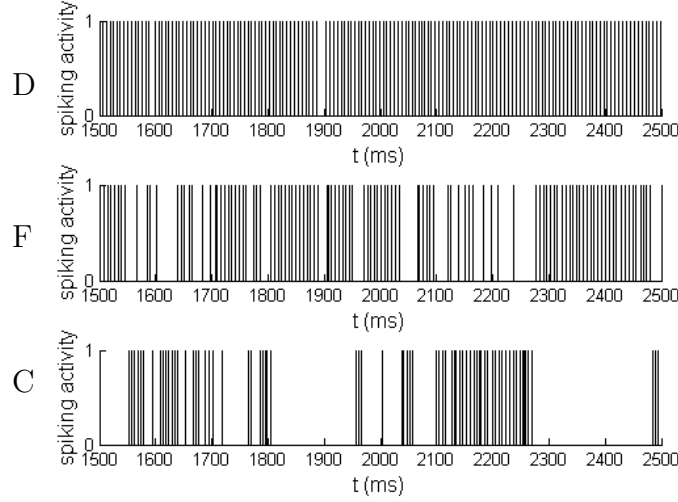


Figure 6.2: A one second sample of the spike trains generated for neuron D, neuron F, and neuron C. Neuron D was excitatory, whereas neuron C was inhibitory, in causally influencing neuron F.

Figure 6.2.

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, $\hat{\mathcal{I}}(\mathbf{X} \rightarrow \mathbf{Y})$, were computed. 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 $\hat{\mathcal{H}}(\mathbf{Y})$, which were found using the same procedure with $K = 0$. The same ordered pairs (\mathbf{X}, \mathbf{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 6.3 shows the averaged normalized estimates (5.17) 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

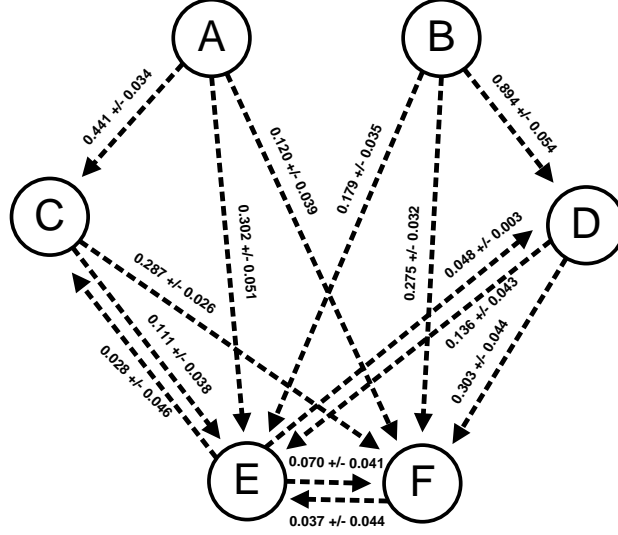


Figure 6.3: Diagram of the averaged non-zero, estimated normalized pairwise directed information rates (with averaged 95% confidence intervals) for the simulated data set, using 20 independently generated samples. The procedure selected the same structure for each sample. The procedure identified all the planned causal relationships (see figure 6.1). No *invalid* causal influences were detected (18 of the possible 30 arrows), nor were any planned causal influences *undetected*. The procedure also identified some indirect or “proxy” influences (i.e. the group B, D, and F) as well as some “cascading” influences (i.e. the group B, D, and E).

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 Figure 6.3 indicates that causal influence was detected ($\hat{K} > 0$), and the corresponding normalized estimate is adjacent to it. Absence of an arrow indicates that $\hat{K} = 0$, so no statistically causal influence was detected. The procedure identified all of the planned causal relationships (see Figure 6.1). Note that no *invalid* causal influences were detected, such as from $A \rightarrow B$ and $D \rightarrow A$. 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, including some “cascading” influences (see Figure 6.4) ($C \rightarrow E$, $E \rightarrow C$, $D \rightarrow E$, $E \rightarrow D$), some “proxy” (see Figure 6.5) influences ($A \rightarrow F$ and $B \rightarrow F$), and some higher order influences ($E \rightarrow F$, $F \rightarrow E$).

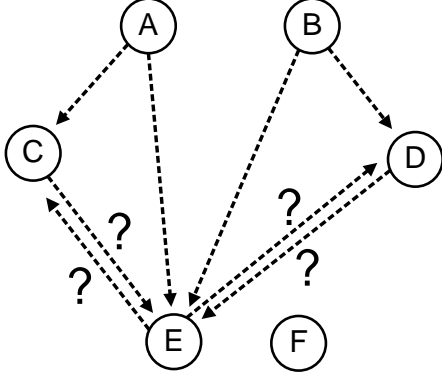


Figure 6.4: Diagram depicting a subgraph, in which cascading influences (denoted by arrows with adjacent “?”) were detected by the pairwise directed information estimates.

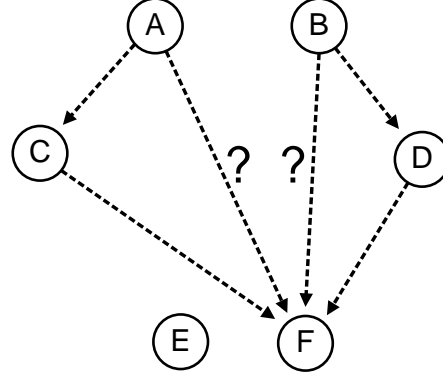


Figure 6.5: Diagram depicting a subgraph, in which proxy influences (denoted by arrows with adjacent “?”) were detected by the pairwise directed information estimates.

After the pairwise estimates were computed, causally conditioned directed information rates were computed and the spurious influences were removed (see Figure 6.6). 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. $\hat{\mathcal{I}}(\mathbf{A} \rightarrow \mathbf{C}||\mathbf{E})$ and $\hat{\mathcal{I}}(\mathbf{E} \rightarrow \mathbf{C}||\mathbf{A})$ were computed by first computing $\hat{\mathcal{H}}(\mathbf{C}||\mathbf{E}, \mathbf{A})$ and then comparing with $\hat{\mathcal{H}}(\mathbf{C}||\mathbf{E})$ and $\hat{\mathcal{H}}(\mathbf{C}||\mathbf{A})$ respectively. For all samples, it was estimated that $\hat{\mathcal{I}}(\mathbf{A} \rightarrow \mathbf{C}||\mathbf{E}) > 0$ and $\hat{\mathcal{I}}(\mathbf{E} \rightarrow \mathbf{C}||\mathbf{A}) = 0$, so $\mathbf{A} \rightarrow \mathbf{C}$ was kept and $\mathbf{E} \rightarrow \mathbf{C}$ was rejected. The same procedure was performed for neuron D with influences B and E, and it was found that $\hat{\mathcal{I}}(\mathbf{B} \rightarrow \mathbf{D}||\mathbf{E}) > 0$ and $\hat{\mathcal{I}}(\mathbf{E} \rightarrow \mathbf{D}||\mathbf{B}) = 0$, so $\mathbf{B} \rightarrow \mathbf{D}$ was kept and $\mathbf{E} \rightarrow \mathbf{D}$ was rejected.

