DIRECTED INFORMATION FOR COMPLEX NETWORK ANALYSIS FROM MULTIVARIATE TIME SERIES

by

Ying Liu

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Electrical Engineering

2012

ABSTRACT

DIRECTED INFORMATION FOR COMPLEX NETWORK ANALYSIS FROM MULTIVARIATE TIME SERIES

by

Ying Liu

Complex networks, ranging from gene regulatory networks in biology to social networks in sociology, have received growing attention from the scientific community. The analysis of complex networks employs techniques from graph theory, machine learning and signal processing. In recent years, complex network analysis tools have been applied to neuroscience and neuroimaging studies to have a better understanding of the human brain. In this thesis, we focus on inferring and analyzing the complex functional brain networks underlying multichannel electroencephalogram (EEG) recordings. Understanding this complex network requires the development of a measure to quantify the relationship between multivariate time series, algorithms to reconstruct the network based on the pairwise relationships, and identification of functional modules within the network.

Functional and effective connectivity are two widely studied approaches to quantify the connectivity between two recordings. Unlike functional connectivity which only quantifies the statistical dependencies between two processes by measures such as cross correlation, phase synchrony, and mutual information (MI), effective connectivity quantifies the influence one node exerts on another node. Directed information (DI) measure is one of the approaches that has been recently proposed to capture the causal relationships between two time series. Two major challenges remain with the application of DI to multivariate data, which include the computational complexity of computing DI with increasing signal length and the accuracy of estimation from limited realizations of the data. Expressions that can simplify the computation of the original definition of DI while still quantifying the causality relationship are needed. In addition, the advantage of DI over conventionally causality measures such as Granger causality has not been fully investigated. In this thesis, we propose time-lagged directed information and modified directed information to address the issue of computational complexity, and compare the performance of this model free measure with model based measures (e.g. Granger causality) for different realistic signal models.

Once the pairwise DI between two random processes is computed, another problem is to infer the underlying structure of the complex network with minimal false positive detection. We propose to use conditional directed information (CDI) proposed by Kramer to address this issue, and introduce the time-lagged conditional directed information and modified conditional directed information to lower the computational complexity of CDI. Three network inference algorithms are presented to infer directed acyclic networks which can quantify the causality and also detect the indirect couplings simultaneously from multivariate data.

One last challenge in the study of complex networks, specifically in neuroscience applications, is to identify the functional modules from multichannel, multiple subject recordings. Most research on community detection in this area so far has focused on finding the association matrix based on functional connectivity, instead of effective connectivity, thus not capturing the causality in the network. In addition, in order to find a modular structure that best describes all of the subjects in a group, a group analysis strategy is needed. In this thesis, we propose a multi-subject hierarchical community detection algorithm suitable for a group of weighted and asymmetric (directed) networks representing effective connectivity, and apply the algorithm to multichannel electroencephalogram (EEG) data.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor, Professor Selin Aviyente, for her guidance, encouragement, and support in every stage of my graduate study. Her knowledge, kindness, patience, passion, and vision affected me a lot and have provided me with lifetime benefits.

I am also grateful to my dissertation committee members, Professor Hayder Radha, Professor Ramakrishna Mukkamala, and Professor Pang-Ning Tan, for their valuable comments and suggestions on the thesis draft, as well as for the experience as a student with these three outstanding teachers. I would also like to thank many faculty members of MSU who were the instructors for the courses I took. The course works have greatly enriched my knowledge and provided the background and foundations for my thesis research.

My PhD study could have never been completed without the help of my fellow graduate students at MSU. I would like to express my special thanks to the colleagues at our lab, Ali Yener Mutlu, Marcos Bolanos, and Suhaily Cardona, for their suggestions, helps, and all the happy and tough time we have been through. I also want to express my thanks to Xi Lu, Lei Zhang, Shenglan Gao, Guanqun Zhang, Di Tang, Meng Cai, Ting Sun, Mingwu Gao, Jiankun Liu, Qiong Huo, Xiaochen Tang and Yuemin Jin, for their kind help during my four years at MSU, who enriched my study and life.

Finally, I would like to express my gratitude to my family. Their endless love and support always encourage me to deal with obstacles in every aspects of my life. In particular, I want to express my deepest gratitude to my dear husband, Jiayin, for his enduring love, encouragement, patience, and understanding.

TABLE OF CONTENTS

LIST OF TABLES . . . . **viii**

LIST OF FIGURES . . . . **ix**

CHAPTER 1 INTRODUCTION **1**

1.1 Measures to quantify the causality 5

1.1.1 Notations 6

1.1.2 Granger causality based measures 7

1.1.3 Information theoretic causality measures 9

1.1.4 Directed information versus other causality measures 12

1.2 EEG data 14

1.3 Overview of the contributions 16

CHAPTER 2 QUANTIFICATION OF EFFECTIVE CONNECTIVITY BY DIRECTED INFORMATION **18**

2.1 Introduction 18

2.2 Modified directed information 22

2.2.1 Problems with the implementation of directed information 22

2.2.2 Time-lagged directed information 25

2.2.3 Modified directed information 28

2.2.4 Modified directed information versus transfer entropy 30

2.2.5 Order selection 33

2.2.6 Normalization and significance testing 34

2.2.7 Performance of modified directed information 35

2.3 Application of DI to bivariate signal models 39

2.3.1 Simulated signal models 40

2.3.2 Biological data 48

2.4 Estimation of directed information 52

2.4.1 Estimation based on entropy estimation 52

2.4.2 Estimation based on mutual information and multi-information . . . 55

2.4.3 Performance comparison of estimators 58

2.5 Conclusions 61

CHAPTER 3 DIRECTED NETWORK INFERENCE BASED ON DIRECTED INFORMATION **64**

3.1 Introduction 64

3.2 Conditional directed information 66

3.2.1 Background 66

3.2.2 Motivational example 67

3.2.3 Computation of conditional directed information 68

3.2.4 Modified conditional directed information 71

3.3 Network inference algorithms . . . 73
3.3.1 Algorithm based on directed information and conditional directed information . . . 74
3.3.2 Algorithm based on time lag . . . 74
3.3.3 Algorithm based on modified time-lagged directed information and conditional directed information . . . 77
3.3.4 Validation . . . 78
3.4 Results . . . 80
3.4.1 Synthetic data: Linear network . . . 80
3.4.2 Synthetic data: Nonlinear network . . . 83
3.5 Discussions . . . 85
3.5.1 Problems with current algorithms . . . 85
3.5.2 Comparison with existing algorithms . . . 86
3.6 Conclusions . . . 87

CHAPTER 4 COMMUNITY DETECTION FOR DIRECTIONAL NEURAL NETWORKS . . . 88
4.1 Introduction . . . 88
4.2 Background . . . 91
4.2.1 Modularity . . . 91
4.2.2 Overview of community detection methods . . . 92
4.2.3 Group analysis approaches . . . 94
4.3 Algorithm for community detection . . . 95
4.3.1 Algorithm for community detection in weighted directed networks . . . 96
4.3.2 Algorithm for community detection for multiple subjects . . . 98
4.4 Results . . . 99
4.4.1 Directed vs. undirected networks . . . 99
4.4.2 Group analysis on synthetic data . . . 102
4.4.3 Group analysis on EEG Data . . . 104
4.5 Conclusions . . . 110

CHAPTER 5 SUMMARY AND FUTURE WORK . . . 113
5.1 Summary . . . 113
5.2 Future work . . . 115

APPENDIX A CONDITIONAL DIRECTED INFORMATION IN TWO GENERAL TRIVARIATE MODELS . . . 118

APPENDIX B COMPUTATION OF TIME-LAGGED CONDITIONAL DIRECTED INFORMATION . . . 121

BIBLIOGRAPHY . . . 129

LIST OF TABLES

2.1 Performance comparison for single order linear model 36

2.2 Performance comparison for multi-order linear model 38

2.3 Performance comparison for nonlinear model 38

2.4 Electrode pairs in the region of interest with significant DI values 51

2.5 Performance Comparison for linear model (128 realizations) 60

2.6 Performance Comparison for linear model (1024 realizations) 60

2.7 Performance Comparison for nonlinear model (128 realizations) 61

2.8 Performance Comparison for nonlinear model (1024 realizations) 61

3.1 Average F-score for three proposed algorithms and PDC for linear network . . . 82

3.2 Average F-score for three algorithms of nonlinear network 85

4.1 Average false discovery rate for group analysis methods 104

4.2 Comparison of Modularity for CRN with for random graphs 108

4.3 Comparison of Modularity for ERN with for random graphs 108

LIST OF FIGURES

2.1 Average information flow over 100 simulations for single order linear model computed using the original definition of DI, TLDI and MDI. 36

2.2 Modified DI for bivariate linear autoregressive model with multiple time lags. . . 37

2.3 Modified DI for nonlinear autoregressive model with multiple time lags. 39

2.4 Application of directed information and Granger causality to bivariate linear autoregressive model. 41

2.5 Application of directed information and Granger causality to bivariate nonlinear autoregressive model. 42

2.6 Application of directed information and Granger causality to linear mixing for both linear and nonlinear autoregressive models. 44

2.7 Application of directed information and Granger causality to single source model. 47

2.8 Application of directed information and Granger causality to two asymmetric coupled Lorenz oscillators. 48

2.9 Application of directed information and Granger causality to EEG data. 50

2.10 Averaged information flow for linear model using different estimators(128 realizations) 59

3.1 The DI value and time lag for both hub and chain models. 69

3.2 Algorithm 1: directed network inference based on DI and CDI 75

3.3 Algorithm 2: directed network inference based on TLDI and TL 77

3.4 Algorithm 3: directed network inference based on MDI and MCDI 79

3.5 The performance of proposed algorithms for linear network inference. 82

3.6 The performance of proposed algorithms for nonlinear network inference. 84

4.1 Hierarchical optimization of modularity by Blondel *et al.* [1]. 96

4.2 Algorithm 4: community detection of weighted networks 98

4.3 Algorithm 5: community detection of multiple weighted networks 100

4.4 Community detection of different representation of the network. 102

4.5 Average false discovery rate of the community detection algorithm for a simulated directed weighted network. 103

4.6 Applying the multi-subject community detection algorithm to 10 subjects for CRN. 106

4.7 Applying the multi-subject community detection algorithm to 10 subjects for ERN. 107

4.8 The information flow ratio of both response types and their corresponding random networks. 111

Chapter 1

INTRODUCTION

Complex networks are abound in nature and engineering, ranging from neural networks and protein interaction networks in biology to social networks in sociology and the internet in communication [2, 3, 4]. A network is referred to as a complex network not only because of its size, but also because of the nature of the interactions (e.g. nonlinear) between its subsystems and the behavior of the individual network nodes (dynamic) [5]. The analysis of complex networks employs techniques from graph theory, machine learning, statistical physics, and signal processing [6, 7, 8]. Complex networks across a range of applications are found to have similar macroscopic behavior such as small-world topology and scale-free distribution. However, these properties are not sufficient for a comprehensive understanding of the network at an intermediate scale. There is an interest to understand how these networks are structurally organized and change dynamically over time and frequency. In recent years, complex network analysis tools have been applied to neuroscience and neuroimaging studies and have resulted in a better understanding of the brain at a system level. In this dissertation, we focus on the analysis of the complex brain network using information-theoretic measures and attempt to gain some insights into the functionality of the brain.

One common approach to analyze the brain as a complex network is to describe it as a graph. Each node in the graph represents a particular region and the neuronal oscillations associated with it. Different neuroimaging modalities, such as electroencephalogram (EEG), magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI), can be used to record the brain activity and extract complex network characteristics from the human brain. However, compared to other neuroimaing modalities, EEG is able to record brain activity with higher temporal resolution and accuracy [9]. Once the neuroimaging recordings are collected, the first goal for brain network analysis is to determine the edges between the nodes or connectivity from the observation of multivariate time series. Three kinds of brain connectivity have been studied recently to determine and quantify the strength of the edge: anatomical connectivity, functional connectivity and effective connectivity [5]. Anatomical connectivity is the set of physical or structural connections linking neuronal units at a given time and can be obtained from measurements of the diffusion tensor [10]. Functional connectivity captures the statistical dependence between distributed and often spatially remote neuronal units by measuring their correlations in either time or frequency domain. Effective connectivity describes the set of causal effects of one neural system over another [5, 11], which can reflect both the interaction and the direction of the information flow in the system. These three types of connectivity help in the understanding of the functional segregation and functional integration of the brain. Functional segregation refers to a cortical area of the brain specializing for some aspects of perceptual or motor processing and this specialization is anatomically segregated within the cortex [12]. When a function involves different specialized areas, the union of these areas is mediated by functional integration [12]. The assessment of functional integration is important to understand how different areas of the brain coordinate with each other for a particular task. Both functional connectivity and effective connectivity can be used to describe the functional integration of the brain. In fact, most of the research in the analysis of brain networks focus on functional connectivity where the edges correspond to the pairwise correlation and result in undirected graphs, which is not usually sufficient to describe the actual neurological processes. On the contrary, as Friston pointed out functional integration of the brain can be better understood through effective connectivity since it reflects the dynamic (activity dependent and time dependent) characteristics of a system. In this sense, the brain network can be well described by effective networks where the edges of the graph have direction and the corresponding association matrix is no longer symmetric. Therefore, we expect that using effective connectivity would reveal new topological characteristics of the brain. This dissertation focuses mainly on the effective connectivity and the related network inference and community detection problems.

Measures to quantify the effective connectivity can be categorized into three groups, dynamic causal modeling (DCM), Granger causality based measures, and information-theoretic measures. Dynamic causal modeling employs a generative model to explain how activity in one brain area is affected by activity in another by using differential equations in continuous time [11, 13]. The parameters of these equations encode the strength of connections and how they change with experimental factors. DCM tries to find the best model that explains the data but it requires *a priori* knowledge of the system, such as the input of the system and hidden states. In addition, DCM is limited to networks with small size [12]. Granger causality is defined as a stochastic process $\mathbf{X}$ causing another process $\mathbf{Y}$ if the prediction of $\mathbf{Y}$ at the current time point, Y_n, is improved when taking into account the past samples of $\mathbf{X}$. Different from DCM, measures based on Granger causality assumes the data reflect states that cause each other and capture the dependencies among the observations directly. Therefore, Granger causality based methods are more viable and can be applied directly to any time series without knowing any knowledge about how the generating data are structured. However, in practice, Granger causality is usually implemented within a linear framework, e.g. bivariate or multivariate autoregressive models, and yielding methods such as directed transfer function (DTF), and partial directed coherence (PDC) [14, 15, 16]. These methods are limited to capturing linear relations and suffer from the common problems of parametric model, such as determination of the order. However, EEG recordings are simultaneously recorded at different locations of the brain, and are known to have nonlinear dependencies between recordings from different sites [17]. Measures that can address the issue of model dependency are needed. Recently, information theoretic tools [18, 19, 20], such as transfer entropy, directed transinformation, and directed information, address the issue of model dependency and have found numerous applications in neuroscience [21, 22, 23]. Transfer entropy (TE) proposed by Schreiber quantifies causality as the deviation of the observed data from the generalized Markov condition. Transfer entropy is based on a Markov assumption and the performance of transfer entropy depends on the estimation of transition probabilities, which requires the selection of order or memory of the Markov processes $\mathbf{X}$ and $\mathbf{Y}$ [24]. Directed transinformation (DT) introduced by Saito [19] measures the information flow from the current sample of one signal to the future samples of another signal given the past samples of both signals. However, this measure does not discriminate between totally dependent and independent processes [25]. Recently, directed information proposed by Marko [26] and later re-formalized by Massey, Kramer, Tatikonda and others has attracted attention for quantifying directional dependencies [20, 26, 27, 28, 29]. Directed information theory has been mostly aimed towards the study of communication channels with feedback. In recent years, new theoretical developments motivated the use of this measure in quantifying causality between two time series. In particular, Amblard *et al.* [29] recently showed how directed information and Granger causality are equivalent for linear Gaussian processes and proved key relationships between existing causality measures and the directed information. Therefore, there has been a growing interest in applying this measure to applications in signal processing, neuroscience and bioinformatics. One major issue remaining with the application of directed information is the estimation and computation of directed information from limited amount of data [20, 26, 27, 30]. Therefore, a simplified expression to reduce the dimensionality of DI estimation is needed. In addition, the comparison of DI with existing measures, in particular the model dependent measures based on Granger causality, is needed to verify its effectiveness for the analysis of neuroscience data.

Although directed information is effective at quantifying the relationship between pairs of neuronal populations, it is not sufficient to reveal the actual network structure. The DI value between two processes by itself can not reflect the true structure of the network. A large DI value does not guarantee direct causality between two time series, i.e., one signal may affect the other through the third signal [31]. Therefore, we use causal conditional directed information introduced by Kramer [27] to address this problem and propose multiple algorithms to infer the directed network. The inferred network can demonstrate the true effective connectivity between two processes and the system topology.

In most applications, discovering the global topology of the network is not sufficient. Motifs that can reflect the local organizational features of the network have also been of interest, using community detection and network classification methods [32, 33]. Previous work has shown that functional brain networks exhibit scale free and small-world properties and have a hierarchical structure [34]; and that the community structure of human brain changes with age and the task at hand [35]. For example, Fair *et al.* showed that young children and young adults have different community structures in functional brain networks from the study of resting state fMRI data [35]. Similarly, Ferrarini *et al.* showed that the resting-state human brain has a hierarchical functional module structure [36] and Meunier *et al.* revealed age-related changes in the modular structure of human brain functional networks from fMRI [37]. Chavez *et al.* pointed out that the modular structure of the human brain provides important information on the functional organization of brain areas during normal and pathological neural activities [38]. Therefore, in order to discover the underlying organization of the network, the partition of the brain network into small functional modules is needed. Traditional clustering algorithms require a *priori* knowledge about the number of clusters [39]. Therefore, modularity based algorithms are widely used to choose the best partitions of a network by maximizing the modularity. In many studies, it is important to discover these functional modules across multiple subjects. In this dissertation, we extend a greedy algorithm, Louvain method [1], to weighted and directed networks to find the functional communities of the brain across subjects.

1.1 Measures to quantify the causality

In this section, we will give a brief introduction to some popular Granger causality and information-theoretic based causality measures.

1.1.1 Notations

Before introducing the various measures to quantify the effective connectivity, we will first review some common notations and definitions that will be used throughout this dissertation. Let $\mathbf{X} = X^n = X_{1:n} = (X_1, \cdots, X_n)$ be a random process with length n and $p(x_1, \cdots, x_n) = p(x^n) = p(x_{1:n})$ be the joint probability of random variables $(X_1, \cdots, X_n)$. $DX^n = X^{n-1} = (0, X_1, \cdots, X_{n-1})$ will be used to define the time delayed version of sequence X^n, which is also equivalent to $X_{1:n-1}$.