Neurons E and F both had 5 influences, but those influences were not all connected. For example, the subsets $\{\mathbf{A}, \mathbf{C}\}$ and $\{\mathbf{B}, \mathbf{D}\}$ 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. $\hat{\mathcal{I}}(\mathbf{C} \rightarrow \mathbf{E}||\mathbf{A})$ was found to be 0 as was $\hat{\mathcal{I}}(\mathbf{D} \rightarrow \mathbf{E}||\mathbf{B})$, for all of the samples, so $\mathbf{C} \rightarrow \mathbf{E}$ and $\mathbf{D} \rightarrow \mathbf{E}$ were rejected.

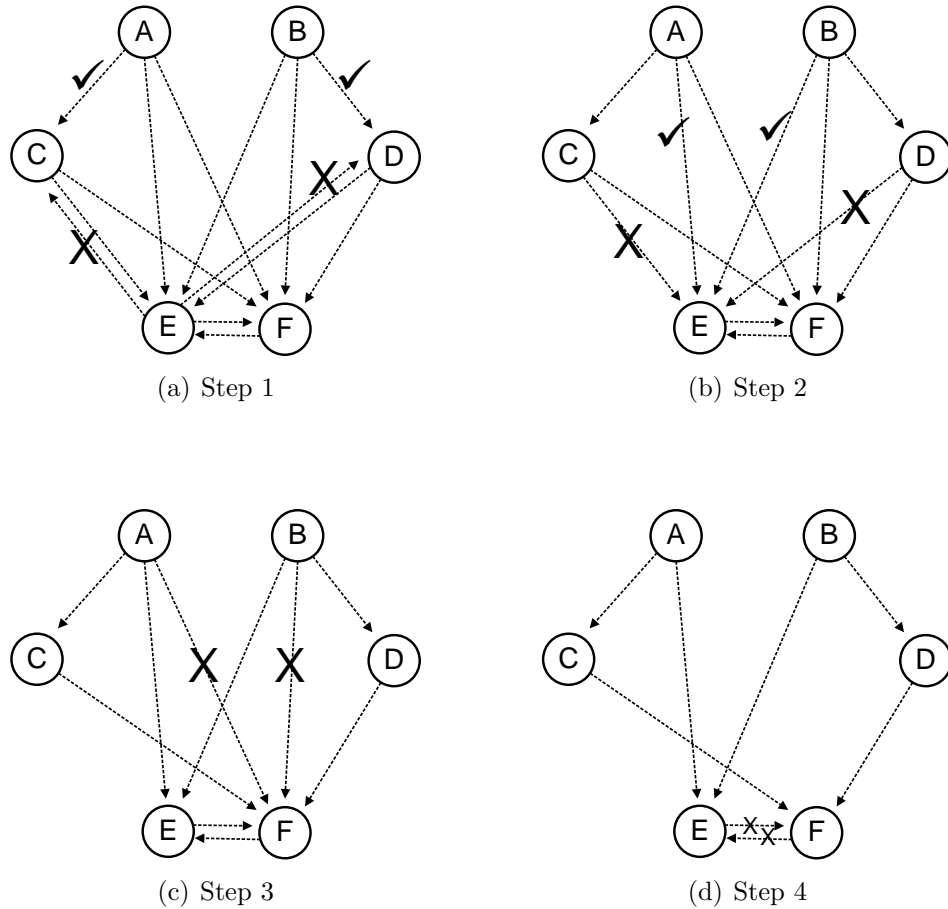


Figure 6.6: Steps of the algorithm to identify which of the detected (pairwise) causal influences are *direct* causal influences. A checkmark is placed next to influences that were tested and kept at that stage in the algorithm. An X is placed over the influences which were determined to not be direct, 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.

Since A and B did not have any detected influences between them, $A \rightarrow E$ and $B \rightarrow E$ 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, $\hat{\mathcal{I}}(A \rightarrow F | C, D)$ and $\hat{\mathcal{I}}(B \rightarrow F | C, D)$ were both computed and found to be 0 for all the samples, and consequently $A \rightarrow F$ and $B \rightarrow F$ were rejected.

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, $\hat{\mathcal{I}}(F \rightarrow E | A, B)$ was computed. For 15 of the 20 samples, it was 0, and thus $F \rightarrow E$ was rejected, with $A \rightarrow E$ and $B \rightarrow E$ kept. E's influences were now unambiguous. For the 5 samples where the estimated rate was greater than 0, $\hat{\mathcal{I}}(A \rightarrow E | F, B)$ and $\hat{\mathcal{I}}(B \rightarrow E | A, F)$ were both computed and found to be nonzero, so $F \rightarrow E$, $A \rightarrow E$, and $B \rightarrow E$ were all kept, and E's influences were now unambiguous. A similar procedure was done for F, and for 12 of the 20 samples, $E \rightarrow F$ was rejected, leaving $C \rightarrow F$ and $D \rightarrow F$; for the rest $E \rightarrow F$ was also kept. Two of the samples kept both $E \rightarrow F$ and $F \rightarrow E$. All of the influences for each of the neurons were thus resolved. The remaining influences were taken to be the direct, causal influences between the neurons (see Figure 6.7).

Figure 6.7 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 Figure 6.1) 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 planned influences and no spurious ones. Only two spurious influences, E to F and F to E, were detected among 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 $E \rightarrow F$ for only 8 samples, and $F \rightarrow E$ for only 5 samples (two of these had both). The pattern of these two arrows (short-dashed) differs from the others to depict this. Enforcing the criterion that only reliable estimated rates would be accepted would re-

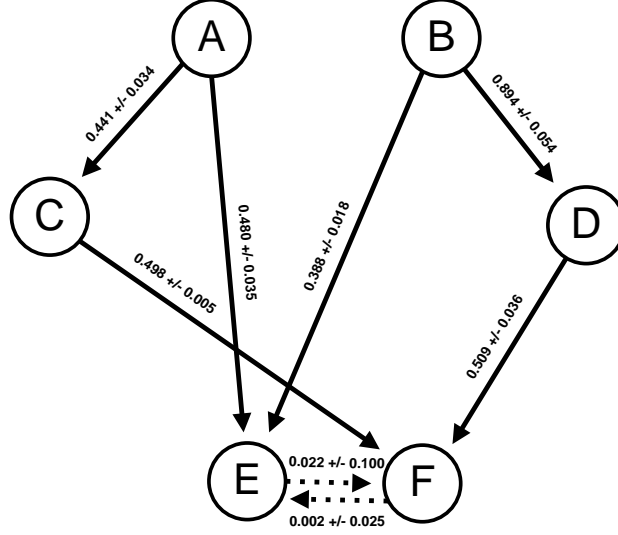


Figure 6.7: Diagram of the averaged non-zero normalized causally conditioned estimated directed information rates for the simulated data set (with averaged 95% confidence intervals). For each of the samples, all of the planned direct, causal influences (see Figure 6.1) were detected. These are depicted with solid arrows. Only two spurious influences, E to F and F to E, were detected among any of the samples, and their estimated rates were small and found to be unreliable (rates much smaller than confidence interval). The pattern of these two arrows (short-dashed) differ from the others to depict this.