Given two continuous random variables X and Y, the mutual information (MI) is defined as follows[1]:

$$I(X;Y) = \int\int p(x,y) \log \frac{p(x,y)}{p_x(x)p_y(y)} dxdy, \tag{1.1}$$

where $p(x,y)$ is the joint probability density function (pdf) of X and Y, and $p_x(x)$, $p_y(y)$ are the marginal pdfs of X and Y, respectively. $I(X;Y) \geq 0$ with equality if and only if X and Y are independent [40]. In information theory, mutual information can be interpreted as the amount of uncertainty about X that can be reduced by the observation of Y or the amount of information Y can provide about X, i.e., $I(X;Y) = H(X) - H(X|Y)$. Since $I(X;Y) \geq 0$, $H(X|Y) \leq H(X)$ with equality if and only if X and Y are independent, i.e., conditioning reduces entropy [40].

Mutual information has a natural generalization to multiple variables defined as multi-information (total correlation) [41]:

$$I_r[P_{1\cdots r}(y_1, y_2, \cdots, y_r)] = \int \cdots \int P_{1\cdots r}(y_1, y_2, \cdots, y_r) \log[\frac{P_{1\cdots r}(y_1, y_2, \cdots, y_r)}{p_1(y_1) \cdots p_r(y_r)}] d^r y. \tag{1.2}$$

Multi-information captures more collective properties than just pairwise relations as quantified by mutual information. The relationship between multi-information and mutual information is as follows [42]:

$$I(X^N, Y^N) = I(X^N; Y^N) + I(X^N) + I(Y^N), \tag{1.3}$$

[1] All integrals are from $-\infty$ to $+\infty$ unless otherwise specified.

where $I(X^N, Y^N)$ is the multi-information between $2N$ random variables $X_1, \cdots, X_N, Y_1, \cdots, Y_N$, and $I(X^N)$, $I(Y^N)$ are the multi-information between N random variables $X_1, \cdots, X_N$ and $Y_1, \cdots, Y_N$, respectively, while $I(X^N; Y^N)$ is the mutual information between two random vectors of length N.

For any three random variables X, Y and Z, if the conditional distribution of Z depends only on Y and is conditionally independent of X, i.e. $p(z|y) = p(z|yx)$, then X, Y and Z are said to form a Markov chain, denoted by $X \rightarrow Y \rightarrow Z$. In this case, the conditional mutual information between X and Y given Z defined as $I(X; Z|Y) = H(Z|Y) - H(Z|X, Y)$ is equal to 0 [40].

1.1.2 Granger causality based measures

Granger causality is widely used to describe the causality between two time series. It is defined as a stochastic process $\mathbf{X}$ causing another process $\mathbf{Y}$ if the prediction of $\mathbf{Y}$ at the current time point, Y_n, is improved when taking into account the past samples of $\mathbf{X}$. This approach is appealing but gives rise to many questions on how to apply this definition to real data [43]. Granger causality has commonly been implemented within a linear prediction framework using a bivariate autoregressive model. In this framework, the improvement of predicting Y_n is assessed by the change of the variances of the prediction errors when the signals are fitted by univariate and bivariate autoregressive models. For two univariate signal models,

$$X(n) = \sum_{i=1}^{p_1} \alpha_{xi} X(n-i) + \eta_x(n), \tag{1.4}$$

$$Y(n) = \sum_{i=1}^{p_2} \beta_{yi} Y(n-i) + \eta_y(n), \tag{1.5}$$

where p_1 and p_2 are the order of the random processes $\mathbf{X}$ and $\mathbf{Y}$, respectively, α_{xi} and β_{yi} are the autoregressive coefficients, and η_x (η_y) are the noise processes. In this model, the prediction of the current sample of Y_n only depends on the past samples of itself. While for a bivariate AR model,

$$
\begin{aligned}
X(n) &= \sum_{i=1}^{p_1} \alpha_{xi} X(n-i) + \sum_{i=1}^{p_3} \gamma_{xi} Y(n-i) + \eta_x(n), && (1.6) \\
Y(n) &= \sum_{i=1}^{p_2} \beta_{yi} Y(n-i) + \sum_{i=1}^{p_4} \gamma_{yi} X(n-i) + \eta_y(n), && (1.7)
\end{aligned}
$$

the prediction of each signal depends on the past samples of both signals.

Granger employs variance to evaluate the improvement of the prediction and the Granger causality from $\mathbf{X}$ to $\mathbf{Y}$ can be quantified as:

$$G_{\mathbf{X}\rightarrow\mathbf{Y}} = \ln\left(\frac{var(Y_n|Y^{n-1})}{var(Y_n|X^{n-1}Y^{n-1})}\right), \tag{1.8}$$

where if $var(Y_n|Y^{n-1}) > var(Y_n|X^{n-1}Y^{n-1})$, $\mathbf{X}$ causes $\mathbf{Y}$. If the past of $\mathbf{X}$ does not improve the prediction of Y_n, $G_{\mathbf{X}\rightarrow\mathbf{Y}}$ is close to zero.

The original definition of Granger causality is quantified in the time domain and is limited to bivariate models. Later, researchers introduced multivariate autoregressive (MVAR) models for multiple simultaneously recorded time series analysis and proposed methods such as directed transfer function (DTF), partial directed coherence (PDC) and directed partial correlation [44, 14, 16] to quantify Granger causality both in the time and the frequency domains in a multivariate setting. We will only brief introduce PDC here since it is the most widely used measure and will be used for comparisons with DI for network inference in Chapter 3.

Partial directed coherence

Consider an m-dimensional MVAR process with order p as follows:

$$\begin{pmatrix} X_1(n) \\ X_2(n) \\ \vdots \\ X_m(n) \end{pmatrix} = \sum_{r=1}^{p} \mathbf{A}_r \begin{pmatrix} X_1(n-r) \\ X_2(n-r) \\ \vdots \\ X_m(n-r) \end{pmatrix} + \begin{pmatrix} u_1(n) \\ u_2(n) \\ \vdots \\ u_m(n) \end{pmatrix} \tag{1.9}$$

where $u_i(n)$ with $i = 1, \cdots, m$ represents independent Gaussian white noise with covariance matrix $\sum$, and $\mathbf{A}_r$ with $r = 1, \cdots, p$ is the $m \times m$ coefficient matrix. The PDC measure from signal j to i is given by:

$$\pi_{i,j}(f) = \frac{\overline{a}_{i,j}(f)}{\sqrt{\overline{\mathbf{a}}_j^H(f)\overline{\mathbf{a}}_j(f)}}, \tag{1.10}$$

where $\overline{a}_{i,j}(f)$ is the i, jth entry of $\overline{\mathbf{A}}(f) = I - \mathbf{A}(f) = [\overline{\mathbf{a}}_1(f), \cdots, \overline{\mathbf{a}}_m(f)]$, $\mathbf{A}(f)$ is the Fourier transform of the coefficients, $\overline{\mathbf{a}}_j^H(f)$ is the Hermtian transpose of $\overline{\mathbf{a}}_j(f)$. The computation of $\pi_{i,j}(f)$ relies on the parameters of the MVAR model and thus the performance of PDC depends on the fitness of the MVAR model to the signal, which requires a proper choice of order p for the model and enough number of time samples to estimate the parameters.

Overall, all of these measures, e.g. Granger causality and PDC, are limited to capturing linear relations or require *a priori* knowledge about the underlying signal models [23].

1.1.3 Information theoretic causality measures

Mutual information can be extended to random vectors or sequences X^N and Y^N as $I(X^N; Y^N)$, where $I(X^N; Y^N) = H(X^N) - H(X^N|Y^N) = H(Y^N) - H(Y^N|X^N)$. However, mutual information is a symmetric measure and does not reveal any directionality or causality between two random sequences. Information theoretic tools [18, 19, 20], such as transfer entropy, directed transinformation and directed information, address the issue of model dependency and evaluate the prediction improvement by 'information (entropy)' directly.

Transfer entropy

Transfer entropy (TE) proposed by Schreiber computes causality as the deviation of the observed data from the generalized Markov condition and is defined as [18],

$$T^n_{\mathbf{X}\to\mathbf{Y}} = \sum_{y_n, y_{n-l:n-1}, x_{n-m:n-1}} p(y_n y_{n-l:n-1} x_{n-m:n-1}) \log \frac{p(y_n|y_{n-l:n-1} x_{n-m:n-1})}{p(y_n|y_{n-l:n-1})}, \tag{1.11}$$

where m and l are the orders (memory) of the Markov processes $\mathbf{X}$ and $\mathbf{Y}$, respectively and $p(y_n y_{n-l:n-1} x_{n-m:n-1})$ is the joint probability of random variables $(Y_n, Y_{n-l:n-1}, X_{n-m:n-1})$. When $n > \max(l, m)$, transfer entropy can be expressed in terms of mutual information as follows,

$$\begin{aligned} T^n_{\mathbf{X}\to\mathbf{Y}} &= \sum p(y_n y_{n-k:n-1} x_{n-l:n-1}) \log \frac{p(y_n|y_{n-k:n-1} x_{n-l:n-1})}{p(y_n|y_{n-k:n-1})} \\ &= \sum p(y_n y_{n-k:n-1} x_{n-l:n-1}) \log \frac{p(x_{n-l:n-1} y_n|y_{n-k:n-1})}{p(x_{n-l:n-1}|y_{n-k:n-1}) p(y_n|y_{n-k:n-1})} \\ &= I(X_{n-l:n-1}; Y_n|Y_{n-k:n-1}), \end{aligned} \tag{1.12}$$

where the last equality follows the definition of conditional mutual information, i.e. $I(X;Y|Z) = \sum p(xyz) \log \frac{p(xy|z)}{p(x|z)p(y|z)}$. The relationship between transfer entropy and conditional mutual information shown in the above equation has also been verified in [29]. It is important to note that transfer entropy is usually defined for a physical recording system, therefore, instantaneous information exchange is not considered. In addition, the definition of TE implies a stationary Markov assumption for the underlying system such that the state of Y_n only depends on the past l states of itself and the past m states of process $\mathbf{X}$, i.e., $p(y_n|y_{1:n-1}x_{1:n-1}) = p(y_n|y_{n-l:n-1}x_{n-m:n-1})$.

Directed transinformation

Directed transinformation (T) introduced by Saito [19] measures the information flow from the current sample of one signal to the future samples of another signal given the past samples of both signals. Directed transinformation is defined as,

$$DT(\mathbf{X} \to \mathbf{Y}) = \sum_{n=1}^{N} I(X_n; Y_{n+1:n+F}|X_{n-P:n-1} Y_{n-P:n-1} Y_n), \tag{1.13}$$

where $Y_{n+1:n+F} = (Y_{n+1} \cdots Y_{n+F})$ are the F future samples of $\mathbf{Y}$, the value of F and P changes with the current time sample n constrained by $F+P+1 = N$ where N is the length of the signal. Different from previously introduced measures, directed transinformation measures the influence of the current sample of $\mathbf{X}$ on the future samples of $\mathbf{Y}$. The definition of directed transinformation does not make any assumptions about the underlying model for the interactions. However, the computation of each term of the above equation requires the information of the whole time series, i.e. the joint probability estimation of $2N$ random variables, which is computationally very complex. In addition, directed transinformation can not discriminate between independent and identical processes since its value is equal to 0 for both cases [25].

Directed information

Massey addressed the issue of symmetry for mutual information by defining the directed information from a length N sequence $X^N = (X_1, \cdots, X_N)$ to $Y^N = (Y_1, \cdots, Y_N)$ [20] as follows:

$$\begin{aligned} DI(X^N \to Y^N) &= H(Y^N) - H(Y^N \| X^N) \\ &= \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}), \end{aligned} \tag{1.14}$$

where $H(Y^N \| X^N)$ is the entropy of the sequence Y^N causally conditioned on the sequence X^N and $H(Y^N \| X^N)$ is defined as:

$$H(Y^N \| X^N) = \sum_{n=1}^{N} H(Y_n | Y^{n-1} X^n), \tag{1.15}$$

which differs from $H(Y^N | X^N) = \sum_{n=1}^{N} H(Y_n | Y^{n-1} X^N)$ in that X^n replaces X^N in each term on the right-hand side of equation (1.15), i.e. only the causal influence of the time series $\mathbf{X}$ up to the current time sample n on the process $\mathbf{Y}$ is considered.

An alternative definition of the directed information is proposed by Tatikonda in terms of Kullback-Leibler (KL) divergence [28]. It shows that the difference between mutual information and directed information is the introduction of feedback in the definition of directed information [20, 29, 28]. Mutual information and directed information expressed by KL divergence are written as:

$$I(X^N; Y^N) = D_{KL}(p(x^N, y^N) \| p(x^N) p(y^N)), \tag{1.16}$$

$$DI(X^N \to Y^N) = D_{KL}(p(x^N, y^N) \| \overleftarrow{p}(x^N | y^N) p(y^N)), \tag{1.17}$$

where $\overleftarrow{p}(x^N|y^N) = \prod_{n=1}^{N} p(x_n|x^{n-1}y^{n-1})$ is the feedback factor influenced by the feedback in the system, i.e., the probability that the input $\mathbf{X}$ at current time is influenced by the past values of both itself and $\mathbf{Y}$. If there is no feedback, then $p(x_n|x^{n-1}y^{n-1}) = p(x_n|x^{n-1})$ and $\overleftarrow{p}(x^N|y^N) = p(x^N)$. In fact, $p(x^N, y^N) = \overleftarrow{p}(x^N|y^N)\overrightarrow{p}(y^N|x^N)$, where $\overrightarrow{p}(y^N|x^N) = \prod_{n=1}^{N} p(y_n|x^n y^{n-1})$ and is defined as the feedforward factor affected by the memory of the system. If the system is memoryless, then $p(y_n|x^n y^{n-1}) = p(y_n|x_n)$.

Entropy and mutual information are extensive quantities, which grow with the length of the signal. Thus, Shannon introduced entropy rate for stochastic processes. The entropy rate of a stochastic process $\{X_i\}$ is defined as $H(\mathscr{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n)$. In addition, when $\{X_i\}$ is a stationary stochastic process, the entropy rate is also the limit of the conditional entropy [40], i.e.,

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n) = \lim_{n\to\infty} H(X_n|X_1, \cdots, X_{n-1}), \tag{1.18}$$

When dealing with discrete valued processes, one can establish that, assuming stationarity, the directed information rate can be written as [27, 29],

$$\begin{aligned} DI_\infty(\mathbf{X}^N \to \mathbf{Y}^N) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} I(X^n; Y_n|Y^{n-1}), \\ &= \lim_{N\to\infty} I(X^N; Y_N|Y^{N-1}). \end{aligned} \tag{1.19}$$

which can further be decomposed into two parts as:

$$\begin{aligned} DI_\infty(\mathbf{X}^N \to \mathbf{Y}^N) &= \lim_{N\to\infty} I(X_{1:N-1}; Y_N|Y_{1:N-1}) + \lim_{N\to\infty} I(X_N; Y_N|X_{1:N-1}, Y_{1:N-1}) \\ &= DI_\infty(DX^N \to Y^N) + DI_\infty(X^N \to Y^N||X^{N-1}) \end{aligned} \tag{1.20}$$

where $DI_\infty(DX^N \to Y^N)$ is the directed information rate from delayed version of $\mathbf{X}^N$ to $\mathbf{Y}^N$, and $DI_\infty(X_{1:n} \to Y_{1:n}||X_{1:n-1})$ is the instantaneous information exchange rate.

1.1.4 Directed information versus other causality measures

Directed information has been mostly aimed towards the study of communication channels with feedback. In recent years, new theoretical developments motivated the use of this measure in quantifying causality between two time series. Amblard *et al.* [29] recently showed how directed information and Granger causality are equivalent for linear Gaussian processes and proved key relationships between existing causality measures and the directed information. Based on Granger's definition of causality, Geweke introduced the Geweke's indices to quantify the causal linear dependencies under Gaussian assumptions [45]. Amblard *et al.* proved that the directed information rate and Geweke's indices are equal for Gaussian processes [29] as indicated by,

$$DI_{\infty}(DX^{N} \rightarrow Y^{N}) = \frac{1}{2} \log \frac{\varepsilon_{\infty}^{2}(Y_{N}|Y^{N-1})}{\varepsilon_{\infty}^{2}(Y_{N}|Y^{N-1}X^{N-1})} = F_{X^{N} \rightarrow Y^{N}}, \tag{1.21}$$

where N is the length of the signal, $DI_{\infty}(X^{N} \rightarrow Y^{N})$ is the directed information rate, $\varepsilon_{\infty}^{2}(Y_{N}|Y^{N-1}) = \lim_{N\rightarrow\infty} \varepsilon^{2}(Y_{N}|Y^{N-1})$ is the asymptotic variance of the prediction residue when predicting Y_N from the observation of Y^{N-1}, and $F_{X^{N} \rightarrow Y^{N}}$ refers to the linear feedback measure from random processes X^N to Y^N defined by Geweke [29]. Moreover, directed information and Granger's approach are equivalent for multivariate time series in the case of Gaussian distributions [29]. In addition, Amblard *et al.* proved that for a stationary process without considering the instantaneous information exchange, the directed information rate is equal to $DI_{\infty}(DX^{N} \rightarrow Y^{N})$ and is equivalent to the transfer entropy when $l = m = n-1$ in equation (1.11). Al-khassaweneh *et al.* derived the relationship between directed information and directed transinformation that $DT(\mathbf{X} \rightarrow \mathbf{Y}) - DT(\mathbf{Y} \rightarrow \mathbf{X}) = DI(\mathbf{X} \rightarrow \mathbf{Y}) - DI(\mathbf{Y} \rightarrow \mathbf{X})$, which indicates that both measures reveal the same information about the difference of information flow in two directions [25]. However, compared to other measures, DI has several advantages. First, different from Granger causality implemented in an AR setting, DI is a model free measure and can quantify both the linear and nonlinear directional information flow. Second, transfer entropy and directed transinformation are equal to 0 for both independent and identical processes and they can not discriminate between these two types of processes. For DI when the two processes are independent, $DI = 0$; and when the two processes are identical, $DI = H(Y_n|Y_{n-1})$. Therefore, DI can discriminate between totally dependent and independent processes. On the other hand, when we consider two time series without instantaneous information exchange, DI is equal to $DI(DX \to Y)$ and cannot discriminate between identical and independent processes, either. However, in this case, the two processes will never be identical and we do not have to worry about it not discriminating between totally dependent and independent cases. Overall, DI can be applied to any class of signals without assumptions about the signal model (e.g. stationarity) or the interactions between signals (e.g. linear). Therefore, in this dissertation, we focus on the estimation and computation of DI, and apply this measure to network inference and community detection problems encountered in multichannel EEG recordings.

1.2 EEG data

With the advance of neuroimaging technology, EEG is able to record brain activity with higher temporal resolution and accuracy than ever before. In this dissertation, we analyze the brain network based on the EEG data provided by Dr. Jason Moser from the Department of Psychology at Michigan State University. Here we give a brief overview of the methods used for EEG data collection, which include subject recruitment, task and data reduction.

Participants

EEG data from ten undergraduates were drawn from an ongoing study of relationships between the error-related negativity (ERN) and individual differences[2]. ERN is a brain potential response that occurs following performance errors in a speeded reaction time task [48]. All participants retained for analysis made at least six errors for computation of stable ERNs, as in [49]. No participants discontinued their involvement once the experiment had begun.

Task

Participants completed a letters version of the Eriksen Flanker task [48]. Stimuli were presented on a Pentium R Dual Core computer, using Presentation software (Neurobehavioral systems, Inc.) to control the presentation and timing of stimuli, the determination of response accuracy, and the measurement of reaction times. During the task, participants

[2] Participants for the present analysis were drawn from samples reported on in [46, 47]

were presented with a string of five letters. Each five-letter string was either congruent (e.g. FFFFF) or incongruent (e.g. EEFEE) and participants were required to respond to the center letter (target) via the left or right mouse button. Trial types were varied randomly such that 50% of the trials were congruent. Letters were displayed in a standard white font on a black background and subtended 1.3° of visual angle vertically and 9.2° horizontally. A standard fixation mark (+) was presented during the inter-trial interval (ITI). Each trial began with the presentation of the flanking letters (i.e. EE EE). Flanking letters remained on the screen for 35 ms and were followed by the target (i.e. EEFEE), which remained for 100 ms (135 ms total presentation time). Each trial was followed by a variable ITI ($1200 - 1700$ ms). The entire experimental session consisted of 480 trials grouped into six blocks of 80 trials each. The letters constituting each string were varied between blocks (e.g., M and N in block 1 and E and F in block 2) and response-mappings were reversed at the midpoint of each block (e.g., left mouse-button click for M through 40 trials of block 1, then right-mouse button click for M for the last 40 trials of block 1) in order to elicit a sufficient number of errors for ERN calculation.

Psychophysiological Data Recording, Reduction and Analysis

Continuous electroencephalographic (EEG) activity was recorded by 64 Ag-AgCl electrodes placed in accordance with the 10/20 system. Electrodes were fitted in a BioSemi (BioSemi, Amsterdam, The Netherlands) stretch-lycra cap. In addition, two electrodes were placed on the left and right mastoids. The electro-oculogram (EOG) generated by eye-movements and blinks were recorded by FP1, as well as by electrodes placed below the right eye and on the left and right outer canthi, all approximately 1 cm from the pupil. During data acquisition, the Common Mode Sense active electrode and Driven Right Leg passive electrode formed the ground, as per BioSemi's design specifications. All bioelectric signals were digitized at 512 Hz using ActiView software (BioSemi). Offline Analysis were performed using BrainVision Analyzer 2 (BrainProducts, Gilching, Germany). Scalp electrode recordings were re-referenced to the numeric mean of the mastoids and band-pass filtered with cutoffs of 0.1 and 30 Hz (12 dB/oct rolloff). Ocular artifacts were then corrected using the regression method developed by Gratton *et al.* [50]. Response-locked data were segmented into individual epochs beginning 200 ms prior to the response and continued for 1000 ms. Individual trials were rejected on the basis of excessive physiological activity: a voltage step exceeding 50 μV between contiguous sampling points, a voltage difference of more than 200 μV within a trial, or a maximum voltage difference less than 0.5 V within a trial. Finally, the response-locked EEG was averaged across trials to yield error- and correct-trial ERPs for each site.

1.3 Overview of the contributions

The contributions of this dissertation can be divided into three parts: computation and estimation of directed information, directed network inference using directed information and conditional directed information, and community detection for multiple weighted directed networks.

In chapter 2, computation and estimation of directed information are realized through simplification of the definition of DI. The major contribution of this work can be summarized as follows:

1. Present the time-lagged directed information and modified directed information to reduce the computational complexity of computing DI while still quantifying the causal dependencies. Prove the relationship between the modified directed information and transfer entropy. Evaluate the performance of modified DI for quantifying causality for various realistic signal models with linear, nonlinear, and dynamic interactions.

2. Introduce a new directed information estimation method based on multi-information. Provide a quantitative comparison of various DI estimation methods.

In the second part of the proposed research, network inference algorithms based on directed information and conditional directed information are introduced with the following contributions:

1. Derive time-lagged conditional directed information and modified conditional directed information to reduce the computational complexity of estimating directed relationships from real data.

2. Propose three network inference algorithms for linear and nonlinear network inference and evaluate their performances on simulated network models.

In the third part of the proposed research, an improved community detection algorithm for weighted and directed networks is introduced with the following contributions:

1. Extend a hierarchical community detection algorithm from undirected networks to the directed case for identifying the modules in the effective brain network.

2. Propose a group analysis method to obtain a common community structure across subjects.

3. Evaluate the performance of the proposed community detection algorithm on both simulated and real EEG data sets for understanding the organization of the effective connectivity networks in the brain.

Chapter 2
QUANTIFICATION OF EFFECTIVE CONNECTIVITY BY DIRECTED INFORMATION

2.1 Introduction

In this chapter, we will focus on the quantification of effective connectivity to get a better understanding of the functional influence in the brain. The main approaches used to quantify the effective connectivity between two time series are model based measures and information-theoretic measures [51]. Granger causality based methods and dynamic causal modeling [13] are two widely used model based measures. Granger causality is a widely used measure to describe the causality between two time series. It defines a stochastic process $\mathbf{X}$ causing another process $\mathbf{Y}$ if the prediction of $\mathbf{Y}$ at the current time point, Y_n, is improved when taking into account the past samples of $\mathbf{X}$. This approach is appealing but gives rise to many questions on how to apply this definition to real data [43]. Granger causality has been mostly applied within a linear prediction framework using a multivariate autoregressive (MVAR) model yielding methods such as directed transfer function (DTF), partial directed coherence (PDC) and directed partial correlation [44, 14, 16, 52]. For example, Hesse *et al.* applied time-varying Granger causality to EEG data and found that conflict situation generates directional interactions from posterior to anterior cortical sites [14]. Kaminski *et al.* applied DTF to EEG recordings of human brain during stage 2 sleep and located the main source of causal influence [16]. Schelter *et al.* employed PDC to EEG recordings from a patient suffering from essential tremor [53]. The extensions of Granger-causality based methods, such as kernel Granger causality, generalized PDC (gPDC) and extended PDC (ePDC) have also found numerous applications in neuroscience [54, 55, 56]. However, Granger-causality based methods, especially those developed from MVAR models, are limited to capturing linear relations or require *a priori* knowledge about the underlying signal models [23]. These approaches may be misleading when applied to signals that are known to have nonlinear dependencies, such as EEG data [17]. DCM, on the other hand, can quantify nonlinear interactions by assuming a bilinear state space model. However, DCM requires *a priori* knowledge about the input to the system [13, 57] and is limited to a network with small size [12]. Thus, a model-free measure detecting both linear and nonlinear relationships is desired.

Information theoretic tools [18, 19, 20], such as transfer entropy [18], address the issue of model dependency and have found numerous applications in neuroscience [21, 22, 23]. Transfer entropy (TE) proposed by Schreiber computes causality as the deviation of the observed data from the generalized Markov condition. Sabesan *et al.* employed TE to i-dentify the direction of information flow for the intracranial EEG data and suggested that transfer entropy plays an important role in epilepsy research [24]. Wibral *et al.* applied TE to magnetoencephalographic data to quantify the information flow in cortical and cerebellar networks [58]. Vicente *et al.* extended the definition of TE and measured the information flow from $\mathbf{X}$ to $\mathbf{Y}$ by introducing a general time delay u and showed that TE has a better performance in detecting the effective connectivity for nonlinear interactions and signals affected by volume conduction such as real EEG/MEG recordings compared to linear methods [57]. The performance of transfer entropy depends on the estimation of transition probabilities, which requires the selection of order or memory of the Markov processes $\mathbf{X}$ and $\mathbf{Y}$ [24]. Directed transinformation (DT) introduced by Saito [19] measures the information flow from the current sample of one signal to the future samples of another signal given the past samples of both signals. Hinrichs *et al.* used this measure to analyze causal interactions in event related EEG-MEG experiments [23]. However, this measure does not discriminate between totally dependent and independent processes [25]. Recently, directed information proposed by Marko [26] and later re-formalized by Massey, Kramer, Tatikonda and others have attracted attention for quantifying directional dependencies [20, 26, 27, 28, 29]. Directed information theory has been mostly aimed towards the study of communication channels with feedback. In recent years, new theoretical developments motivated the use of this measure in quantifying causality between two time series. In particular, Amblard *et al.* [29] recently showed how directed information and Granger causality are equivalent for linear Gaussian processes and proved key relationships between existing causality measures and the directed information. Therefore, there has been a growing interest in applying this measure to applications in signal processing, neuroscience and bioinformatics. For example, it has been successfully used to infer genomic networks [3] and to quantify effective connectivity between neural spike data in neuroscience [4, 29, 59]. In order to detect both linear and nonlinear relationships, in this chapter, we propose directed information as a powerful measure to quantify the effective connectivity in the brain.

The theoretical advantages of DI over existing measures have been noted in literature [4, 29, 59]. However, until now the implementation and benefits of using DI for capturing the effective connectivity in the brain through neurophysiological data have not been illustrated thoroughly and formally. We will mainly address three issues in this chapter. First, one major issue with the application of directed information is the practical computation from limited data. Current studies of directed information focus on the stationary Gaussian processes due to the fact that the DI of a Gaussian process can be easily obtained even with longer time series and limited sample sizes [30]. However, most complex systems are nonlinear and not all nodes of the network follow Gaussian distributions. In this case, directed information can be expressed in terms of mutual information or joint entropy and estimators such as k-nearest neighborhood based methods or m-spacing based estimators can be used. However, when the length of the signal increases, the computational complexity and the bias of the DI estimator increases immensely. Therefore, a simplified expression for DI to reduce the dimensionality of estimation is needed. In previous work, DI measure was applied either to limited time series such as every two time samples [2] or to a known parametric signal model to overcome this problem [4]. In this chapter, we show that applying directed information to short-time windows such as every two time samples may lose most of the causal dependencies between two random processes. In order to address this issue, we propose modified directed information to simplify the expression of DI and reduce the computational complexity while still quantifying the causal dependencies. In addition, we prove some key relationships between transfer entropy and the modified directed information. Second, since DI can be expressed in terms of entropy or mutual information, current applications, such as genomic network inference and neural network inferrence, compute DI using either entropy or mutual information based estimators [3, 60, 31]. However, traditional joint entropy estimation methods of multiple random variables are inaccurate and inefficient when the data space is sparse. In order to overcome the inadequacy of entropy estimators for DI estimation, we introduce an alternative representation of DI in terms of multi-information, which can be estimated by extending the mutual information estimator proposed by Darbellay [61] to multiple random variables using an adaptive partitioning of the observed space [42]. Moreover, different applications put different constraints on the estimation methods, so it is important to evaluate the performance of the estimators in terms of bias and variance, and select an appropriate DI estimator for different systems or applications. In this chapter, we offer an extensive analysis of different DI estimation methods in terms of the bias, variance, computational efficiency and discrimination power through simulations.

Finally, once the problem for computation and estimation of DI has been addressed, we focus on the application of DI on EEG data. Because of the relationship between Granger causality and directed information, in this chapter, we will also compare the performance of these two measures and investigate the advantage of DI over Granger-causality based model measures. Theoretical developments only proved the equivalence between these two measures for the case that the time series are distributed as Gaussian in a linear model. However, to date there has not been much work that compares the actual performance of DI and Granger causality based measures for realistic signal models, including both linear and nonlinear interactions. This chapter addresses this issue by evaluating the performance of DI and Granger causality based methods under a common framework without making any assumptions about the data distribution.

In this chapter, we first illustrate the problems related to the computation of DI. We then propose a modified directed information measure that simplifies the DI computation by reducing the order of the joint entropy terms while still quantifying the causal dependencies. In addition, we provide a DI estimator based on multi-information. We then evaluate the performance of DI for quantifying the effective connectivity for linear and nonlinear autoregressive models, linear mixing models, single source models and dynamic chaotic oscillators in comparison to existing causality measures, in particular with Granger causality. Finally, we apply our method to multichannel EEG data to detect the effective connectivity in the brain.

2.2 Modified directed information

2.2.1 Problems with the implementation of directed information

To apply directed information to real data, the first issue we need to solve is the computation of DI. According to the definition of DI in equation (1.14), in practice we need to estimate the conditional mutual information $I(X^n; Y_n|Y^{n-1})$, where $n = 1, \cdots, N$. In fact, $I(X^n; Y_n|Y^{n-1})$ quantifies the causal information flow from X to Y at time point n, since $I(X^n; Y_n|Y^{n-1}) = DI(X^n \to Y^n) - DI(X^{n-1} \to Y^{n-1})$. In addition, as $n \to \infty$, $I(X^n; Y_n|Y^{n-1})$ is the directed information rate for a stationary process. $I(X^n; Y_n|Y^{n-1})$ can be expanded using entropy, mutual information, and multi-information. Therefore, DI can be expressed:

- In terms of entropy as:

$$\begin{aligned} DI(X^N \to Y^N) &= \sum_{n=1}^{N}[(H(Y^n) - H(Y^{n-1})) - (H(X^nY^n) - H(X^nY^{n-1}))], \\ &= \sum_{n=1}^{N}[H(X^nY^{n-1}) - H(X^nY^n)] + H(Y^N). \end{aligned} \tag{2.1}$$

- In terms of mutual information as:

$$DI(X^N \to Y^N) = \sum_{n=1}^{N} [I(X^n; Y^n) - I(X^n; Y^{n-1})]. \tag{2.2}$$

- In terms of multi-information as:

$$\begin{aligned} DI(X^N \to Y^N) &= \sum_{n=1}^{N} [(I(X^n, Y^n) - I(X^n, Y^{n-1})) - (I(Y^n) - I(Y^{n-1}))], \\ &= \sum_{n=1}^{N} [(I(X^n, Y^n) - I(X^n, Y^{n-1}))] - I(Y^N). \end{aligned} \tag{2.3}$$

From the above equations, we can observe that the computation of DI requires the estimation of joint probabilities of high dimensional random variables over time. If X_n and Y_n are normally distributed, the joint entropy can be estimated based on the covariance matrices as follows:

$$DI(X^N \to Y^N) = \sum_{n=1}^{N} \frac{1}{2} \log \frac{|cov(X_1 \cdots X_n Y_1 \cdots Y_{n-1})||cov(Y_1 \cdots Y_n)|}{|cov(X_1 \cdots X_n Y_1 \cdots Y_n)||cov(Y_1 \cdots Y_{n-1})|}, \tag{2.4}$$

where $|cov(X_1, X_2, \cdots, X_n)|$ is the determinant of the covariance matrix of n random variables $X_1, X_2, \cdots, X_n$ and N is the length of the signal. The complexity of computing the original definition of DI through equation (2.4) is $O(N^4)$ (using LU decomposition [62]). However, for EEG data, the distribution is usually not Gaussian. The non-parametric entropy and mutual information estimators, such as plug-in estimator, m-spacing estimator, and Kozachenko and Leonenko (KL) estimator, can be used to estimate DI [63, 61]. When the length of the signal increases, the computational complexity, the bias, and the variance of these estimators increase immensely with limited sample sizes. Methods that can reduce the dimension and simplify the computation of DI are needed.

In order to simplify the estimation of DI, we first clarify the connection between the definition of DI used in information theory and the definition as it applies to physical time series. In a physical recording system, if $\mathbf{X}$ starts to influence $\mathbf{Y}$ after p_1 time points or with a delay of p_1 samples, we need to record at least $N + p_1$ time points to obtain N points of the time sequence $\mathbf{Y}$ that have been affected by $\mathbf{X}$. The directed information rate from time series X^{N+p_1} to Y^{N+p_1} can be defined as [27],

$$\begin{aligned}
& DI_{\infty}(X^{N+p_1} \rightarrow Y^{N+p_1}) \\
= & \lim_{N+p_1 \rightarrow \infty} \frac{1}{N+p_1} \sum_{n=1}^{N+p_1} I(X^n; Y_n | Y^{n-1}) && (2.5) \\
= & \lim_{N+p_1 \rightarrow \infty} I(X^{N+p_1}; Y_{N+p_1} | Y^{N+p_1-1}) && (2.6) \\
= & \lim_{N+p_1 \rightarrow \infty} [H(Y_{N+p_1} | Y^{N+p_1-1}) - H(Y_{N+p_1} | X^{N+p_1} Y^{N+p_1-1})] && (2.7) \\
= & \lim_{N+p_1 \rightarrow \infty} [H(Y_{N+p_1} | Y_{p_1+1:N+p_1-1}) - H(Y_{N+p_1} | X^{N+p_1} Y_{p_1+1:N+p_1-1})] && (2.8) \\
= & \lim_{N+p_1 \rightarrow \infty} [H(Y_{N+p_1} | Y_{p_1+1:N+p_1-1}) - H(Y_{N+p_1} | X_{1:N} Y_{p_1+1:N+p_1-1})] && (2.9) \\
= & \lim_{N+p_1 \rightarrow \infty} I(X_{1:N}; Y_{N+p_1} | Y_{p_1+1:N+p_1-1}) && (2.10) \\
= & \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^{N} I(X^n; Y_{n+p_1} | Y_{p_1+1:n+p_1-1}) && (2.11) \\
= & DI_{\infty}(X_{1:N} \rightarrow Y_{p_1+1:p_1+N}), && (2.12)
\end{aligned}$$

where equation (2.8) comes from the fact that $Y_{1:p_1}$ is independent of Y_{N+p_1}, and equation (2.9) is derived using the fact that $X_{N+1:N+p_1}$ has no effect on Y_{N+p_1} because of the time delay p_1 between these two time series. For two physical recordings $\mathbf{X}$ and $\mathbf{Y}$ with length $N + p_1$ and a lag of p_1, the last equation shows that DI rate for these two time series is equivalent to DI rate for two random processes with length N that are not synchronized in time. In fact, $Y_{p_1+1:p_1+N}$ may be indexed as $Y_{1:N}$ when using the information theoretic indexing, which indexes the signal not according to the physical time point but based on when the receiver receives its first piece of information. Therefore, directed information rate computed by using physical time indices is equivalent to the directed information rate using information theoretic indices for two systems that interact through a time delay. Moreover, when the length of the signal is long enough, the directed information value using both indices will be equivalent. Once the definition of directed information is extended from random vectors to two physical time series, we propose time-lagged DI and modified DI to simplify the computation of DI.