sult in only the planned, direct causal influences being accepted. The values in the graph are: $\hat{\mathcal{I}}(\mathbf{A} \rightarrow \mathbf{C})$, $\hat{\mathcal{I}}(\mathbf{B} \rightarrow \mathbf{D})$, $\hat{\mathcal{I}}(\mathbf{A} \rightarrow \mathbf{E}||\mathbf{B})$, $\hat{\mathcal{I}}(\mathbf{B} \rightarrow \mathbf{E}||\mathbf{A})$, $\hat{\mathcal{I}}(\mathbf{C} \rightarrow \mathbf{F}||\mathbf{D})$, $\hat{\mathcal{I}}(\mathbf{D} \rightarrow \mathbf{F}||\mathbf{C})$, $\hat{\mathcal{I}}(\mathbf{F} \rightarrow \mathbf{E}||\mathbf{A}, \mathbf{B})$, and $\hat{\mathcal{I}}(\mathbf{E} \rightarrow \mathbf{F}||\mathbf{C}, \mathbf{D})$.

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 , $\{\alpha_i\}_{i=1}^J$, and $\{\beta_i\}_{i=1}^K$ for each neuron) is difficult and was not done for this 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 equation (5.12) 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 consequently on 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 affecting the conditional intensity more and thus Y's spiking rate. For example, $\hat{\mathcal{I}}(A \rightarrow C) \approx 0.4$. C was designed to depend on A with parameters $J = 6$, $K = 5$,

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

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

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

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.

6.2 Experimental Data

Data Source

In addition to simulated data trials, experimental data from Wu and Hatsopoulos [67] 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 cor-

respondingly moved a cursor, to hit the current target. Each target was presented for a maximum of 2 seconds, 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 microelectrode 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) [67]. 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 [67].

6.2.1 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. Seven second samples of the data selected for neurons 3 and 1 are shown in Figure 6.8. 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 seconds (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 seconds for selected neurons was about 8000 spikes, or approximately one

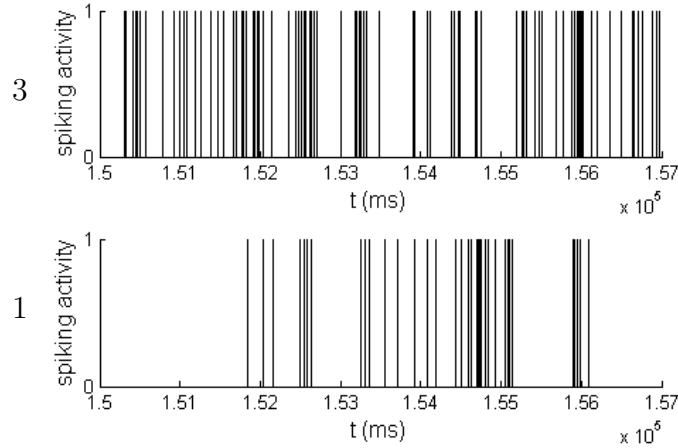


Figure 6.8: 7 second snapshot of spiking activity of neurons 1 and 3 in the data set from [67] used for analysis. The procedure found that neuron 3 causally influences neuron 1, in an excitatory manner.

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 6.9 is a graph of the pairwise results. Each box with a label of $i \in \{1, 2, \dots, 37\}$ 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 ($\hat{K} > 0$). Absence of an arrow between an ordered pair (X, Y) depicts that the estimation procedure detected that there was no statistically causal influence ($\hat{K} = 0$). 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

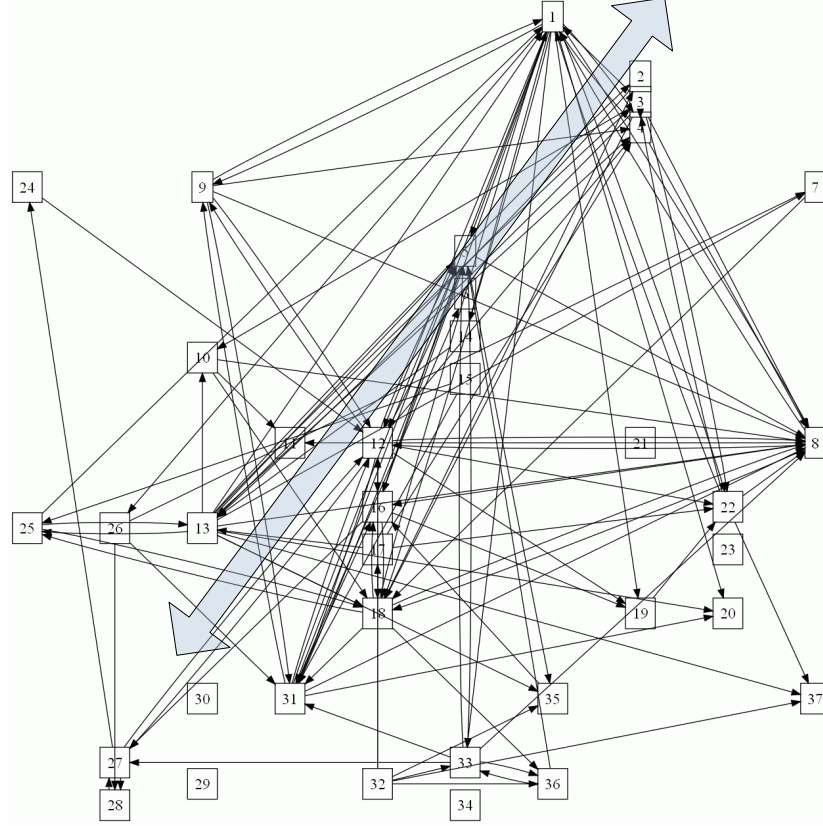


Figure 6.9: Diagram of statistically estimated causal relationships for the 37 neurons used from the subset of electrode recordings in the arm area of a monkey’s primary motor cortex (MI) from [67]. Each box with a number indicates a different neuron. The relative positions of the neurons in the diagram 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, $\hat{K} > 0$). Absence of an arrow from X to Y depicts that the procedure detected no statistically causal relationship from X to Y ($\hat{K} = 0$). 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 [68].

many neighboring neurons exciting or inhibiting them (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 $\{1, 4, 9\}$, $\{3, 10, 13\}$, $\{5, 13, 35\}$, $\{8, 10, 11\}$, $\{12, 16, 27\}$, $\{13, 18, 25\}$, 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 pairwise influences was removed from each. See Figures 6.10 through 6.15. For the other subsets, all of the causally conditioned directed information estimates were greater than 0, and so they were kept.