2.2.2 Time-lagged directed information

As we mentioned before, when the length N of the signal increases, the computational complexity, the bias, and the variance of estimating DI increase immensely with limited sample sizes. In addition, the directed information defined for the physical system is actually a DI with a lag of p_1 samples over a time window with length N. Therefore, an intuitive way to simplify the computation is to apply DI with lag p_1 over a small window. For example, in [2], the authors applied DI to gene $\mathbf{X}$ and $\mathbf{Y}$ at every two time samples, i.e., $DI_n(X_nX_{n+1} \to Y_nY_{n+1})$, with the assumption that the value of gene $\mathbf{X}$ is only influenced by the values of the other genes at one previous time step, i.e. a first order Markov model assumption. However, when $\mathbf{X}$ influences $\mathbf{Y}$ with a delay of p_1, we apply DI to every two samples of these two time series X^N and Y^N at the nth time sample with a time delay of p_1 $(n > p_1)$:

$$\begin{aligned}
&DI_n(X_{n-p_1}X_{n-p_1+1} \to Y_nY_{n+1}) \\
&= I(X_{n-p_1}; Y_n) + I(X_{n-p_1}X_{n-p_1+1}; Y_{n+1}|Y_n) \\
&= H(X_{n-p_1}) + H(X_{n-p_1}X_{n-p_1+1}Y_n) + H(Y_nY_{n+1}) \\
&- H(X_{n-p_1}Y_n) - H(X_{n-p_1}X_{n-p_1+1}Y_nY_{n+1}),
\end{aligned} \tag{2.13}$$

where $n = p_1 + 1, \cdots, N - 1$. However, in practice, the actual time lag of the two time series is unknown, and the estimated time lag d is used to compute DI for every two samples, i.e. replacing p_1 with d in equation (2.13). Thus the main question we need to answer is how much actual information flow is captured if the estimated time lag d is used in equation (2.13). To answer this question, we first consider a single order bivariate linear autoregressive model with delay p_1 (in this case, the maximum order of the model is also equal to p_1) as,

$$\begin{aligned} x_i &= u_i; \\ y_i &= bx_{i-p_1} + v_i \end{aligned} \tag{2.14}$$

where u_i and v_i are white Gaussian noise samples following $N(0, \sigma^2)$ and the order (delay) of the model is p_1. The actual directed information value can be computed by estimating the covariance matrix. For this model each term in the DI expression in equation (2.4) is simplified as:

$$\begin{aligned} \frac{1}{2}\log\frac{|cov(X_1\cdots X_nY_1\cdots Y_{n-1})|}{|cov(X_1\cdots X_nY_1\cdots Y_n)|} &= \begin{cases} -\frac{1}{2}\log{(b^2+1)\sigma^2},\ n=1,\cdots,p_1 \\ 0,\ n>p_1 \end{cases} \\ \frac{1}{2}\log\frac{|cov(Y_1\cdots Y_n)|}{|cov(Y_1\cdots Y_{n-1})|} &= \frac{1}{2}\log{(b^2+1)\sigma^2}. \end{aligned} \tag{2.15}$$

Therefore,

$$\begin{aligned} I(X^n;Y_n|Y_{1:n-1}) &= \frac{1}{2}\log\frac{|cov(X_1\cdots X_nY_1\cdots Y_{n-1})||cov(Y_1\cdots Y_n)|}{|cov(X_1\cdots X_nY_1\cdots Y_n)||cov(Y_1\cdots Y_{n-1})|} \\ &= \begin{cases} 0,\ n=1,\cdots,p_1 \\ \frac{1}{2}\log{(b^2+1)\sigma^2},\ n>p_1 \end{cases} \end{aligned} \tag{2.16}$$

Based on equations (1.19) and (2.16), when $p_1 \ll N$, the directed information rate is:

$$DI_\infty(X^N \to Y^N) = \lim_{N\to\infty} I(X^N;Y_N|Y^{N-1}) = \frac{1}{2}\log{(b^2+1)\sigma^2} \tag{2.17}$$

On the other hand, when DI measure is computed over every two samples of X^N and Y^N, the time-lagged DI given by equation (2.13) when replacing p_1 with d can be simplified as follows:

$$DI_n(X_{n-d}X_{n-d+1} \to Y_nY_{n+1}) = \begin{cases} 0,\ d<p_1-1 \text{ or } d\geq p_1+1 \\ \frac{1}{2}\log{(b^2+1)\sigma^2},\ d=p_1-1 \\ \log{(b^2+1)\sigma^2},\ d=p_1 \end{cases} \tag{2.18}$$

The time-lagged DI rate per sample is defined as:

$$\overline{DI_d} = \lim_{N\to\infty} \frac{1}{2(N-d-1)} \sum_{n=d+1}^{N-1} DI_n(X_{n-d}X_{n-d+1} \to Y_nY_{n+1}), \tag{2.19}$$

where d is the delay variable, $d = 0, \cdots, p$ and p is the largest possible time delay ($p \geq d$). The factor of 2 is introduced since the time-lagged DI rate is averaged over samples. In addition, when $d = p_1$, $\overline{DI_{p_1}} = \frac{1}{2}\log(b^2+1)\sigma^2$, which is equal to the directed information rate, i.e.,

$$DI_\infty(X^N \to Y^N) = \overline{DI_{p_1}}. \tag{2.20}$$

Therefore, motivated by the model in equation (2.14), for any general autoregressive model with a single order, we define the time-lagged directed information over the whole time series as follows:

$$\begin{aligned} TLDI_d(X^N \to Y^N) &= \sum_{n=1}^{d} I(X^n; Y_n|Y^{n-1}) + \overline{DI_d} \times (N-d) \\ &= \sum_{n=1}^{d} I(X^n; Y_n|Y^{n-1}) + \lim_{N\to\infty} \frac{N-d}{2(N-d-1)} \sum_{n=d+1}^{N-1} DI_n(X_{n-d}X_{n-d+1} \to Y_nY_{n+1}) \\ &= \sum_{n=1}^{d} I(X^n; Y_n|Y^{n-1}) + \lim_{N\to\infty} \frac{1}{2} \sum_{n=d+1}^{N-1} DI_n(X_{n-d}X_{n-d+1} \to Y_nY_{n+1}) \end{aligned} \tag{2.21}$$

where $d = 0, \cdots, p$. The first part of righthand side of above equation is the initial information flow from X^N to Y^N when $n \leq d$. The second part is motivated by the fact that when the information flow becomes stable, DI along the time series can be approximated by the time-lagged DI rate times the number of time samples. When $d = p_1$, where p_1 is the actual time delay between $\mathbf{X}$ and $\mathbf{Y}$, the first term in the righthand side of the above equation is 0. The TLDI will capture the same amount of information as computing DI over the whole time series. Moreover, it performs much better than the two sample DI (d equals to 0) used in previous work, especially when p_1 is large. For example, for the model in equation (2.14), when $p_1 \geq 2$, $\overline{DI_0} = 0$, which does not capture any of the causal dependencies.

Therefore, compared to equation (2.1), the computational complexity of computing DI using equation (2.21) is highly reduced. If $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed, we see that the complexity of using the original definition of DI in equation (2.4) is $O(N^4)$ (using LU decomposition [62]), while the complexity of computing TLDI for two time samples is $O(N)$. Therefore, time-lagged DI is more computationally efficient. However, the time-lagged DI is equivalent to the original definition of DI when the estimated delay d is equal to the actual time delay of the system p_1, i.e. the signals $\mathbf{X}$ and $\mathbf{Y}$ follow a single order model, and Y_n only depends on one past sample of itself, Y_{n-1}. However, these assumptions are not always true. Therefore, we propose the modified DI to address these issues.

2.2.3 Modified directed information

Consider a general Markov model, where X^N and Y^N are time series with a lag of p_1 and $p(Y_n|X_{1:n-p_1}, Y_{p_1+1:n-1}) = p(Y_n|X_{n-p_2:n-p_1}, Y_{n-p_3:n-1})$, where $p_2 \geq p_1$, $p_3 \geq 1$, p_2 is the order of the process $\mathbf{X}$, and p_3 is the order of the process $\mathbf{Y}$. In this model, it is assumed that $\mathbf{X}$ starts to influence $\mathbf{Y}$ with a delay of p_1 samples and the order of the model is $\max(p_2, p_3)$. When the length of the signal N is large enough, then equation (2.12) can be further simplified as,

$$\begin{aligned} DI(X^N \to Y^N) &= DI(X_{1:N-p_1} \to Y_{p_1+1:N}) \\ &= \sum_{n=p_1+1}^{N} I(X_{1:n-p_1}, Y_n|Y_{p_1+1:n-1}) \\ &= \sum_{n=p_1+1}^{N} [H(Y_n|Y_{p_1+1:n-1}) - H(Y_n|X_{1:n-p_1}Y_{p_1+1:n-1})]. \end{aligned} \tag{2.22}$$

Since

$$p(Y_n|X_{1:n-p_1}, Y_{p_1+1:n-1}) = p(Y_n|X_{n-p_2:n-p_1}, Y_{n-p_3:n-1}), \tag{2.23}$$

then $X_{1:n-p_2-1}Y_{p_1+1:n-p_3-1} \to X_{n-p_2:n-p_1}, Y_{n-p_3:n-1} \to Y_n$ follows a Markov chain.

According to Markov chain property,

$$\begin{aligned}
&I(X_{1:n-p_2-1}Y_{1:n-p_3-1};Y_n|X_{n-p_2:n-p_1}Y_{n-p_3:n-1}) \\
&= H(Y_n|X_{n-p_2:n-p_1}Y_{n-p_3:n-1}) - H(Y_n|X_{1:n-p_1}Y_{p_1+1:n-1}) = 0,
\end{aligned} \tag{2.24}$$

which means $H(Y_n|X_{n-p_2:n-p_1}Y_{n-p_3:n-1}) = H(Y_n|X_{1:n-p_1}Y_{p_1+1:n-1})$. Therefore,

$$\begin{aligned}
DI(X^N \to Y^N) &= \sum_{n=p_1+1}^{N} [H(Y_n|Y_{p_1+1:n-1}) - H(Y_n|X_{1:n-p_1}Y_{p_1+1:n-1})] \\
&= \sum_{n=p_1+1}^{N} [H(Y_n|Y_{p_1+1:n-1}) - H(Y_n|X_{n-p_2:n-p_1}Y_{n-p_3:n-1})] \\
&\leq \sum_{n=p_1+1}^{N} [H(Y_n|Y_{n-p_3:n-1}) - H(Y_n|X_{n-p_2:n-p_1}Y_{n-p_3:n-1})] \\
&= \sum_{n=p_1+1}^{N} I(X_{n-p_2:n-p_1};Y_n|Y_{n-p_3:n-1}),
\end{aligned} \tag{2.25}$$

where the second equality uses the Markov property and the inequality comes from the fact that conditioning reduces entropy. For a general Markov model, where X^N and Y^N are stationary statistical processes without instantaneous interaction, e.g. $p(Y_n|X_{1:n-p_1}, Y_{p_1+1:n-1})$ $= p(Y_n|X_{n-p_2:n-p_1}, Y_{n-p_3:n-1})$, the modified directed information (MDI) is defined as the upper bound of DI:

$$MDI(X^N \to Y^N) = \sum_{n=p+1}^{N} I(X_{n-p_2} \cdots X_{n-p_1}; Y_n|Y_{n-p_3} \cdots Y_{n-1}), \tag{2.26}$$

where in practice we let $p_1 = 1$, $p = \max(p_2, p_3)$ to reduce the number of parameters. Note that letting $p_1 = 1$ does not lose any of the information flow compared to using the actual time delay, $p_1 > 1$. The only drawback of letting $p_1 = 1$ is that the computational complexity of estimating the joint entropies increases since the length of the window to compute MDI increases and the dimensionality increases. The main reason why we let $p_1 = 1$ is because estimating the actual value for the delay accurately is not practical when the amount of data is limited. In a lot of similar work such as in [57], different values of p_1 are tested to choose the best one which is not computationally efficient either.

According to equation (2.25), modified directed information is the upper bound of directed information, i.e. $MDI \geq DI$. Moreover, MDI is a more general extension of time-lagged DI and has two major advantages. First, MDI considers the influence of multiple past samples of $\mathbf{Y}$ on the DI value. Second, it takes into account models with multiple orders, i.e. $\mathbf{Y}$ is influenced by different time lags of $\mathbf{X}$. The modified directed information extends the length of the window from 2 to p, which is closer to the actual information flow. When $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed, the computational complexity of the MDI is $O(p^3N)$ and is more efficient than that of the original definition of DI.

2.2.4 Modified directed information versus transfer entropy

Both the modified DI and transfer entropy are defined based on a Markov signal model. Therefore, in this subsection, we will explore the relationship between them in detail. Based on the definition of transfer entropy given in equation (1.11), we should note that transfer entropy is defined for a physical recording system, therefore, instantaneous information exchange is not considered. In addition, the definition of TE implies a Markov assumption of the system that the state of Y_n only depends on the past l states of itself and the past m states of process $\mathbf{X}$, i.e. $p(y_n|y_{1:n-1}x_{1:n-1}) = p(y_n|y_{n-l:n-1}x_{n-m:n-1})$.

Therefore, to explore the relationship between transfer entropy and directed information, the DI should be derived under the same assumptions of no instantaneous information exchange and a Markov model. For two random processes $\mathbf{X}$ and $\mathbf{Y}$ without instantaneous information exchange, the directed information and directed information rate are expressed as,

$$\begin{aligned} DI(X^N \to Y^N) = DI(DX^N \to Y^N) &= \sum_{n=1}^{N} I(X^{n-1}; Y_n|Y^{n-1}), \\ DI_\infty(X^N \to Y^N) = DI_\infty(DX^N \to Y^N) &= \lim_{N\to\infty} I(X^{N-1}; Y_N|Y^{N-1}). \end{aligned} \tag{2.27}$$

Based on the assumption that the system can be approximated by a Markov process, we derive a formula to show the relationship between the rate of directed information and transfer entropy as shown in the following theorem.

Theorem 1 *If X^N and Y^N are two stationary Markov processes with $p(y_n|y_{1:n-1}x_{1:n-1}) = p(y_n|y_{n-l:n-1}x_{n-m:n-1})$, then the upper bound of the directed information rate, i.e., modified DI rate, is equal to the transfer entropy.*

proof 1

Based on the Markov assumption of these two processes, i.e. $p(y_n|y_{1:n-1}x_{1:n-1}) = p(y_n|y_{n-l:n-1}x_{n-m:n-1})$, $X_{1:n-m-1}Y_{1:n-l-1} \to X_{n-m:n-1}Y_{n-l:n-1} \to Y_n$ follows a Markov chain. According to Lemma 1,

$$\begin{aligned} &I(Y_n; X_{1:n-m-1}Y_{1:n-l-1}|X_{n-m:n-1}Y_{n-l:n-1}) \\ &= H(Y_n|X_{n-m:n-1}Y_{n-l:n-1}) - H(Y_n|X_{1:n-1}Y_{1:n-1}) = 0, \end{aligned} \tag{2.28}$$

implying $H(Y_n|X_{n-m:n-1}Y_{n-l:n-1}) = H(Y_n|X_{1:n-1}Y_{1:n-1})$. Therefore,

$$\begin{aligned} I(X^{n-1}; Y_n|Y^{n-1}) &= H(Y_n|Y_{1:n-1}) - H(Y_n|X_{1:n-1}Y_{1:n-1}) \\ &= H(Y_n|Y_{1:n-1}) - H(Y_n|X_{n-m:n-1}Y_{n-l:n-1}) \\ &\leq H(Y_n|Y_{n-l:n-1}) - H(Y_n|X_{n-m:n-1}Y_{n-l:n-1}) \\ &= I(X_{n-m:n-1}; Y_n|Y_{n-l:n-1}) \\ &= T^n_{\mathbf{X}\to\mathbf{Y}}, \end{aligned} \tag{2.29}$$

where the last equality comes from equation (1.12). The inequality follows from the fact that conditioning reduces entropy and the equality holds when $Y_{1:n-l-1}$ is conditionally independent of Y_n given $Y_{n-l:n-1}$, i.e. when l is large enough that the influence of $Y_{1:n-l-1}$ on Y_n can be ignored, or when $n \leq l$, i.e. $Y_{n-l:n-1} = Y_{1:n-1}$.

The directed information rate in a physical recording system can be expressed as,

$$\begin{aligned} DI_\infty(DX^N \to Y^N) &= \lim_{N\to\infty} I(X^{N-1}; Y_N|Y^{N-1}) \\ &\leq \lim_{N\to\infty} I(X_{N-m:N-1}; Y_N|Y_{N-l:N-1}) \\ &= \lim_{N\to\infty} T^N_{\mathbf{X}\to\mathbf{Y}}. \end{aligned} \tag{2.30}$$

Therefore, when $l = m = N-1$, $DI_\infty(DX^N \to Y^N) = \lim_{N\to\infty} T^N_{\mathbf{X}\to\mathbf{Y}}$, which is aligned with previous work in [29]. For stationary Markov processes, when m and l are fixed, $\lim_{N\to\infty} I(X_{N-m:N-1}; Y_N | Y_{N-l:N-1})$ is equal to the rate of modified DI in equation (2.26) with $p_1 = 1$, $p_2 = m$, and $p_3 = l$. Moreover, in order to reduce the computational complexity, we usually let $l = m$. In this way, in practice, the limit (rate) of transfer entropy is the upper bound of directed information rate and is equal to the modified DI rate.

In summary, transfer entropy and directed information are very closely related to each other. Transfer entropy quantifies the information gained at each time step by measuring the deviation of the observed data from the generalized Markov condition. Therefore, the definition of transfer entropy implicitly assumes a stationary Markov process [29]. Compared to transfer entropy, directed information quantifies the sum of information obtained over the whole time series [64] and does not make any assumptions about the underlying signal model. Thus, theoretically, the original definition of directed information can be applied to any signal model. In real applications, in order to simplify the computation of directed information, we usually make certain assumptions about the underlying signal model such as the modified DI proposed in this dissertation, which basically assumes a stationary Markov process similar to transfer entropy. In addition, Amblard *et al.* proved that for a stationary process, directed information rate can be decomposed into two parts, one of which is equivalent to the transfer entropy when $l = m = n-1$ in equation (1.11) and the other to the instantaneous information exchange rate [29]. In another words, for a physical system without instantaneous interactions between its subsystems, the rate of these two measures, directed information and transfer entropy, are equivalent asymptotically as the length of the signal goes to infinity. When l or m is fixed and not equal to $n-1$, transfer entropy rate is the upper bound of directed information rate and is equal to the modified DI rate.

2.2.5 Order selection

For the implementation of MDI, we need to determine the maximum order of the model p. Criterions such as Akaike's Final Prediction Error (FPE) can be used to determine the order of the signal model p. However, this criterion is based on the assumption that the original signal follows a linear AR model and may lead to false estimation of the order when the underlying signal model is nonlinear. Therefore, model-free order selection methods, such as the embedding theorem [65], are needed. For the simplification of computation or parameter estimation, we are only interested in a limited number of variables that can be used to describe the whole system. Suppose we have a time series $(X_1, \cdots, X_n)$, the time-delay vectors can be reconstructed as $(X_n, X_{n-\tau}, X_{n-2\tau}, \cdots, X_{n-(d-1)\tau})$. Projecting the original system to this lower dimensional state space depends on the choice of d and τ, and the optimal embedding dimension d is related to the order of the model $p = d$ [57]. A variety of measures such as mutual information can be used to determine τ. For discrete time signals, usually the best choice of τ is 1 [66]. To determine d, Cao criterion based on the false nearest neighbor procedure [57] is used to determine the local dimension. The underlying concept of nearest neighbor is that: if d is the embedding dimension of a system, then any two points that stay close in the d-dimensional reconstructed space are still close in the $(d+1)$-dimensional reconstructed space; otherwise, these two points are false nearest neighbors [66, 57]. The choice of d, i.e., the model order p, is important for DI estimation. If d is too small, we will lose some of the information flow from $\mathbf{X}$ to $\mathbf{Y}$. If it is too large, the computational complexity of MDI will be very high, causing the bias and the variance of the estimators to increase.

2.2.6 Normalization and significance testing

Since $DI(X^N \to Y^N) + DI(Y^N \to X^N) = I(X^N;Y^N) + DI(X^N \to Y^N||DX^N)$ and $DI(X^N \to Y^N) = DI(DX^N \to Y^N) + DI(X^N \to Y^N||DX^N)$ [27], then

$$\begin{aligned} DI(X^N \to Y^N) + DI(Y^N \to X^N) &= DI(DX^N \to Y^N) + DI(X^N \to Y^N||DX^N) \\ &\quad + DI(DY^N \to X^N) + DI(Y^N \to X^N||DY^N). \end{aligned} \tag{2.31}$$

Therefore,

$$DI(DX^N \to Y^N) + DI(DY^N \to X^N) + DI(Y^N \to X^N||DY^N) = I(X^N;Y^N), \tag{2.32}$$

where $DI(Y^N \to X^N||DY^N) = DI(X^N \to Y^N||DX^N)$ indicating the instantaneous information exchange between processes $\mathbf{X}$ and $\mathbf{Y}$. For a physical system without instantaneous causality, i.e. $I(X^N \to Y^N||DX^N) = 0$, then $DI(X^N \to Y^N) + DI(Y^N \to X^N) = I(X^N;Y^N)$ and $0 \leq DI(X^N \to Y^N) \leq I(X^N;Y^N) < \infty$. A normalized version of DI, which maps DI to the $[0,1]$ range is used for comparing different interactions,

$$\rho_{DI}(X^N \to Y^N) = \frac{DI(X^N \to Y^N)}{I(X^N;Y^N)} = \frac{DI(X^N \to Y^N)}{DI(X^N \to Y^N) + DI(Y^N \to X^N)}, \tag{2.33}$$

where for a unidirectional system $\mathbf{X} \to \mathbf{Y}$ with no instantaneous interaction between $\mathbf{X}$ and $\mathbf{Y}$, $\rho_{DI}(X^N \to Y^N) = 1$ and $\rho_{DI}(Y^N \to X^N) = 0$; otherwise, if there is no causal relationship between the two signals, the values of $\rho_{DI}(X^N \to Y^N)$ and $\rho_{DI}(Y^N \to X^N)$ are very close to each other.

In order to test the null hypothesis of noncausality, the causal structure between $\mathbf{X}$ and $\mathbf{Y}$ is destroyed. For each process with multiple trials, we shuffle the order of the trials of the time series $\mathbf{X}$ 100 times to generate new observations $\mathbf{X}_m^*$, $m = 1, \cdots, 100$. In this way, the causality between $\mathbf{X}$ and $\mathbf{Y}$ for each trial is destroyed, and the estimated joint probability changes [67]. We compute the DI for each pair of data ($\mathbf{X}_m^*$ and $\mathbf{Y}$). A threshold is obtained at a $\alpha = 0.05$ significance level such that 95% of the directed information for randomized pairs of data ($DI(\mathbf{X}_m^* \to \mathbf{Y})$) is less than this threshold. If the DI value of the original pairs of data is larger than this threshold, then it indicates there is significant information flow from $\mathbf{X}$ to $\mathbf{Y}$.

2.2.7 Performance of modified directed information

In this section, we compare the performances of three different approaches, computing DI using the original definition, proposed time-lagged DI and modified DI. For the proposed TLDI and MDI, the order of the model p is determined by the Cao criteria. TLDI is computed over different time lags. The comparison is based on three different simulation models. Without loss of generality, we repeat each simulation 100 times to quantify the mean and variance of different computation approaches. In addition, for the linear models, we obtain the bias by comparing the means of different approaches with the theoretical DI value.

First, we test the performance of the proposed time-lagged DI for a bivariate linear autoregressive model given as follows:

$$\begin{aligned} X_i &= u_i; \\ Y_i &= 0.5 \times X_{i-2} + v_i; \end{aligned} \tag{2.34}$$

where u_i and v_i are white Gaussian noise samples following $N(0, 1)$ and the order of the model is 2. We generate 2048 realizations for each time series, and compute the DI value using three different approaches over a $N = 8$ block of time samples, respectively. The order of the model p is equal to 2. TLDI is computed over different time lags $d = 0, \cdots, 3$. From Figure. 2.1, we can observe that when $d = 2$, TLDI_2 reaches its maximum value, which is aligned with the order of the model. However, when $d \neq 2$, the TLDI_d will lose most of the causal dependencies. Therefore, computing DI over two samples with $d = 0$, as is done in previous work [2], is not sufficient for high order models. In addition, from Table 2.1, we can observe that TLDI_2 and MDI have lower bias and are more computationally efficient (less computation time) compared to the original definition of DI. MDI is larger than the theoretical DI as anticipated by the theoretical bound in Section 2.2. DI using the original definition has the largest bias, which is due to the fact that the bias of each term in equation (2.4) increases with the length of the signal for limited number of realizations.

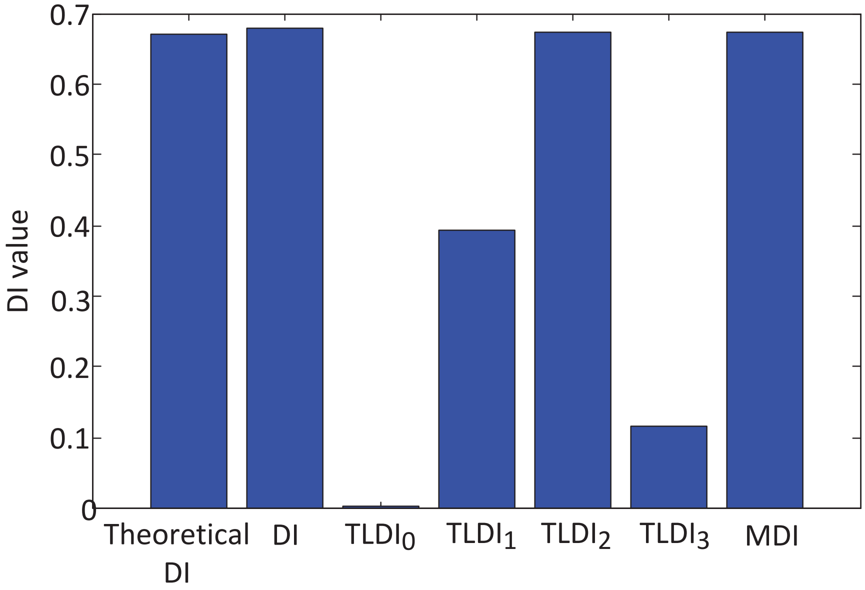


Figure 2.1: Average information flow over 100 simulations for single order linear model computed using the original definition of DI, TLDI and MDI. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Table 2.1: Performance comparison for single order linear model

	Bias	Variance ($\times 10^{-3}$)	Computation time (s)
DI	0.0095	0.6032	0.0511
TLDI$_2$	0.0032	0.6356	0.0233
MDI	0.0036	0.5937	0.0235

Next we test the performance of DI computation over different time windows on a multiple order bivariate linear autoregressive model as follows:

$$\begin{aligned} X_i &= 0.5 \times X_{i-1} + 0.4 \times X_{i-3} + u_i; \\ Y_i &= 0.6 \times X_{i-2} + 0.3 \times X_{i-4} + 0.5 \times Y_{i-3} + v_i; \end{aligned} \tag{2.35}$$

where u_i and v_i are white Gaussian noise samples following $N(0,1)$ and the maximum time delay between $\mathbf{X}$ and $\mathbf{Y}$ is 4. We generate 2048 realizations for each time series and compute DI using different measures over a length of time samples $N = 12$. The causal information flow from $\mathbf{X}$ to $\mathbf{Y}$ at time sample i ($I(X^i; Y_i|Y^{i-1})$) is plotted in Figure. 2.2(a). We can observe that, beginning at time sample 4 when $\mathbf{X}$ starts to influence $\mathbf{Y}$, the MDI at each time point is larger than DI, which is aligned with the inequality in equation (2.25). The DI, TLDI with different time delays $d = 0, \cdots, 4$, and MDI ($p = 4$), averaged over simulations, are shown in Figure. 2.2(b). We can observe that MDI is slightly larger than the theoretical value of DI , because MDI is the upper bound of directed information. Moreover, we apply the sign test to test whether there is a statistically significant difference between MDI and the actual DI, and the hypothesis that there is no difference between the two measures can not be rejected at the 5% significance level. Therefore, MDI can be used to replace DI to reduce the computational complexity. From Table 2.2, we can observe that MDI has lower bias and variance than TLDI, and is more computationally efficient than using the original definition of DI with the computation time being cut in half. Moreover, for the TLDI, it is hard to choose a proper d when the model is multi-order.

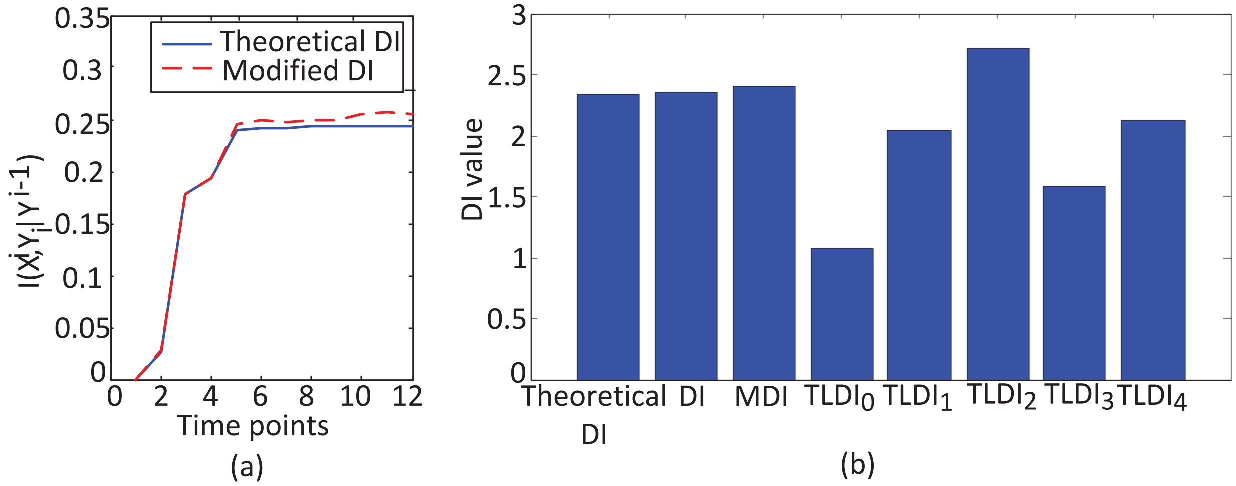


Figure 2.2: Modified DI for bivariate linear autoregressive model with multiple time lags. (a) The increase in the amount of information flow ($I(X^i; Y_i|Y^{i-1})$) at time point i. (b) Average total information flow over 100 simulations using the original definition of DI, TLDI and MDI.

Finally, we evaluate the performance of MDI for a nonlinear autoregressive model,

$$\begin{aligned} X_i &= 0.3 \times X_{i-1} + u_i; \\ Y_i &= 0.8 \times \sqrt{X_{i-1}} + \frac{0.2}{0.1 - 4 \times e^{X_{i-2}}} + v_i; \end{aligned} \tag{2.36}$$

Table 2.2: Performance comparison for multi-order linear model

	Bias	Variance	Computation time (s)
DI	0.0216	0.0014	0.1037
TLDI$_2$	0.3751	0.0029	0.0333
TLDI$_4$	-0.2148	0.0022	0.0360
MDI	0.0699	0.0015	0.0458

where u_i and v_i are white Gaussian noise samples following $N(0,1)$ and the maximum time delay between $\mathbf{X}$ and $\mathbf{Y}$ is 2. We generate 8192 realizations for each time series and compute DI and MDI ($p = 2$) over a length of time samples $N = 6$. The results averaged over 100 simulations are shown in Figure. 2.3. We can observe that the MDI is larger than computing DI using the original definition. Although the DI value computed over the whole time series does not necessarily reflect the actual information flow, for a stationary model, when $i > 2$, the information flow in the system becomes stable, i.e., $I(X^i, Y_i|Y^{i-1})$ should not change much. However, using the original definition of DI, this value is not stable as shown in Figure. 2.3(a). This is due to the MI estimator used for computing DI, which has bias and variance that increase with the number of joint variables for a limited sample size [61]. On the other hand, the dimension and the number of joint pdfs are fixed for MDI as seen in equation (2.26), which also leads to the lower variance of MDI in Table 2.3. In addition, the computation time of MDI is only one fourth of the original DI, which is important for detecting the nonlinear causality of a complex network with large number of nodes. Therefore, the MDI is preferred over DI computed over the whole time series because of its reduced computational complexity and stable performance.

Table 2.3: Performance comparison for nonlinear model

	Mean	Variance	Computation time (s)
DI	0.6109	0.0069	95.3133
MDI	1.2342	0.0010	24.4649

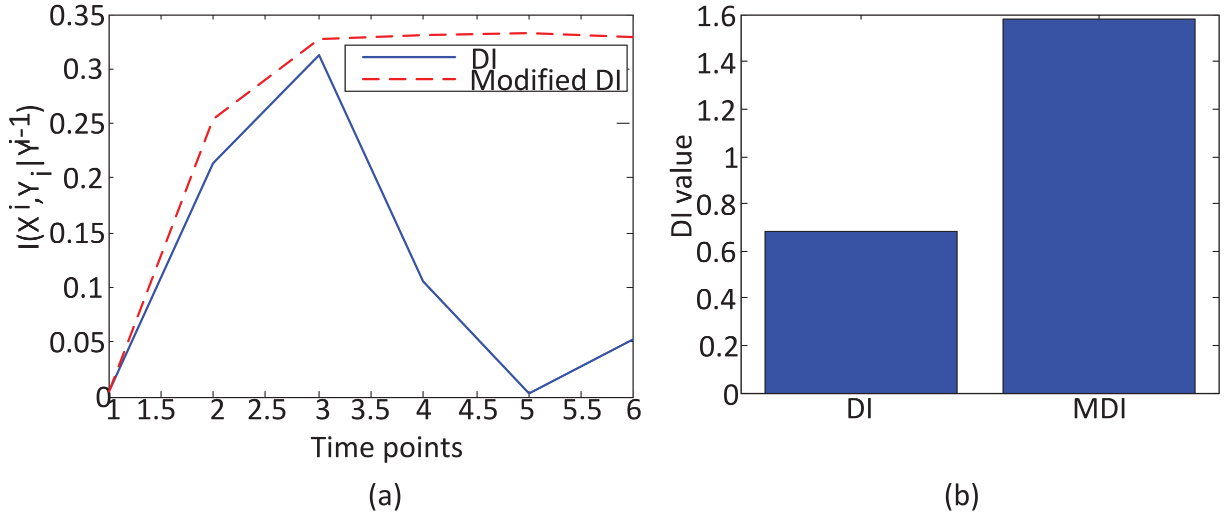


Figure 2.3: Modified DI for nonlinear autoregressive model with multiple time lags. (a) The increase in the amount of information flow $(I(X^i; Y_i|Y^{i-1}))$ at time point i. (b) Average total information flow over 100 simulations using the original definition of DI and MDI.

2.3 Application of DI to bivariate signal models

After we address the computation problem of DI, in this section, we test the validity and evaluate the performance of DI for quantifying the effective connectivity. We generate five different simulations. We use these simulation models to compare DI with classical Granger causality (GC) for quantifying causality of both linear and nonlinear autoregressive models, linear mixing models, single source models, and dynamic Lorenz systems. The Matlab toolbox developed by Seth is used to compute the GC value in the time domain. GC is also normalized to the $[0, 1]$ range for comparison purposes [68]. The performance of GC depends on the length of the signal, whereas the performance of DI relies on the number of realizations of time series. Therefore, for each simulation, the length of the generated signal for implementing GC is equal to the number of realizations for DI. The significance of DI values are evaluated by shuffling along the trials, while the significance of GC values are evaluated by shuffling along the time series.

2.3.1 Simulated signal models

Example 1: Multiple order bivariate linear autoregressive model

In this example, we evaluate the performance of DI on a general bivariate linear model,

$$X(n) = \sum_{i=1}^{p_4} \alpha_i X(n-i) + \sigma_x \eta_x(n-1), \tag{2.37}$$

$$Y(n) = \sum_{i=1}^{p_3} \beta_i Y(n-i) + \gamma \sum_{i=p_1}^{p_2} X(n-i) + \sigma_y \eta_y(n-1). \tag{2.38}$$

In this bivariate AR model with a delay p_1 and order $p_2 - p_1 + 1$, γ controls the coupling strength between the signals $\mathbf{X}$ and $\mathbf{Y}$. The initial values of $\mathbf{X}$ and $\mathbf{Y}$, and the noise η_x and η_y are all generated from a Gaussian distribution with mean 0 and standard deviation 1. All coefficients (α_i, β_i, σ_x and σ_y) are generated from Gaussian distributions with zero mean and unit variance with unstable systems being discarded. To evaluate the performance of directed information, we generate the bivariate model 4096 times with the same parameters but different initial values. γ is varied from 0.1 to 1 with a step size of 0.1, $p_1 = 1$ and $p_2 = p_3 = p_4 = 5$, i.e. $\mathbf{Y}$ is influenced by $\mathbf{X}$ through multiple time lags. Without loss of generality, we repeat the simulation 10 times, and average $DI(X^N \rightarrow Y^N)$ and $DI(Y^N \rightarrow X^N)$ over 10 simulations for different γ values. For each simulation, the threshold is evaluated by trial shuffling and the average threshold is obtained. For GC, the length of the generated signal is chosen as 4096, which is the same as the number of realizations for DI. The GC values in two directions and the corresponding thresholds at the 5% significance level are obtained. The DI value in two directions averaged across 10 simulations with different γ are shown in Figure 2.4(a). The performance of GC is shown in Fig 2.4(b). The estimated order of the model is $p = 5$, which is in accordance with the simulation model. We observe that $DI(X^N \rightarrow Y^N)$ is significant for all values of γ. On the contrary, $DI(Y^N \rightarrow X^N)$ is less than the threshold, which indicates the acceptance of the null hypothesis that there is no significant causal information flow from $\mathbf{Y}$ to $\mathbf{X}$. Since GC uses a linear autoregressive framework for quantifying causality, in this example, GC detects the causality relationship between $\mathbf{X}$ and $\mathbf{Y}$ successfully, i.e. the information flow from $\mathbf{X}$ to $\mathbf{Y}$ is significant for all γ while it is insignificant for the opposite direction. It is also interesting to note that GC and DI exhibit similar behavior across different values of γ, indicating the equivalency of the two measures for linear Gaussian signal models.