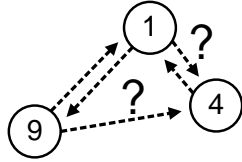


Figure 6.10: Diagram depicting the induced subgraph of neurons 1, 9, and 4. Both 1 and 9 have pairwise influences into 4, one of which might be due to an indirect influence. A question mark is drawn adjacent to the arrows in question.

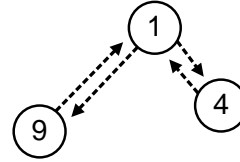


Figure 6.11: The resulting subgraph after computing causally conditioned directed information estimates. $\hat{\mathcal{I}}(1 \rightarrow 4|9) > 0$ and $\hat{\mathcal{I}}(9 \rightarrow 4|1) = 0$, so $9 \rightarrow 4$ was removed, and $1 \rightarrow 4$ was kept.

A strong structure can be seen in the graph (Figure 6.9). 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 among 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

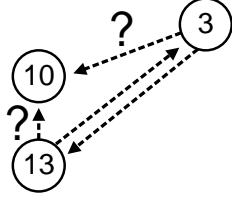


Figure 6.12: Diagram depicting the induced subgraph of neurons 3, 10, and 13. Both 3 and 13 have pairwise influences into 10, one of which might be due to an indirect influence. A question mark is drawn adjacent to the arrows in question.

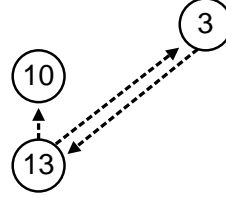


Figure 6.13: The resulting subgraph after computing causally conditioned directed information estimates. $\hat{\mathcal{I}}(3 \rightarrow 10|13) = 0$ and $\hat{\mathcal{I}}(13 \rightarrow 10|3) > 0$, so $3 \rightarrow 10$ was removed, and $13 \rightarrow 10$ was kept.

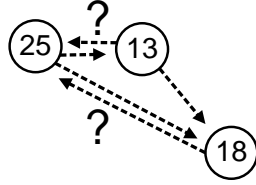


Figure 6.14: Diagram depicting the induced subgraph of neurons 13, 18, and 25. Both 13 and 18 have pairwise influences into 25, one of which might be due to an indirect influence. A question mark is drawn adjacent to the arrows in question.

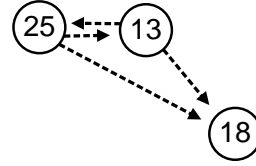


Figure 6.15: The resulting subgraph after computing causally conditioned directed information estimates. $\hat{\mathcal{I}}(13 \rightarrow 25|18) > 0$ and $\hat{\mathcal{I}}(18 \rightarrow 25|13) = 0$, so $18 \rightarrow 25$ was removed, and $13 \rightarrow 25$ was kept.

that there might be *some* physical pathway between the two neurons, such that the spiking 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 [67]. Similar to the analysis of the simulated data set, even among the recorded neurons, it is unclear what influences are “direct,” and which might be accounted for by “proxy” or “cascading” effects (see Figure 4.3).

In addition to the number of detected influence relationships between the neurons, there is also a visibly dominant orientation of the connections (see Figure 6.9). While the 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 [68]. 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 facilitate information transfer between intra- and inter-cortical regions during movement preparation and execution [68]. Other studies have found that in the turtle visual cortex, these waves were present during the introduction of visual stimuli [69] and have been shown to encode information related to target position [70]. Similar wave-like spatiotemporal activity has been observed in other areas of the nervous systems of a variety of animals and are thought to play an important role in the communication between different areas of the brain [71]. Physically, beta oscillations are believed to correspond to the summed effects of multiple, synchronous postsynaptic potentials from neurons close to the recording electrode [68]. Little is known about the precise mechanisms through which the propagation of these waves occur [68]. The proposed estimation procedure could provide insight into these mechanisms. The procedure could potentially identify both the local propagation pathways (by detecting structure as in Figure 6.9) as well as the specific relationship dynamics between the recorded neurons (by identifying the coefficients of the conditional intensity function, the α_i s and β_j s).

CHAPTER 7

APPROXIMATING DISCRETE PROBABILITY DISTRIBUTIONS WITH CAUSAL DEPENDENCE TREES

7.1 Introduction

Numerous statistical learning, inference, prediction, and communication problems require storing the joint distribution for a large number of random variables. As the number of variables increases linearly, the number of elements in the joint distribution increases multiplicatively by the cardinality of the alphabets. Thus, for storage and analysis purposes, it is often desirable to approximate to the full joint distribution.

Bayesian networks is an area of research within which methods have been developed to approximate or simplify a full joint distribution with an approximating distribution [26]. In general, there are various choices of the structure of the approximating distribution. Chow and Liu developed a method of approximating a full joint distribution with a dependence tree distribution [16]. The joint distribution is represented as a product of marginals, where each random variable is conditioned on at most one other random variable and no loops are allowed in the corresponding graph. For Bayesian networks, graphical models are often used to represent distributions, with variables represented as nodes and undirected edges between pairs of correlated variables. The edge set is such that a variable is independent of all the nodes it is not connected to, conditioned on the nodes it is connected to. The dependence tree distributions have graphical representations as trees (a graph without any loops). Chow and Liu’s procedure efficiently computes the “best” approximating tree for a given joint distribution, where “best” is defined in terms of the Kullback-Liebler (KL) divergence between the original joint distribution and the approximating tree distribution [16]. They also showed that finding the “best” fitting tree was equivalent to maximizing a sum of mutual informations [16].

This work will consider the specific setting where there are random processes with a common index set (interpreted as timings). Applying the Chow and Liu procedure directly would result in intermixing the variables between the different processes. For some research problems, it is desired to keep the processes separate. Also, the Chow and Liu procedure would ignore the index set, which would result in a loss of timing information and thus causal structure. We will present a procedure similar to Chow and Liu's, but for this setting. Analogous to the result of Chow and Liu, finding the "best" fitting causal dependence tree (between the processes) is equivalent to maximizing a sum of directed informations. There will also be an algorithm presented, similar to that in Chow and Liu's work, to efficiently identify the optimal causal dependence tree.