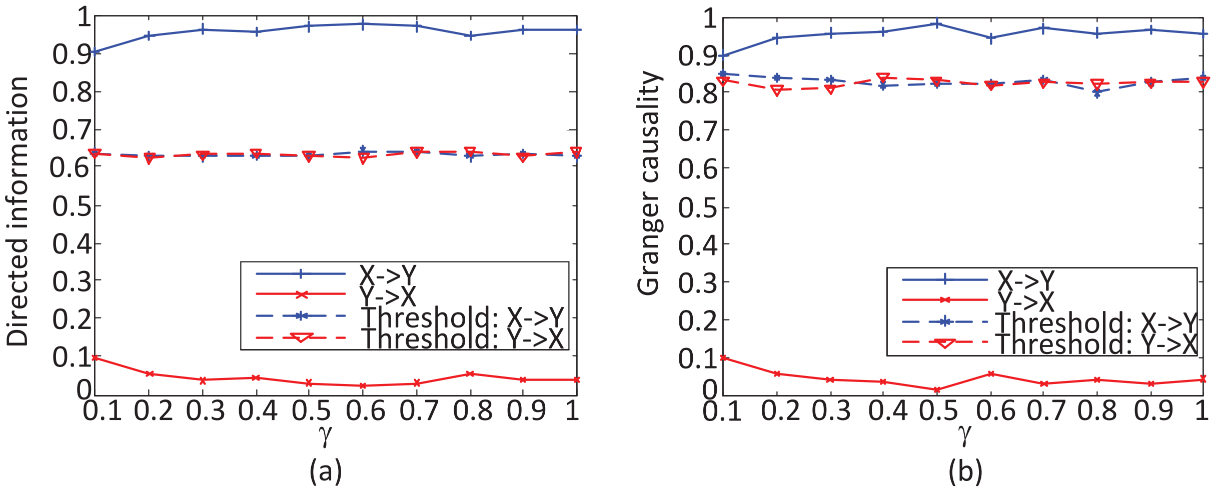


Figure 2.4: Application of directed information and Granger causality to bivariate linear autoregressive model. (a) Directed information with different γ. (b) Granger causality with different γ.

Example 2: Multiple order bivariate nonlinear autoregressive model

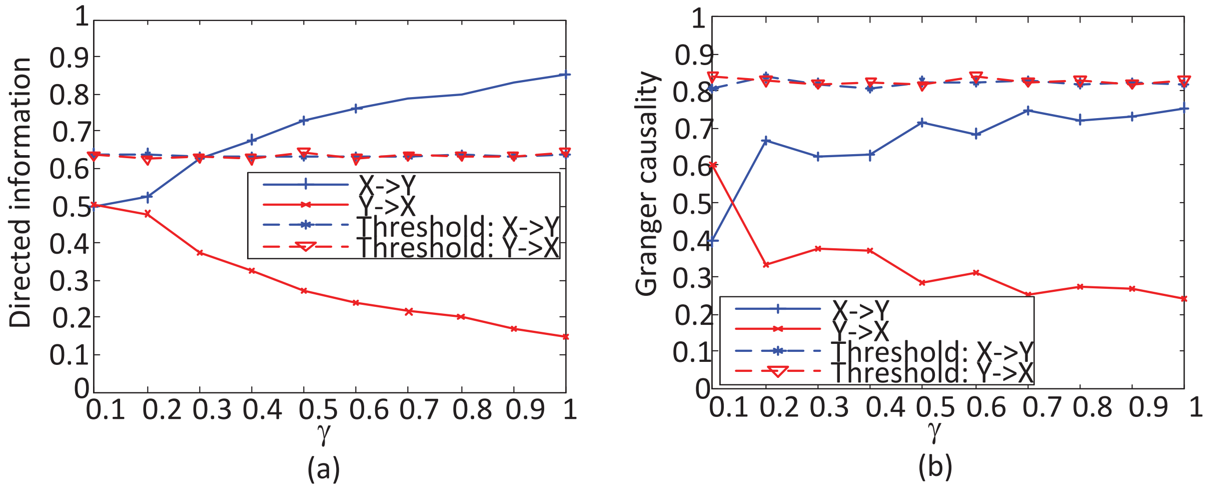
In this example, we evaluate the performance of DI on a general bivariate nonlinear model,

$$X(n) = \sum_{i=1}^{p_4} \alpha_i X(n-i) + \sigma_x \eta_x(n-1), \tag{2.39}$$

$$Y(n) = \sum_{i=1}^{p_3} \beta_i Y(n-i) + \gamma \sum_{i=p_1}^{p_2} \frac{1}{1+\exp(b_1 + b_2 X(n-i))} + \sigma_y \eta_y(n-1). \tag{2.40}$$

For this bivariate nonlinear AR model, the setting for the coupling strength γ and the generation of $\mathbf{X}$, $\mathbf{Y}$, η_x, η_y, α_i, β_i, σ_x, σ_y, p_1, p_2, p_3 and p_4 are the same as in Example 1. $\mathbf{Y}$ and $\mathbf{X}$ interact nonlinearly through the sigmoid function. Parameters of this function b_1 and b_2 control the threshold level and slope of the sigmoidal curve, respectively. We set $b_1 = 0$ and $b_2 = 50$. DI value and its threshold are averaged over 10 simulations for different

γ. The GC values in two directions and the corresponding thresholds at 5% significance level are obtained. The performance of DI and GC for the nonlinear autoregressive model in equations (2.39) and (2.40) averaged across 10 simulations with different γ are evaluated as shown in Figure 2.5. The estimated order of the model is 5. We observe that when γ is less than 0.3, the coupling strength between $\mathbf{X}$ and $\mathbf{Y}$ is weak and the DI value in both directions is not significant. As γ increases, $DI(X^N \rightarrow Y^N)$ increases and becomes significant. $DI(Y^N \rightarrow X^N)$ decreases with increasing γ and is still less than the threshold as expected. The results indicate increased unidirectional information flow from $\mathbf{X}$ to $\mathbf{Y}$ with increasing γ and show that detecting the information flow in nonlinear processes is more difficult especially when the coupling strength is low. GC fails to detect the information flow from $\mathbf{X}$ to $\mathbf{Y}$ for all γ. Since GC is implemented in a linear framework, the estimated order and the model itself do not match with the nonlinearity of the signal. Therefore, it cannot detect nonlinear causality.



(a)

(b)

Figure 2.5: Application of directed information and Granger causality to bivariate nonlinear autoregressive model. (a) Directed information with different γ. (b) Granger causality with different γ.

Example 3: Linear Mixing model

In this example, we test the effectiveness of DI in inferring effective connectivity when there is linear mixing between two signals. Linear instantaneous mixing is known to exist in human noninvasive electrophysiological measurements such as EEG or MEG. Instantaneous mixing from coupled signals onto sensor signals by the measurement process degrades signal asymmetry [57]. Therefore, it is hard to detect the causality between the two signals. For unidirectional coupled signal pairs $\mathbf{X} \rightarrow \mathbf{Y}$ described in equations (2.37) to (2.40), we create two linear mixtures $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ as follows,

$$X_\epsilon(n) = (1-\epsilon)X(n) + \epsilon Y(n), \tag{2.41}$$

$$Y_\epsilon(n) = \epsilon X(n) + (1-\epsilon)Y(n), \tag{2.42}$$

where ϵ controls the amount of linear mixing and is varied from 0.05 to 0.45 with a step size of 0.05, and γ is fixed to 0.8 for both models. When $\epsilon = 0.5$, the two signals are identical. Both DI and GC are used to quantify the information flow between $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ in the two directions. The DI value and GC value averaged across 10 simulations with changing linear mixing coefficient ϵ for both linear and nonlinear AR models are shown in Figure 2.6. The estimated order of the model is 5 as before. When $\epsilon = 0.5$, the two observed mixing signals are identical and we expect to see no significant information flow in the two directions. We observe that for the linear AR model, directed information detects the causality between $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ when ϵ is smaller than 0.4. When ϵ is larger than 0.4, the causality between $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ is hard to detect because of the strong mixing, i.e., $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ are almost identical, and the information flow in both directions becomes insignificant. Compared to DI, GC only detects the causality from $\mathbf{X}_\epsilon$ to $\mathbf{Y}_\epsilon$ when the mixing is weak ($\epsilon < 0.2$), indicating that GC is more vulnerable to linear mixing. It is probably due to the fact that GC is sensitive to the mixture of signals and the assumed signal model does not match with the original signal [69]. For the nonlinear AR model, DI fails to detect causality when ϵ is larger than 0.1, which indicates that linear mixing of nonlinear source models makes it harder to detect effective connectivity compared to mixing of linear source models. On the other hand, GC fails to detect any causality even when $\epsilon = 0$, since it cannot detect nonlinear interactions.

Example 4: Single source model

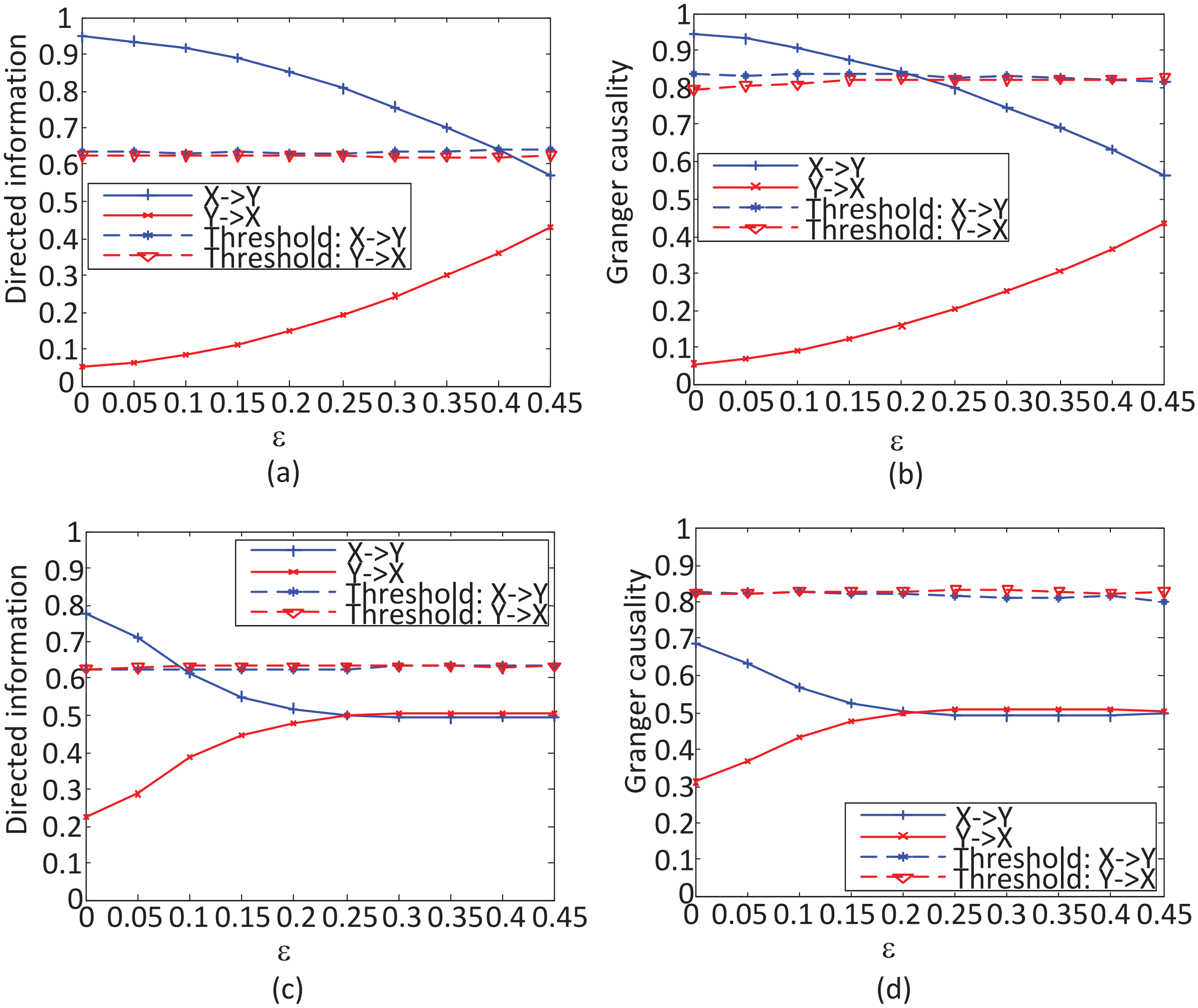


Figure 2.6: Application of directed information and Granger causality to linear mixing for both linear and nonlinear autoregressive models. (a) Directed information with different ϵ for the linear mixing of linear AR model. (b) Granger causality with different ϵ for the linear mixing of linear AR model. (c) Directed information with different ϵ for the linear mixing of nonlinear AR model. (d) Granger causality with different ϵ for the linear mixing of nonlinear AR model.

A single source is usually observed on different signals (channels) with individual channel noises [57], which is common in EEG signals due to the effects of volume conduction. In this case, false positive detection of effective connectivity occurs for methods such as Granger causality [69], which means GC has low specificity. We generate two signals $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ as follows to test the specificity of DI when there is no significant information flow from one signal to the other signal.

$$
\begin{aligned}
S(n) &= \sum_{i=1}^{p_4} \alpha_i S(n-i) + \eta_S(n); &(2.43)\\
X_\epsilon(n) &= S(n); &(2.44)\\
Y_\epsilon(n) &= (1-\epsilon)S(n) + \epsilon\eta_Y(n); &(2.45)
\end{aligned}
$$

where $S(n)$ is the common source generated by an autoregressive model, order $p_4 = 5$, α_i and $\eta_S(n)$ are generated from a Gaussian distribution with mean 0 and standard deviation 1. $S(n)$ is measured on both sensors $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$. $\mathbf{Y}_\epsilon$ is further corrupted by independent Gaussian noise $\eta_Y(n)$ with 0 mean and unit variance. ϵ controls the signal to noise ratio (SNR) in $\mathbf{Y}_\epsilon$ and is varied from 0.1 to 0.9 with a step size of 0.1, corresponding to SNR in the range of $-19 \sim 19$ dB. The DI value and GC value averaged across 100 simulations for changing ϵ for a single source model are shown in Figure 2.7. The estimated order of the model is 5. In addition, the false positive rate using both DI and Granger causality with increasing ϵ is also calculated. We observe that the information flow in two directions using DI are less than the threshold for all values of ϵ, which indicates the acceptance of the null hypothesis that there is no significant causal information flow from $\mathbf{X}_\epsilon$ to $\mathbf{Y}_\epsilon$ or $\mathbf{Y}_\epsilon$ to $\mathbf{X}_\epsilon$. Note that DI is normalized by the mutual information. For a common source model, the instantaneous information exchange between $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ contributes mostly to the mutual information between $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$. Thus according to equation (2.32), $DI(DX_\epsilon^N \rightarrow Y_\epsilon^N)$ and $DI(DY_\epsilon^N \rightarrow X_\epsilon^N)$ normalized by mutual information are close to 0 and less than the threshold from the randomized data pairs. The false positive rate of DI is 0 for all ϵ. Therefore, DI is able to discriminate between instantaneous mixing from actual causality and is very robust to noise. For GC, when ϵ is small (< 0.2) or large (> 0.9), the value of GC is less than or very close to the threshold in both directions thus indicating that there is no causal information flow between the two processes. However, GC fails to accept the null hypothesis when ϵ is between 0.3 to 0.9 and detects a non-existing effective connectivity. GC reaches its maximum value when $\epsilon = 0.5$. This is due to the fact that GC is close to

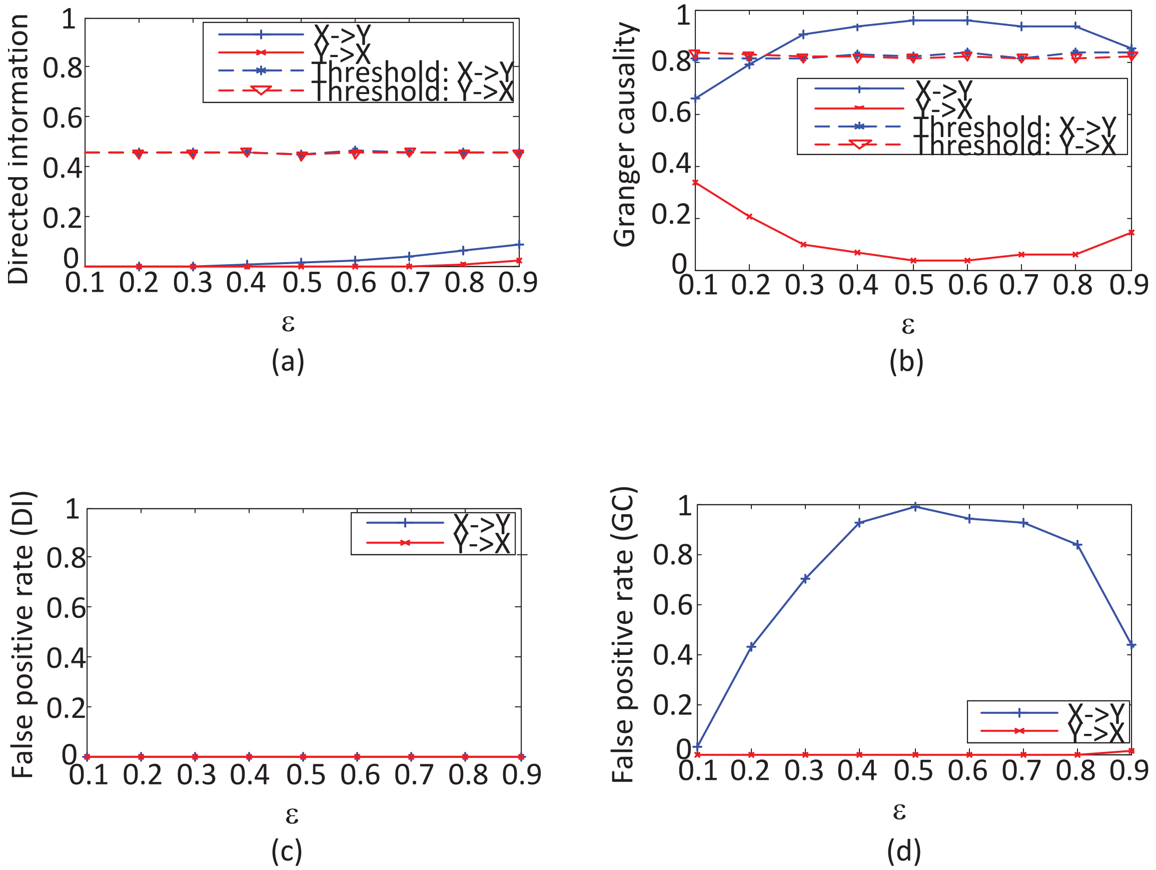
0 when two processes $\mathbf{X}_\epsilon$ and $\mathbf{Y}_\epsilon$ are independent or identical, i.e. when $\epsilon = 1$ and $\epsilon = 0$. Based on the definition of GC, the prediction of $\mathbf{Y}_\epsilon$ at the current time point will not be improved by taking into account the past samples of $\mathbf{X}_\epsilon$ for these processes [58]. Therefore, as ϵ increases from 0 to 0.5, $\mathbf{X}_\epsilon$ becomes the most different from $\mathbf{Y}_\epsilon$, therefore it can provide more new information about $\mathbf{Y}_\epsilon$ and the GC increases. As ϵ increases from 0.5 to 1, $\mathbf{X}_\epsilon$ becomes independent of $\mathbf{Y}_\epsilon$, and the GC decreases. The false positive rate of GC is not equal to 0 for all values of ϵ, which indicates that it has lower specificity compared to DI. Therefore, GC is not robust to the effect of a common source and may infer false positive effective connectivity. This simulation indicates that DI is more sensitive and discriminative about the information flow patterns in the presence of volume conduction, which means it is a more promising method to capture the effective connectivity for real EEG data.