7.2 Background: Dependence Tree Approximations

Given a set of n discrete random variables $X^n = \{X_1, X_2, \dots, X_n\}$, possibly over different alphabets $\{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_n\}$, the chain rule is

$$\begin{aligned} P_{X^n}(\cdot) &= P_{X_n|X^{n-1}}(\cdot) P_{X_{n-1}|X^{n-2}}(\cdot) \cdots P_{X_1}(\cdot) \\ &= P_{X_1}(\cdot) \prod_{i=2}^n P_{X_i|X^{i-1}}(\cdot) \end{aligned}$$

For the chain rule, the order of the random variables does not matter, so for any permutation $\pi(\cdot)$ on $\{1, \dots, n\}$,

$$P_{X^n}(\cdot) = \prod_{i=1}^n P_{X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \dots, X_{\pi(1)}}(\cdot).$$

Chow and Liu developed an algorithm to approximate a known, full joint distribution by a product of second order distributions [16]. For their procedure, the chain rule is applied to the joint distribution, and all the terms of the form $P_{X_{\pi(i)}|X_{\pi(i-1)}, X_{\pi(i-2)}, \dots, X_{\pi(1)}}(\cdot)$ are approximated (possibly exactly) by $P_{X_{\pi(i)}|X_{\pi(j(i))}}(\cdot)$ where $j(i) \in \{1, \dots, i-1\}$, such that the conditioning is on at most one variable. This product of second order distributions serves

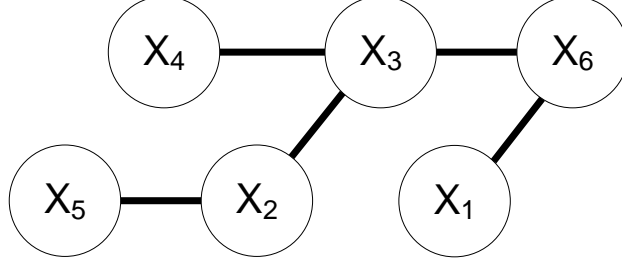


Figure 7.1: Diagram of a joint distribution with an underlying dependence tree structure. In this example, $P_{X_1^6}(\cdot) = P_{X_6}(\cdot)P_{X_1|X_6}(\cdot)P_{X_3|X_6}(\cdot) \times P_{X_4|X_3}(\cdot)P_{X_2|X_3}(\cdot)P_{X_5|X_2}(\cdot)$. There is a similar factorization taking any other node (other than X_6 as root).

as an approximation of the full joint.

$$P_{X^n}(\cdot) \approx \prod_{i=1}^n P_{X_{\pi(i)}|X_{\pi(j(i))}}(\cdot).$$

This approximation has a tree dependence structure. This follows because application of the chain rule induces a dependence structure which has no loops (e.g., no terms of the form $P_{A|B}(a|b)P_{B|C}(b|c)P_{C|A}(c|a)$), and this is a reduction of that structure, so it does not introduce any loops. An example of a joint distribution with an underlying tree dependence structure is shown in Figure 7.1. In general, the approximation will not be exact. Denote each tree approximation of $P_{X^n}(x^n)$ by $\hat{P}_T(x^n)$. Each choice of $\pi(\cdot)$ and $j(\cdot)$ over $\{1, \dots, n\}$ completely specifies a tree structure T . Thus, the tree approximation of the joint using the particular tree T is

$$\hat{P}_T(x_1^n) \triangleq \prod_{i=1}^n P_{X_{\pi(i)}|X_{\pi(j(i))}}(x_{\pi(i)}|x_{\pi(j(i))}). \quad (7.1)$$

Chow and Liu's method obtains the “best” such model T , where the “goodness” is defined in terms of Kullback-Liebler (KL) distance between the original distribution and the approximating distribution [16]. Their method involves minimizing the KL distance over all trees T (permutations π and

functions $j(\cdot)$) for a given joint distribution P :

$$\arg \min_T D(P_{X^n} || \hat{P}_T) = \arg \min_T E_P \left[\log \frac{P_{X^n}(X^n)}{\hat{P}_T(X^n)} \right] \quad (7.2)$$

$$= \arg \max_T \sum_{i=1}^n I(X_{\pi(i)}; X_{\pi(j(i))}). \quad (7.3)$$

Equation (7.2) follows from definition of KL distance; (7.3) follows from [16]. The optimization objective is to maximize a sum of mutual informations.

They propose an efficient algorithm to identify this approximating tree [16]. Calculate the mutual information between each pair of random variables. Now consider a complete, undirected graph, in which each of the random variables is represented as a node. The mutual information values can be thought of as weights for the corresponding edges. Finding the dependence tree distribution that maximizes the sum (7.3) is equivalent to the graph problem of finding a tree of maximal weight [16]. Kruskal's minimum spanning tree algorithm [72] can be used to reduce the complete graph to a tree with the largest sum of mutual informations [16]. If mutual information values are not unique, there could be multiple solutions.

7.3 Main Result: Causal Dependence Tree Approximations

In situations where there are multiple random processes, the Chow and Liu method can be used. However, it will consider all possible arrangements of all the variables, "mixing" the processes and timings to find the best approximation. An alternative approach, which would maintain causality and keep the processes separate, is to find an approximation to the full joint probability by identifying causal dependencies between the processes themselves. In particular, consider finding a causal dependence tree structure, where instead of conditioning on a variable using one auxilliary variable as in Chow and Liu, the conditioning is on a *process* using one auxilliary process.

Consider the joint distribution $P_{\{\mathbf{A}_h\}_{h=1}^N}$ of N random processes $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N$, each of length n , where $\mathbf{A}_h = \{A_{h,i}\}_{i=1}^n$ and $A_{h,i}$ is a discrete random variable over the finite alphabet \mathcal{A}_h , with $1 \leq h \leq N$ (so if $N = 4$,

we might denote the processes as $\{W^n, X^n, Y^n, Z^n\}$. Denote realizations of the random variable $A_{h,i}$ by $a_{h,i}$. By causal conditioning,

$$P_{\mathbf{A}_{h_1}||\mathbf{A}_{h_2}}(a_{h_1}^n||a_{h_2}^n) \triangleq \prod_{i=1}^n P_{A_{h_1,i}||A_{h_1}^{i-1}, A_{h_2}^i}(a_{h_1,i}|a_{h_1}^{i-1}, a_{h_2}^i) \quad (7.4)$$

for any two random processes \mathbf{A}_{h_1} and \mathbf{A}_{h_2} . This chain rule for causal conditioning is similar to the normal chain rule, except that “future” outcomes in the other processes are not conditioned upon.

The joint distribution of the processes can be approximated in a similar manner as before, except that instead of permuting the index of the set of random variables, consider permutations on the index set of the processes themselves. For a given joint probability distribution $P_{\{\mathbf{A}_h\}_{h=1}^N}(\{a_h^n\}_{h=1}^N)$ and tree T , denote the corresponding approximating causal dependence tree induced probability to be

$$\widehat{P}_T(\{a_h^n\}_{h=1}^N) = \prod_{h=1}^N P_{\mathbf{A}_{\pi(h)}||\mathbf{A}_{\pi(j(h))}}(a_{\pi(h)}^n||a_{\pi(j(h))}^n)$$

As before, the goal is to obtain the “best” such model T , where the “goodness” is defined in terms of KL distance between the original distribution and the approximating distribution. The objective is to optimize, for a given

joint distribution P :

$$\arg \min_T D(P || \hat{P}_T) = \arg \min_T E_P \left[\log \frac{P_{\{\mathbf{A}_h\}_{h=1}^N}(\{\mathbf{A}_h\}_{h=1}^N)}{\hat{P}_T(\{\mathbf{A}_h\}_{h=1}^N)} \right] \quad (7.5)$$

$$= \arg \min_T E_P \left[\log P_{\{\mathbf{A}_h\}_{h=1}^N}(\{\mathbf{A}_h\}_{h=1}^N) \right] + E_P \left[-\log \prod_{h=1}^N P_{\mathbf{A}_{\pi(h)} || \mathbf{A}_{\pi(j(h))}} \right] \quad (7.6)$$

$$= \arg \min_T -H(\{\mathbf{A}_h\}_{h=1}^N) + \sum_{h=1}^N E_P \left[-\log P_{\mathbf{A}_{\pi(h)} || \mathbf{A}_{\pi(j(h))}} \right] \quad (7.7)$$

$$= \arg \min_T \sum_{h=1}^N H(\mathbf{A}_{\pi(h)} || \mathbf{A}_{\pi(j(h))}) \quad (7.8)$$

$$= \arg \min_T \sum_{h=1}^N H(\mathbf{A}_{\pi(h)} || \mathbf{A}_{\pi(j(h))}) + (H(\mathbf{A}_{\pi(h)}) - H(\mathbf{A}_{\pi(h)})) \quad (7.9)$$

$$= \arg \min_T \sum_{h=1}^N -I(\mathbf{A}_{\pi(j(h))} \rightarrow \mathbf{A}_{\pi(h)}) + \sum_{h=1}^N H(\{\mathbf{A}_h\}) \quad (7.10)$$

$$= \arg \min_T -\sum_{h=1}^N I(\mathbf{A}_{\pi(j(h))} \rightarrow \mathbf{A}_{\pi(h)}) \quad (7.11)$$

$$= \arg \max_T \sum_{h=1}^N I(\mathbf{A}_{\pi(j(h))} \rightarrow \mathbf{A}_{\pi(h)}). \quad (7.12)$$

Equation (7.5) follows from definition of KL distance, (7.6) breaks up log and uses def of \hat{P}_T , (7.7) uses definition of entropy and properties of log, (7.8) joint entropy of processes is not dependent on tree T , and uses definition of causal entropy, (7.9) adds 0, (7.10) rearranges, uses identity of directed information, and re-orders sum over entropy of each process, and (7.11) the sum is independent of tree T . Our optimization objective is to maximize a sum of directed informations.

In Chow and Liu's work, Kruskal's minimum spanning tree algorithm performs the analogous optimization procedure efficiently, after having computed the mutual information between each pair [16]. A similar procedure can be done in this setting. First, compute the directed information between

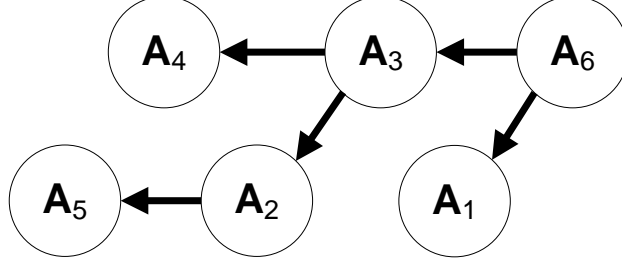


Figure 7.2: Diagram of a joint distribution with an underlying causal dependence tree structure. In this example, $P_{\{\mathbf{A}_i\}_{i=1}^6}(\cdot) = P_{\mathbf{A}_6}(\cdot) \times P_{\mathbf{A}_1|\mathbf{A}_6}(\cdot) P_{\mathbf{A}_3|\mathbf{A}_6}(\cdot) P_{\mathbf{A}_4|\mathbf{A}_3}(\cdot) P_{\mathbf{A}_2|\mathbf{A}_3}(\cdot) P_{\mathbf{A}_5|\mathbf{A}_2}(\cdot)$.

each ordered pair of processes. This can be represented as a graph, where each of the nodes represents a process. This graph will have a directed edge from each node to every other node (thus is a complete, directed graph), and the value of edge from node \mathbf{X} to node \mathbf{Y} will be $I(\mathbf{X} \rightarrow \mathbf{Y})$. The analog to a spanning tree for undirected graphs is an “arborescence” for directed graphs [73]. An arborescence is a connected, directed graph, where each node has at most one incoming edge, and there is one node with no incoming edges, called the “root.” An example of a joint distribution with an underlying causal dependence tree, which is depicted as an arborescence, is in Figure 7.2. There are several efficient algorithms which can be used to find the maximum weight (sum of directed informations) arborescence of a directed graph [74], such as Chu and Liu [75] (which was independently discovered by Edmonds [76] and Bock [77]) and a distributed algorithm by Humblet [78]. Note that in some implementations, a root is required a priori. For those, the implementation would need to be applied for each node in the graph as a root, and then the arborescence which has maximal weight among all of those would be selected.

CHAPTER 8

CONCLUSION

This work examines an information theoretic quantity known as directed information, which is shown to be a general quantity (applicable to arbitrary probability distributions). It is interpreted in terms of prediction, communication with feedback, source coding with feed forward, control over noisy channels, and other settings. It is also shown to be consistent with Granger’s philosophical definition. Formal definitions for causal and direct, causal influences were presented. Two applications of directed information were then investigated.

In the first, a procedure to consistently estimate the directed information between neural spike trains was developed. The procedure was tested on simulated data and applied to experimental data, both with promising results. This technique could become a practical, provably-good, and philosophically well-grounded means of identifying the statistically causal, complex relationships between neurons in large data sets of simultaneous, multiple electrode recordings.

In the second, a procedure, similar to Chow and Liu’s, was developed for finding the “best” approximation (in terms of KL divergence) of a full, joint distribution over a set of random processes, using a causal dependence tree distribution. Chow and Liu’s method had been shown to be equivalent to maximizing a sum of mutual informations, and the procedure presented here is shown to be equivalent to maximizing a sum of directed informations. An algorithm was presented for efficiently finding the optimal causal tree, similar to that in Chow and Liu’s work.