Example 5*: Nonlinear dynamic system*

In this example, we illustrate the applicability of DI to coupled Lorenz oscillators with a certain delay. The Lorenz oscillator is a three-dimensional dynamic system that exhibits chaotic behavior. Synchronization of two Lorenz systems has been widely investigated for the analysis of EEG data, because the dynamic interactions related to the behavior of the cortex can be exemplified by these coupled systems [70]. In the following, we examined two asymmetric coupled Lorenz oscillators (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2) as follows [71],

$$\begin{aligned}
\dot{X}_1(t) &= -A(X_1(t) - Y_1(t)), && (2.46)\\
\dot{Y}_1(t) &= RX_1(t) - Y_1(t) - X_1(t)Z_1(t), && (2.47)\\
\dot{Z}_1(t) &= X_1(t)Y_1(t) - BZ_1(t), && (2.48)\\
\dot{X}_2(t) &= -A(X_2(t) - Y_2(t)) + \beta X_1(t - t_p), && (2.49)\\
\dot{Y}_2(t) &= RX_2(t) - Y_2(t) - X_2(t)Z_2(t), && (2.50)\\
\dot{Z}_2(t) &= X_2(t)Y_2(t) - BZ_2(t), && (2.51)
\end{aligned}$$

where each equation is a first-order differential equation. $A = 10$, $R = 28$, $B = \frac{8}{3}$, and $t_p = 0.02$ represents the time delay between two coupled components of these two oscillators,



(a)

(b)

(c)

(d)

Figure 2.7: Application of directed information and Granger causality to single source model. (a) Directed information with different ϵ for the single source model. (b) Granger causality with different ϵ for the single source model. (c) False positive rate for directed information with different ϵ for the single source model. (d) False positive rate for Granger causality with different ϵ for the single source model.

i.e. $\mathbf{X_1}$ and $\mathbf{X_2}$. β corresponds to the coupling strength and is varied from 0.1 to 1 with a step size of 0.2. The differential equations are numerically integrated with a time step of 0.01 using Euler's method [72], corresponding to a delay of 2 time samples between $\mathbf{X_1}$ and $\mathbf{X_2}$. The initial conditions of these six components are randomly generated from a Gaussian distribution with zero mean and unit variance. We generate 100 samples and the first 90 samples are discarded to eliminate the initial transients. We compute the information flow in two directions over 10 time points and the significance of the obtained DI value is verified by trial shuffling. The DI values and GC values between $\mathbf{X_1}$ and $\mathbf{X_2}$ of two asymmetric coupled Lorenz systems are computed with coupling strength β being set from 0.1 to 1. The estimated order of the model is 3. Though this is larger than the actual model order, our method will not lose any information except for the increased computational complexity. The results are shown in Figure 2.8. The results show that DI values from $\mathbf{X_1}$ to $\mathbf{X_2}$ increase with the coupling strength β and are significant for all values of β. In addition, there is no significant causal information flow from $\mathbf{X_2}$ to $\mathbf{X_1}$. Therefore, DI can effectively detect the causality in a nonlinear dynamic system. On the contrary, GC can not detect any significant information flow for all β values. It is due to the fact that the model selected for implementing GC is not consistent with the dynamic characteristics of the system.

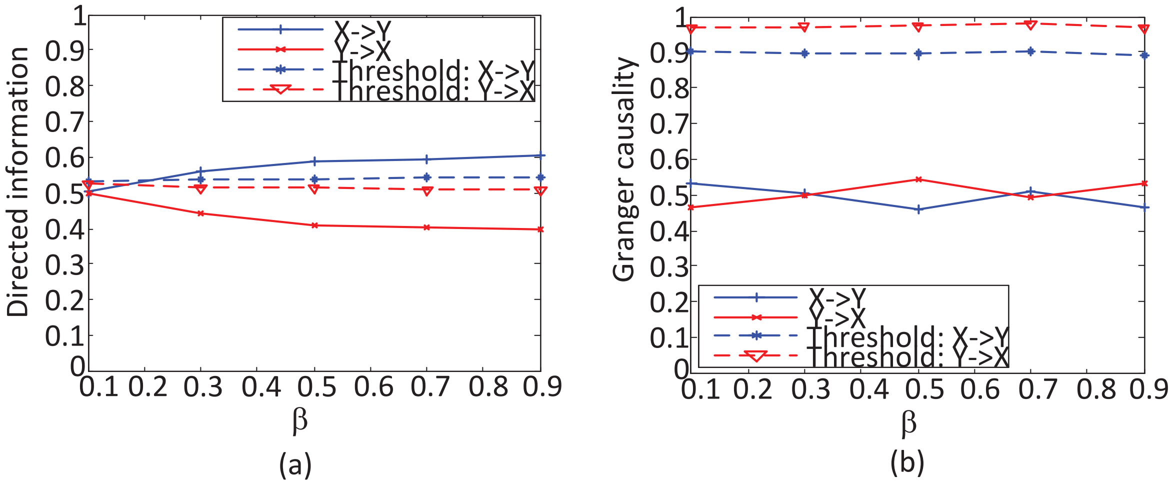


Figure 2.8: Application of directed information and Granger causality to two asymmetric coupled Lorenz oscillators. (a) Directed information with different β. (b) Granger causality with different β.

2.3.2 Biological data

In this subsection, we examine EEG data from ten undergraduates at Michigan State University drawn from an ongoing study of relationships between the error-related negativity (ERN) and individual differences[1] such as worry and anxiety. ERN is a brain potential response that occurs following performance errors in a speeded reaction time task [48]. All EEG data are collected as described in Chapter 1. Once the data are obtained, for each

[1] Participants for the present analysis were drawn from samples reported on in [46, 47].

subject, the EEG data are preprocessed by the spherical spline current source density (CS-D) waveforms to sharpen event-related potential (ERP) scalp topographies and eliminate volume conduction [73]. In addition, a bandpass filter is used to obtain signals in the theta band. In this study we focus on 33 electrodes corresponding to the frontal, central and parietal regions of the brain. For each pair of 33 electrodes $\mathbf{X}$ and $\mathbf{Y}$ for each subject, the effective connectivity is quantified by computing the modified DI over 70 trials and a model order of p in the theta band. The model order or the length of the time window p is determined by the Cao Criterion. We also apply Granger causality to the same data and compare its performance with directed information.

Previous work indicates that there is increased information flow associated with ERN for the theta frequency band ($4-8$ Hz) and ERN time window $25-75$ ms for Error responses compared to correct responses in particular between mPFC and lPFC regions [74]. In addition, Cavanagh *et al.* have shown that there is increased synchronization for error trials between electrode pairs, such as FCz-F5 and FCz-F6, compared to the synchrony between FCz-CP3 and FCz-CP4 [75]. The DI and GC values for each pair of electrodes averaged over 10 subjects are computed over a time window of 53 time points (100ms). The estimated order of the model for each electrode pairs is 3. In order to control the error rates for multiple hypothesis testing for all pairs of electrodes, the method proposed by Genovese *et al.* is used in this dissertation [76]. To implement this procedure, for two electrodes with time series $\mathbf{X}$ and $\mathbf{Y}$, we first shuffle the order of the trials of $\mathbf{X}$ 100 times to generate new observations $\mathbf{X}_m^*$, $m = 1, \cdots, 100$. The P-value of $DI(\mathbf{X} \to \mathbf{Y})$ is obtained by comparing it with DI values from randomized pairs of data $DI(\mathbf{X}_m^* \to \mathbf{Y})$, $m = 1, \cdots, 100$. We then obtain the threshold P_r for all P-values ($33 \times 33 \times 10$) by controlling the FDR bound q as 0.05. For $DI(\mathbf{X} \to \mathbf{Y})$, if the P-value is less than P_r, then the directed information flow from $\mathbf{X}$ to $\mathbf{Y}$ is significant; otherwise, it is not significant. Electrode pairs between which the information flow is significant in at least one of the ten subjects are shown in Figure 2.9(b). We also test the significance of Granger causality in the same way. When the FDR is controlled at 0.05,

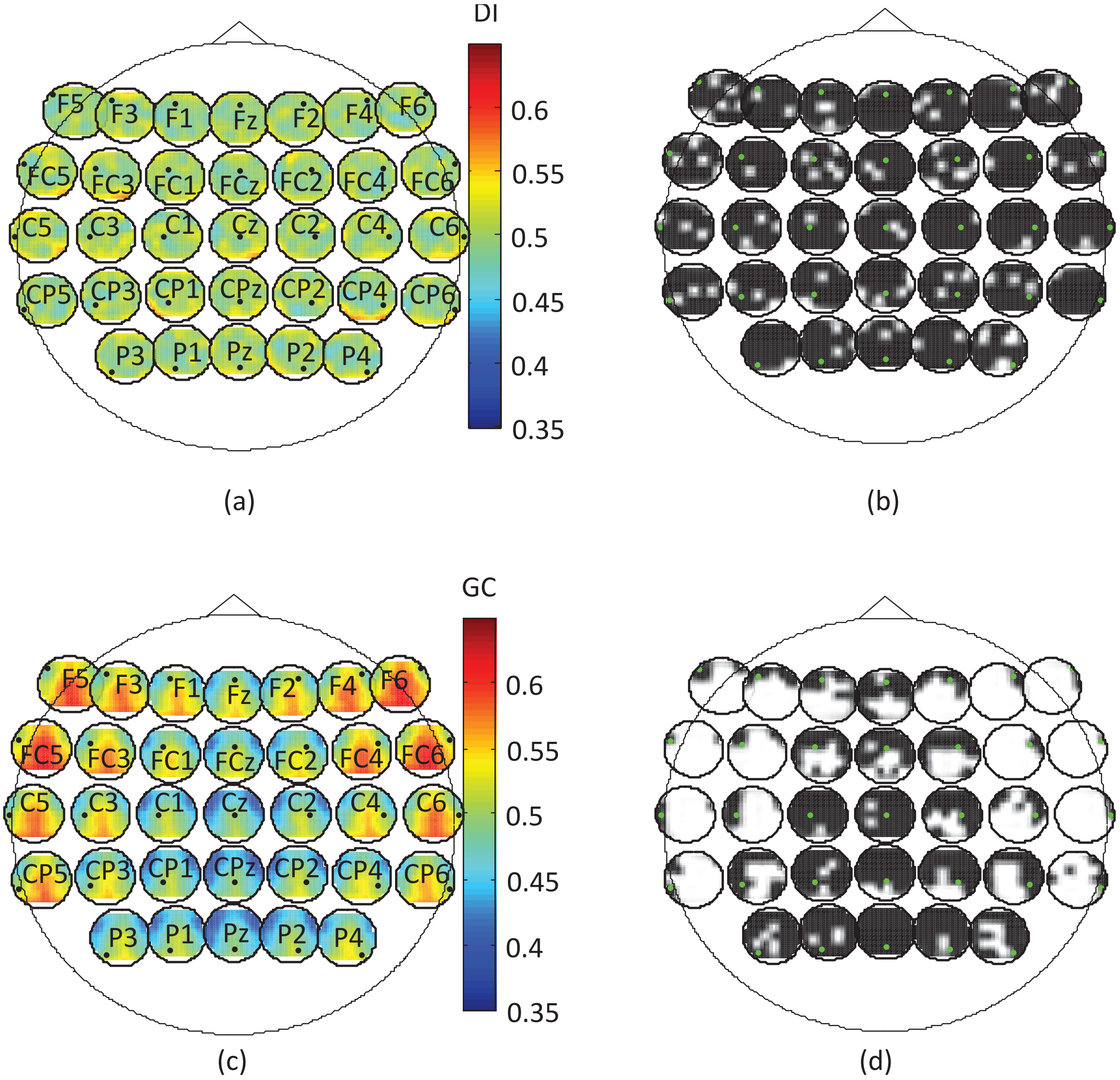


Figure 2.9: Application of directed information and Granger causality to EEG data. (a) Pairwise directed information. (b) Electrode pairs with significant DI values. (c) Pairwise Granger causality. (d) Electrode pairs with significant GC values. For (b) and (d), green dots indicate the location of the particular node and white regions correspond to significant information flow from that particular electrode to other electrodes. The name of each particular node in (b) and (d) is identical to the name in (a) and (c). The details of the significant electrode interactions are shown in Table 2.4.

the information flow between electrode pairs is significant if the P-value of DI or GC is less than 0.01. Electrode pairs that have significant causality relationship using both measures are shown in Figure 2.9. In Figure 2.9(a) and Figure 2.9(c), each small circle shows the

Table 2.4: Electrode pairs in the region of interest with significant DI values

From	To	From	To
F5	F1 FC2 CPz CP4 P3	C5	F6 FC5 Cz CP4
F3	FC3 CP4	C3	FC2 C6 P1
F1	C1 Cz Pz	C1	FC1 C6
FZ	F5	CZ	F5 C2 CP4
F2	FC3 FC6 C5 CP1	C2	FC6
F4	F6 C4	C4	P2
F6	F2 FC3 FCz Cz	C6	Pz
FC5	Fz C3 C2 CP6	CP5	Cz C4 CP3
FC3	CP1	CP3	C5 CPz P4
FC1	F4 FC3 C2 CP1 CP4	CP1	F6 FCz P3
FCZ	C3 CP1	CPz	FC6 C6 CP5 CP4 P1
FC2	F3 C1 C6 CP2 CP4 P3	CP2	F6 FCz FC4 CP1
FC4	C5	CP4	FC5 FCz C4
FC6	C5 C4 CP1	CP6	F5
P3	P4		
P1	F2 C6 CP2		
Pz	F5 F4 FCz		
P2	FC4 C5		
P4	F3 F4 FC3 FC2 FC4 Pz P2		

directed information and Granger causality from a particular electrode to other electrodes. In Figure 2.9(b) and Figure 2.9(d), each small circle shows electrode pairs that have significant causality relationship. The details of the significant electrode interactions are also shown in Table 2.4. The results indicate that DI detects strong information flow from the frontal region (e.g. F5, F6) to the frontal-central region (e.g. FC2, FCz) corresponding to the lateral prefrontal cortex (lPFC) and medial prefrontal cortex (mPFC). In addition, the central-parietal region (e.g. CPz, CP1, CP2) around the midline, corresponding to the motor cortex, has strong influence on the central and frontal regions (e.g. FCz, F6) since this is a speeded response task involving the motor cortex. These results are aligned with the previous work in [75], which shows that error processing is controlled by the communication between the lateral prefrontal cortex and medial prefrontal cortex. When GC is applied to the same data, the information flow pattern around the midline is similar to the DI. However, the information flow from the lateral prefrontal cortex to the rest of the brain is significant. On one hand, the similar patterns of connectivity using both measures verify the validity of proposed DI computation algorithm. On the other hand, GC shows significance over a wide region of the brain especially in the lateral areas compared to DI, which may be due to GC's low specificity to volume conduction in the form of a common source. Previous work and our simulation in Example 4 have indicated that Granger causality based measures may infer erroneous effective connectivity in the case of the common source as seen in EEG data [69, 57]. However, without ground truth, we cannot confirm that some links reported as significant by GC are spurious and due to volume conduction in a conclusive manner, but the results from DI agree more with the suggestions in [75], that most of the increase in connectivity during cognitive control, i.e. ERN, should be between medial prefrontal cortex and lateral prefrontal cortex, compared to the results of GC. Therefore, DI is more sensitive and discriminative about the information flow patterns compared to GC for real neurophysiological data.

2.4 Estimation of directed information

2.4.1 Estimation based on entropy estimation

According to equation (2.1), for DI estimation, the entropy estimator should be applied to both marginal and joint entropy estimation. In this subsection, we will review three commonly used entropy estimators and comment on their applicability to DI estimation.

Plug-in estimator

The entropy $H(f)$ for a continuous probability density function $f(x)$ is given by:

$$H(f) = -\int f(x) \log f(x) dx \tag{2.52}$$

The plug-in estimates of entropy are based on a consistent estimation f_N of f and the estimation of f_N depends on N realizations $(x_1, \cdots, x_N)$ of X. Ordinary histogram is the most commonly used density estimation method. First the minimum to maximum of N realizations of X is divided into bins, where $B_k = [t_k, t_{k+1})$ denotes the k-th bin and $h = t_{k+1} - t_k$, then the approximation of the pdf is given by [77]:

$$\hat{f}(x) = \frac{\nu_k}{Nh} \text{ for } t_k \leq x < t_{k+1} \tag{2.53}$$

where ν_k is the number of data points that fall in the k-th bin. The accuracy of this method depends on both the proper bin size and sample size. When the samples are barely sufficient, it has a large bias, which becomes worse as the dimensionality of the observed space increases.

An alternative to the histogram is Kernel density estimator (KDE) written as [77]:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{x - x_i}{h}) \tag{2.54}$$

where K is the kernel function and h is the width of the kernel (the smoothing parameter). Some common kernel types include Uniform, Epanechnikov, Gaussian, and Laplacian Kernels. The quality of KDE depends on the choice of the smoothing parameter and the choice of kernel. Although KDE becomes time consuming for joint entropy estimation, it can still give an accurate estimate of DI if proper parameters are chosen.

M-spacing estimator

M-spacing estimator is a nonparametric estimator which estimates the entropy directly from i.i.d data samples without approximating the pdf. Consider samples of a scalar random variable $X = (x_1, \cdots, x_N)$ rearranged in non-decreasing order $x_{(1)} \leq \cdots \leq x_{(N)}$. *M*-spacing, is then defined to be $x_{(i+m)} - x_{(i)}$, for $1 \leq i < i + m \leq N$. If m is a function of N, one may define the m_N-spacing as $x_{(i+m_N)} - x_{(i)}$. The m_N-spacing entropy estimator can be defined as [63]:

$$\hat{H}_N(x_1, \cdots, x_N) = \frac{1}{N} \sum_{i=1}^{N-m_N} \log(\frac{N}{m_N}(x_{(i+m_N)} - x_{(i)})). \tag{2.55}$$

The m-spacing estimators of entropy are based on the intuition that sums of small random intervals have consistent behavior. In order to estimate the joint entropies of multiple random variables, multi-dimensional spaces with constant expected probability mass are generated by constructing a Voronoi or Delaunay region [63]. Then the following estimator is used:

$$\hat{H}_{Hyper} = \sum_{i=1}^{m} \frac{C(U^i)}{N} \log \frac{NA(U^i)}{C(U^i)} \tag{2.56}$$

where $C(U^i)$ is the number of (finite volume) Voronoi regions in a Hyper-Region U^i, $N = \sum_i C(U^i)$, $A(U^i)$ is the d-dimensional volume of Voronoi region U^i. The calculation of volumes needed for equation (2.56) is exponential in the dimension. According to equation (2.1), estimation of $H(X_{k-d}X_{k-d+1}Y_kY_{k+1})$ which is a 4-dimensional problem makes the calculation of volumes complicated and slow in DI estimation.