REFERENCES

- [1] C. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
- [2] J. Massey, “Causality, feedback and directed information,” in *Proc. Int. Symp. Information Theory Application (ISITA-90)*, 1990, pp. 303–305.
- [3] H. Marko, “The bidirectional communication theory—a generalization of information theory,” *Communications, IEEE Transactions on*, vol. 21, no. 12, pp. 1345–1351, Dec 1973.
- [4] J. Rissanen and M. Wax, “Measures of mutual and causal dependence between two time series (Corresp.),” *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 598–601, 1987.
- [5] G. Kramer, “Directed information for channels with feedback,” Ph.D. dissertation, University of Manitoba, Canada, 1998.
- [6] S. Tatikonda and S. Mitter, “The capacity of channels with feedback,” *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 323–349, 2009.
- [7] H. Permuter, T. Weissman, and A. Goldsmith, “Finite state channels with time-invariant deterministic feedback,” *IEEE Transactions on Information Theory*, vol. 55, no. 2, pp. 644–662, 2009.
- [8] J. Massey and P. Massey, “Conservation of mutual and directed information,” in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, 2005, pp. 157–158.
- [9] H. Permuter, Y. Kim, and T. Weissman, “On directed information and gambling,” in *IEEE International Symposium on Information Theory, 2008. ISIT 2008*, 2008, pp. 1403–1407.
- [10] H. Permuter, Y. Kim, and T. Weissman, “Interpretations of directed information in portfolio theory, data compression, and hypothesis testing,” *IEEE Transactions on Information Theory*, submitted for publication.

- [11] N. Elia, “When Bode meets Shannon: control-oriented feedback communication schemes,” *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1477 – 1488, sept. 2004.
- [12] N. Martins and M. Dahleh, “Feedback control in the presence of noisy channels: “Bode-like” fundamental limitations of performance,” *Automatic Control, IEEE Transactions on*, vol. 53, no. 7, pp. 1604 –1615, aug. 2008.
- [13] S. Tatikonda, “Control under communication constraints,” Ph.D. dissertation, Massachusetts Institute of Technology, 2000.
- [14] S. Gorantla and T. Coleman, “On reversible Markov chains and maximization of directed information,” *IEEE International Symposium on Information Theory (ISIT)*, submitted.
- [15] R. Venkataramanan and S. Pradhan, “Source coding with feed-forward: rate-distortion theorems and error exponents for a general source,” *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2154–2179, 2007.
- [16] C. Chow and C. Liu, “Approximating discrete probability distributions with dependence trees,” *IEEE transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.
- [17] T. Cover and J. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley-Interscience, 2006.
- [18] E. Brown, R. Barbieri, U. Eden, and L. Frank, “Likelihood methods for neural spike train data analysis,” in *Computational Neuroscience: A Comprehensive Approach*, J. Feng, Ed. New York, NY: Chapman and Hall/CRC, 2003, pp. 253-286.
- [19] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes*. New York: Springer-Verlag, 1988.
- [20] E. Brown, R. Barbieri, V. Ventura, R. Kass, and L. Frank, “The time-rescaling theorem and its application to neural spike train data analysis,” *Neural Computation*, vol. 14, no. 2, pp. 325–346, 2002.
- [21] P. Bremaud, *Point Processes and Queues: Martingale Dynamics*. New York: Springer-Verlag, 1981.
- [22] R. Sundaresan and S. Verdú, “Capacity of queues via point-process channels,” *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2697–2709, June 2006.
- [23] C. Shalizi, “Causal architecture, complexity and self-organization in time series and cellular automata,” Ph.D. dissertation, University of Wisconsin-Madison, 2001.

- [24] M. Al-Khassaweneh and S. Aviyente, “The relationship between two directed information measures,” *Signal Processing Letters, IEEE*, vol. 15, pp. 801–804, 2008.
- [25] Y. Kim, H. Pennuter, and T. Weissman, “Directed information and causal estimation in continuous time,” in *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory-Volume 2*. Institute of Electrical and Electronics Engineers Inc., The, 2009, pp. 819–823.
- [26] J. Pearl, *Causality: Models, Reasoning and Inference*. New York, NY: Cambridge University Press, 2009.
- [27] W. Hesse, E. Möller, M. Arnold, and B. Schack, “The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies,” *Journal of Neuroscience Methods*, vol. 124, no. 1, pp. 27–44, 2003.
- [28] H. Akaike, “An information criterion (AIC),” *Math Sci*, vol. 14, no. 153, pp. 5–9, 1976.
- [29] A. Barron and T. Cover, “Minimum complexity density estimation,” *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1034–1054, 1991.
- [30] M. Kaminski and K. Blinowska, “A new method of the description of the information flow in the brain structures,” *Biological Cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.
- [31] M. Kamiński, M. Ding, W. Truccolo, and S. Bressler, “Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance,” *Biological Cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.
- [32] L. Uddin, A. Clare Kelly, B. Biswal, F. Xavier Castellanos, and M. Milham, “Functional connectivity of default mode network components: correlation, anticorrelation, and causality,” *Hum Brain Mapp*, vol. 30, no. 2, pp. 625–37, 2009.
- [33] R. Goebel, A. Roebroeck, D. Kim, and E. Formisano, “Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping,” *Magnetic Resonance Imaging*, vol. 21, no. 10, pp. 1251–1261, 2003.
- [34] A. Roebroeck, E. Formisano, and R. Goebel, “Mapping directed influence over the brain using Granger causality and fMRI,” *Neuroimage*, vol. 25, no. 1, pp. 230–242, 2005.

- [35] B. Rogers, V. Morgan, A. Newton, and J. Gore, “Assessing functional connectivity in the human brain by fMRI,” *Magnetic Resonance Imaging*, vol. 25, no. 10, pp. 1347–1357, 2007.
- [36] M. Dhamala, G. Rangarajan, and M. Ding, “Analyzing information flow in brain networks with nonparametric Granger causality,” *NeuroImage*, vol. 41, no. 2, pp. 354–362, 2008.
- [37] B. Abler, A. Roebroek, R. Goebel, A. Höse, C. Schönfeldt-Lecuona, G. Hole, and H. Walter, “Investigating directed influences between activated brain areas in a motor-response task using fMRI,” *Magnetic Resonance Imaging*, vol. 24, no. 2, pp. 181–185, 2006.
- [38] A. Korzeniewska, M. Mańczak, M. Kamiński, K. Blinowska, and S. Kasiński, “Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method,” *Journal of neuroscience methods*, vol. 125, no. 1-2, pp. 195–207, 2003.
- [39] X. Wang, Y. Chen, S. Bressler, and M. Ding, “Granger causality between multiple interdependent neurobiological time series: Blockwise versus pairwise methods,” *International Journal of Neural Systems*, vol. 17, no. 2, p. 71, 2007.
- [40] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. Bressler, “Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 26, p. 9849, 2004.
- [41] T. Schreiber, “Measuring information transfer,” *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, 2000.
- [42] M. Chávez, J. Martinerie, and M. Le Van Quyen, “Statistical assessment of nonlinear causality: Application to epileptic EEG signals,” *Journal of Neuroscience Methods*, vol. 124, no. 2, pp. 113–128, 2003.
- [43] B. Gourevitch and J. Eggermont, “Evaluating information transfer between auditory cortical neurons,” *Journal of Neurophysiology*, vol. 97, no. 3, p. 2533, 2007.
- [44] A. Kraskov, “Synchronization and interdependence measures and their application to the electroencephalogram of epilepsy patients and clustering of data,” Ph.D. dissertation, University of Wuppertal, 2004.
- [45] E. Pereda, R. Quiroga, and J. Bhattacharya, “Nonlinear multivariate analysis of neurophysiological signals,” *Progress in Neurobiology*, vol. 77, no. 1-2, pp. 1–37, 2005.

- [46] K. Friston, L. Harrison, and W. Penny, “Dynamic causal modelling,” *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.
- [47] K. Stephan, L. Kasper, L. Harrison, J. Daunizeau, H. den Ouden, M. Breakspear, and K. Friston, “Nonlinear dynamic causal models for fMRI,” *NeuroImage*, vol. 42, no. 2, pp. 649–662, 2008.
- [48] C. Grefkes, S. Eickhoff, D. Nowak, M. Dafotakis, and G. Fink, “Dynamic intra-and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM,” *Neuroimage*, vol. 41, no. 4, pp. 1382–1394, 2008.
- [49] K. Hamandi, H. Powell, H. Laufs, M. Symms, G. Barker, G. Parker, L. Lemieux, and J. Duncan, “Combined EEG-fMRI and tractography to visualise propagation of epileptic activity,” *British Medical Journal*, vol. 79, no. 5, pp. 594–597, 2008.
- [50] B. Schuyler, J. Ollinger, T. Oakes, T. Johnstone, and R. Davidson, “Dynamic Causal Modeling applied to fMRI data shows high reliability,” *Neuroimage*, vol. 49, no. 1, pp. 603–611, 2009.
- [51] T. Bitan, J. Booth, J. Choy, D. Burman, D. Gitelman, and M. Mesulam, “Shifts of effective connectivity within a language network during rhyming and spelling,” *Journal of Neuroscience*, vol. 25, no. 22, p. 5397, 2005.
- [52] O. David, S. Kiebel, L. Harrison, J. Mattout, J. Kilner, and K. Friston, “Dynamic causal modeling of evoked responses in EEG and MEG,” *NeuroImage*, vol. 30, no. 4, pp. 1255–1272, 2006.
- [53] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY: Cambridge University Press, 2006.
- [54] J. Ziv and A. Lempel, “A universal algorithm for sequential data compression,” *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, 1977.
- [55] H. Cai, S. Kulkarni, and S. Verdú, “Universal entropy estimation via block sorting,” *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1551–1561, 2004.
- [56] H. Cai, S. Kulkarni, and S. Verdu, “An algorithm for universal lossless compression with side information,” *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4008–4016, 2006.
- [57] L. Zhao, H. Permuter, Y. Kim, and T. Weissman, “Universal Estimation of Directed Information,” in *Proceedings of the IEEE International Symposium on Information Theory*, to be published.

- [58] F. Perez-Cruz, “Estimation of information theoretic measures for continuous random variables,” in *Proceedings of Neural Information Processing Systems*, 2009, pp. 1257–1264.
- [59] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. New York, NY: Cambridge University Press, 2009.
- [60] W. Truccolo, U. Eden, M. Fellows, J. Donoghue, and E. Brown, “A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects,” *Journal of Neurophysiology*, vol. 93, no. 2, pp. 1074–1089, 2005.
- [61] G. Casella and R. Berger, *Statistical Inference*. Pacific Grove, CA: Duxbury, 2002.
- [62] P. Grünwald and J. Rissanen, *The Minimum Description Length Principle*. Cambridge, MA: The MIT Press, 2007.
- [63] L. Lastras, “An almost sure convergence proof of the sliding-window Lempel-Ziv algorithm,” in *IEEE International Symposium on Information Theory, Proceedings*, 2002, p. 121.
- [64] T. Vogels and L. Abbott, “Signal propagation and logic gating in networks of integrate-and-fire neurons,” *Journal of Neuroscience*, vol. 25, no. 46, p. 10786, 2005.
- [65] L. Paninski, M. Fellows, N. Hatsopoulos, and J. Donoghue, “Spatiotemporal tuning of motor cortical neurons for hand position and velocity,” *Journal of Neurophysiology*, vol. 91, no. 1, p. 515, 2004.
- [66] S. Iyengar and Q. Liao, “Modeling neural activity using the generalized inverse Gaussian distribution,” *Biological Cybernetics*, vol. 77, no. 4, pp. 289–295, 1997.
- [67] W. Wu and N. Hatsopoulos, “Evidence against a single coordinate system representation in the motor cortex,” *Experimental Brain Research*, vol. 175, no. 2, pp. 197–210, 2006.
- [68] D. Rubino, K. Robbins, and N. Hatsopoulos, “Propagating waves mediate information transfer in the motor cortex,” *Nature Neuroscience*, vol. 9, no. 12, pp. 1549–1557, 2006.
- [69] J. Prechtl, L. Cohen, B. Pesaran, P. Mitra, and D. Kleinfeld, “Visual stimuli induce waves of electrical activity in turtle cortex,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 14, p. 7621, 1997.

- [70] X. Du, B. Ghosh, and P. Ulinski, “Encoding and decoding target locations with waves in the turtle visual cortex,” *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 4, pp. 566–577, 2005.
- [71] G. Ermentrout and D. Kleinfeld, “Traveling Electrical Waves in Cortex Insights from Phase Dynamics and Speculation on a Computational Role,” *Neuron*, vol. 29, no. 1, pp. 33–44, 2001.
- [72] J. Kruskal Jr, “On the shortest spanning subtree of a graph and the traveling salesman problem,” *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.
- [73] J. Evans and E. Minieka, *Optimization Algorithms for Networks and Graphs*, 2nd ed. New York, NY: Dekker, 1992.
- [74] H. Gabow, Z. Galil, T. Spencer, and R. Tarjan, “Efficient algorithms for finding minimum spanning trees in undirected and directed graphs,” *Combinatorica*, vol. 6, no. 2, pp. 109–122, 1986.
- [75] Y. Chu and T. Liu, “On the shortest arborescence of a directed graph,” *Science Sinica*, vol. 14, no. 1396-1400, p. 270, 1965.
- [76] J. Edmonds, “Optimum branchings,” *J. Res. Natl. Bur. Stand., Sect. B*, vol. 71, pp. 233–240, 1967.
- [77] F. Bock, “An algorithm to construct a minimum directed spanning tree in a directed network,” *Developments in Operations Research*, vol. 1, pp. 29–44, 1971.
- [78] P. Humblet, “A distributed algorithm for minimum weight directed spanning trees,” *IEEE Transactions on Communications*, vol. 31, no. 6, pp. 756–762, 1983.