Kozachenko and Leonenko (KL) estimator

Nearest neighbor (NN) distances based entropy estimators was introduced by Kozachenko and Leonenko, also known as KL estimator [78]. For a random variable X with N observations, $X = (x_1, \cdots, x_N)$, the distance of each point (observation) x_i to any other points x_j, $j = 1, \cdots, N, j \neq i$, defined as $d_{i,j} = \|x_i - x_j\|$, is found. Then the distance to the $N - 1$ neighbors of each point are ranked: $d_{i,j_1} \leq \cdots \leq d_{i,j_{N-1}}$. $H(X)$ can be estimated as the average distance to the k-nearest neighbors, averaged over all x_i. The KL entropy estimator is defined as:

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log(c_d) + \frac{d}{N} \sum_{i=1} N \log \epsilon(i) \tag{2.57}$$

where $\psi(x)$ is the digamma function, d is the dimension of X and c_d is the volume of the d-dimensional unit ball, $\epsilon(i)$ is twice the distance from x_i to its kth neighbor. The algorithm spends most of its time in spatial queries, which is unacceptable for DI estimation because the dimension of the point x_i can be larger than four.

As it can be seen from the above discussions, the nonparametric entropy estimators are either complex or time consuming when applied to high dimensional data. Therefore, methods with less complexity and high efficiency are needed for estimating DI.

Universal estimator

Recently, Zhao *et al.* proposed an universal algorithm to estimate directed information for stationary ergodic processes by using sequential probability assignment and context tree weighting [79]. For this estimator, DI is obtained by estimating $H(\mathbf{Y})$ and $H(\mathbf{Y}||\mathbf{X})$ separately and $DI(\mathbf{X} \to \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}||\mathbf{X})$. Given the realizations of $\mathbf{X}$ and $\mathbf{Y}$ and the universal source code C_n^Y on $\mathbf{Y}$ and $C_n^{X,Y}$ on $(\mathbf{X}, \mathbf{Y})$, the sequential probability assignments Q_Y and $Q_{X,Y}$ induced by C_n^Y and $C_n^{X,Y}$ are used to calculate the estimate of $H(\mathbf{Y})$ and $H(\mathbf{Y}||\mathbf{X})$. Zhao *et al.* employed the context tree weighting as the universal source coding scheme and the universal probability assignments can be constructed from a universal coding scheme [79]. This algorithm requires both the realizations of the signal and an universal source code scheme with low complexity and fast convergence rates. In addition, the original context tree weighting is for binary sequences, and has to be extended for discrete signals with continuous values and multiple realizations.

2.4.2 Estimation based on mutual information and multi-information

DI can also be expressed in terms of mutual information and multi-information, which require the estimation of the common information between two length N random vectors ($I(X^N; Y^N)$) or among multiple one-dimensional random variables ($I(X^1, \cdots, X^N, Y^1, \cdots, Y^N)$). Estimators based on adaptive partitioning, which do not require parameter selection, region construction and NN-search, are efficient for high dimensional data and are used in both methods for estimating DI.

Mutual information estimation

According to equation (2.2), directed information can be written in terms of mutual information. The most straightforward approach for estimating MI is partitioning the supports of X and Y into bins of finite size, and approximating equation (1.1) by the finite sum:

$$I(X;Y) \approx I_{binned}(X;Y) \equiv \sum_{i,j} p(i,j) \log \frac{p(i,j)}{p_x(i)p_y(j)} \tag{2.58}$$

$I_{binned}(X;Y)$ is obtained by counting the number of samples falling into the various bins, which is similar to the histogram. If we let $N \to \infty$ then all bin sizes tend to zero, and the right hand side of Equation (2.58) will converge to $I(X;Y)$. If the distributions are fractal, this convergence might no longer be true.

Darbellay presented a data-dependent nonparametric estimator of the mutual information based on an adaptive partitioning of the observed space [61]. The basic concept of the method is to build a succession of finer partitions of the high-dimensional observed space and stop the refinement process on any hyperrectangle when the local independence has been achieved.

Multi-information estimator

In equation (2.3), we introduce an alternative representation of DI in terms of multi-information, which can be estimated by extending the adaptive data-dependent partitioning method to multiple random variables [42]. The procedure for multi-information estimation is as follows:

- $X_1, \cdots, X_d$ are d one-dimensional random variables with N observations $(x_i(1), \cdots, x_i(j), \cdots, x_i(N))$, $i = 1, \cdots, d$, $j = 1, \cdots, N$. First, N observations of each random variable are rank ordered separately. $z_i(j)$ is the rank of $x_i(j)$ with respect to the other $N-1$ samples from the same random variable, $z_i(j) \in \{1, \cdots, N\}$. Then the estimation of multi-information among $X_1, \cdots, X_d$ has turned into the estimation of multi-information among $Z_1, \cdots, Z_d$, which is the ranked sample space of $X_1, \cdots, X_d$.

- In the $(Z_1, \cdots, Z_d)$ space, a dyadic partitioning of the space is iteratively done until the sample distribution of each hypercube is conditionally independent. Given N samples of $Z = (Z_1, \cdots, Z_d) \in \mathbb{Z}^d$, let $\mathbb{Z}^d$ be the initial one-cell partition. Then every cell is to be partitioned by marginal equiquantization and not partitioned further unless it achieves conditional independence. For example, for a cell C, each edge (margins) of the d-dimensional cell is divided into α ($\alpha = 2$ in general) intervals with approximately the same number of the points in each marginal subintervals to obtain α^d subcells C_k, $k = 1, \cdots, \alpha^d$. If $\alpha = 2$, the partition point is the midpoint of each edge in the Z space. The lower (L) and upper (U) bounds of each dimension of the subcell C_k are $z_i^k(L)$ and $z_i^k(U)$, i is the ith dimension of C_k, $i = 1, \cdots, d$. The total number of points in cell C_k and the marginal number of points in each dimension are:

$$\begin{aligned}
N_Z(C_k) &= \text{number of points } z \text{ such that} \\
&\qquad z_i^k(L) < z_i < z_i^k(U) \text{ for all } i = 1, \cdots, d \\
N_{Z_1}(C_k) &= \text{number of points } z \text{ such that } z_1^k(L) < z_1 < z_1^k(U) \\
&\vdots \\
N_{Z_d}(C_k) &= \text{number of points } z \text{ such that } z_d^k(L) < z_d < z_d^k(U)
\end{aligned}$$

The χ^2 goodness-of-fit test at the 3% significance level is applied to all subcells to test the local independence of cell C, that is,

$$\sum_{k=1}^{\alpha^d} \frac{(N_{C_k} - N_C/\alpha^d)^2}{N_C/\alpha^d} \leq \chi^2_{0.97}(\alpha^d - 1), \tag{2.59}$$

where N_C is the number of points in cell C, and N_{C_k} is the number of points in subcell C_k, $k = 1, \cdots, \alpha^d$, respectively. If the condition in equation (2.59) is fulfilled, the hypothesis of conditional independence is accepted and the cell C is not subjected to further partitioning.

- Once the conditional independence has been achieved, it can be shown that:

$$\begin{aligned}
\hat{I}(Z_1, \ldots, Z_d) &= \sum_{k=1}^{\alpha^d} P_{Z_1,\cdots,Z_d}(C_k) \log \frac{P_{Z_1,\cdots,Z_d}(C_k)}{P_{Z_1}(C_k), \cdots P_{Z_d}(C_k)} \\
&= \frac{1}{N} \sum_{k=1}^{\alpha^d} N_Z(C_k) \log \frac{N_Z(C_k)}{N_{Z_1}(C_k) \cdots N_{Z_d}(C_k)} + (d-1) \log N,
\end{aligned} \tag{2.60}$$

where $P_{Z_1,\cdots,Z_d}(C_k)$ is the probability that the N d-dimensional points $(z_1, \cdots, z_d)$ falls into each hypercube C_k, $P_{Z_1,\cdots,Z_d}(C_k) = N_Z(C_k)/N$, and $P_{Z_1}(C_k), \cdots, P_{Z_d}(C_k)$ are the corresponding marginal probabilities, $P_{Z_i}(C_k) = N_{Z_i}(C_k)/N$.

Though the mutual information and multi-information offer two different ways to express DI, the procedure for the two estimation methods are very similar. However, according to equation (2.3), DI estimation based on multi-information has one more term $I(Y^N)$ to estimate, which makes it computationally slightly more expensive.

2.4.3 Performance comparison of estimators

In this section, we compare different directed information estimation methods using simulated data to verify their effectiveness. In order to reduce the computational complexity for each simulation, DI for every successive two time points are computed.

In the first example, we consider a first order linear autoregressive (AR) model defined as:

$$\begin{aligned} X_i &= 0.5 \times X_{i-1} + u_i; \\ Y_i &= 0.2 \times Y_{i-1} + 0.8 \times X_{i-1} + v_i, \end{aligned} \tag{2.61}$$

where u_i and v_i are white Gaussian random processes with standard deviation of 0.3 and $\mathbf{X}$ and $\mathbf{Y}$ have zero initial conditions. We generate 128 realizations of $\mathbf{X}$ and $\mathbf{Y}$ and compute the DI over 20 time samples. In order to evaluate the bias and variance of each estimator, the same model is replicated 100 times. From equation (2.61), it is obvious that $\mathbf{X}$ is the driver of $\mathbf{Y}$, so we expect the directed information from $\mathbf{X}$ to $\mathbf{Y}$ to be greater than from $\mathbf{Y}$ to $\mathbf{X}$. First, the averaged DI for each estimator at two successive time points is shown in Figure. 2.10.

Second, in order to compare the performance of different methods and verify the advantages of information based estimators, we take the average DI value along the whole time sequence and evaluate the bias, variance, computational efficiency and discrimination power of each estimator. The bias of each method can be obtained by comparing with using covariance matrices, which has the closest result to the actual DI value since $\mathbf{X}$ and $\mathbf{Y}$ are normally distributed in this simulation. The computational efficiency quantified by the average run time of each method and the discrimination power (DP) quantified by the difference of mean

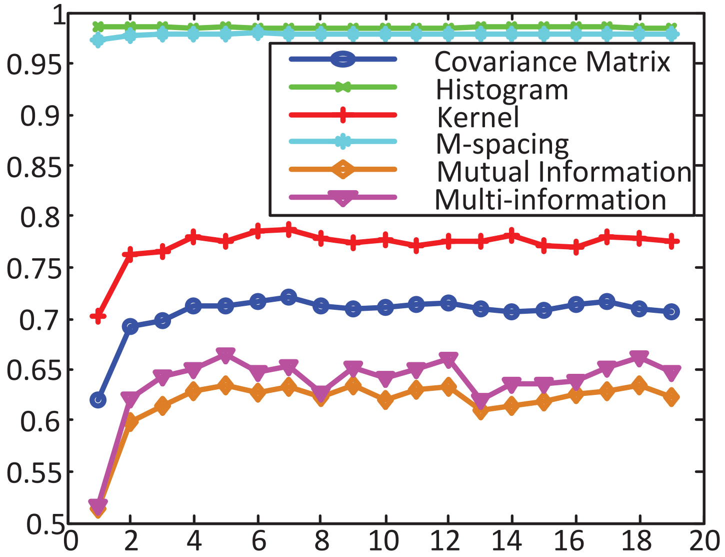


Figure 2.10: Averaged information flow for linear model using different estimators(128 realizations)

between the DIs in the two directions normalized by the standard errors defined as:

$$DP = \frac{\overline{DI(\mathbf{X} \to \mathbf{Y})} - \overline{DI(\mathbf{Y} \to \mathbf{X})}}{\sqrt{\frac{S^2_{DI(\mathbf{X}\to\mathbf{Y})} + S^2_{DI(\mathbf{Y}\to\mathbf{X})}}{N}}}, \tag{2.62}$$

where $\overline{DI(\mathbf{X} \to \mathbf{Y})} = \frac{1}{N-1}\sum_{k=1}^{N-1} DI_k(\mathbf{X} \to \mathbf{Y})$ and $S_{DI(\mathbf{X}\to\mathbf{Y})}$ is the standard deviation of the DI values along the time sequence, $DI_k(\mathbf{X} \to \mathbf{Y})$ is the directed information value from $\mathbf{X}$ to $\mathbf{Y}$ at a small time window $k \sim k+1$. In this simulation we use Gaussian kernel, whose width is chosen by likelihood cross-validation (LCV) for KDE. As it can be seen from Figure. 2.10 and Table 2.5, the histogram and m-spacing estimator have higher bias compared to the other methods. The histogram is very dependent on the number of bins, i.e., if we use a smaller bin size such as 5, it will have a lower bias. Kernel method has a slightly upward bias while MI and multi-information based estimation methods have downward bias, but the absolute difference is nearly the same. However, the latter ones are faster than the Kernel method.

We also increase the number of trials to show the dependency of each estimator on the number of data samples. The result for 1024 realizations is shown in Table 2.6. When

Table 2.5: Performance Comparison for linear model (128 realizations)

	Mean	Variance($\times 10^{-3}$)	Time(s)	DP
Covariance	0.7058	0.1264	0.0155	15.0449
Histogram	0.9856	0.0007	6.4049	2.7736
Kernel	0.7724	0.1607	8.6807	7.9381
M-spacing	0.9790	0.0006	55.6460	11.4727
MI	0.6185	0.4705	2.7546	5.3127
Multi-information	0.6382	0.7489	2.7355	4.8684

the number of trials increases, the bias of all of the estimators decrease. Moreover, the MI and multi-information based methods outperform Kernel methods in bias and get very close to the performance of the Covariance Matrix method without any prior knowledge of the distribution of the data.

Table 2.6: Performance Comparison for linear model (1024 realizations)

	Mean	Variance($\times 10^{-3}$)	Time(s)	DP
Covariance	0.7009	0.0000	0.0113	19.9898
Histogram	0.9033	0.0000	11.6406	12.4591
Kernel	0.7622	0.0000	195.1709	22.6683
M-spacing	0.9809	0.0000	613.2450	27.2449
MI	0.7035	0.0000	8.0961	12.5346
Multi-information	0.7096	0.1000	7.6563	11.3131

In the second example, we consider a first order nonlinear autoregressive model given below and compare the proposed method with the regular histogram estimation for every two pairs of **X** and **Y**.

$$\begin{aligned} X_i &= 0.5 \times X_{i-1} + u_i; \\ Y_i &= 0.2 \times Y_{i-1} + 0.8 \times X_{i-1}^2 + v_i, \end{aligned} \tag{2.63}$$

where u_i and v_i are distributed as in the first example. The performance of each estimator for 128 and 1024 realizations of $\mathbf{X}$ and $\mathbf{Y}$ are shown in Tables 2.7 and 2.8. We can observe that the MI and multi-information based methods run much faster than the others. We should also note that, the multi-information based method has much stronger discrimination power than MI estimator, though the discrimination power is very low when the number of trials is very small, which makes it hard to infer the causality between $\mathbf{X}$ and $\mathbf{Y}$ accurately. However, the discrimination power becomes stronger when the number of trials increases to 1024.

Table 2.7: Performance Comparison for nonlinear model (128 realizations)

	Mean	Variance($\times 10^{-3}$)	Time(s)	DP
Histogram	0.9842	0.0018	2.5665	1.3349
Kernel	0.5503	0.6849	3.5845	2.9192
M-spacing	0.9631	0.0012	30.4351	4.1032
MI	0.3126	1.2990	0.7148	0.0802
Multi-information	0.3184	1.7980	0.6970	0.6136

Table 2.8: Performance Comparison for nonlinear model (1024 realizations)

	Mean	Variance($\times 10^{-3}$)	Time(s)	DP
Histogram	0.8633	0.0640	11.1156	3.0179
Kernel	0.5368	0.0809	191.9125	11.2326
M-spacing	0.9676	0.0001	602.7246	13.7146
MI	0.3507	0.2887	5.5128	2.1319
Multi-information	0.3494	0.5902	5.2984	4.0577

2.5 Conclusions

In this chapter, we presented the time-lagged directed information and modified directed information to reduce the computational complexity of computing DI while still quantifying the causal dependencies. These simplified measures are derived for stationary statistical processes with limited order and it is proven that the rate of modified DI is equal to the transfer entropy rate. The simulation results presented above indicate that the MDI measure is more suitable for the approximation of DI when the order of the model is unknown or when there are multiple time lags compared to TLDI. Even though the MDI is shown to be an upper bound for the actual DI, it achieves a better performance compared to TLDI in terms of bias and is comparable in terms of the computational complexity. Moreover, we also introduce a new directed information estimation method based on multi-information and provide a quantitative comparison of various DI estimation methods. Considering various factors including bias, variance, computational speed and discrimination power, the MI and multi-information based DI estimation methods have similar performance and are better than the others. Moreover, the multi-information based estimator outperforms MI estimator in discriminating nonlinear causal relationships. Finally, in order to illustrate the advantages of DI, we applied directed information measure to identify the causality relationships for both linear and nonlinear AR models, linear mixing models, single source models and Lorenz systems, and compare its performance with Granger causality. Directed information is shown to be more effective in detecting the causality of different systems compared to Granger causality. We also applied the directed information measure on EEG data from a study containing the error-related negativity to infer the information flow patterns between different regions. The results showed that the directed information measure can capture the effective connectivity in the brain between the mPFC and lPFC areas as predicted by previous work.

There are still remaining issues with the implementation of directed information. First, the performance of directed information relies on accurate estimation from limited sample sizes that introduces bias to the estimated values. This problem can be addressed by either using parametric density models or improving existing mutual information and entropy estimators. Recently, Zhao *et al.* proposed an universal algorithm to estimate directed information for stationary ergodic processes by using sequential probability assignment, which may be used to improve the effective connectivity results discussed in this dissertation [79]. Current applications of this algorithm are for binary sequences, therefore, it has to be extended for discrete signals with continuous values and multiple realizations. Second, the performance of directed information relies on the selection of the model order. If the order of the model is too small, it will lose the information from $\mathbf{X}$ to $\mathbf{Y}$. If it is too large, the computational complexity is very high. In addition to classical embedding dimension determination methods such as the Cao criterion used in this dissertation, Faes *et al.* proposed a sequential procedure to determine the embedding dimension of multivariate series [80]. This method is based on an information-theoretic technique and shows promising performances for various signal models, which may be extended to DI computation in the future. Third, directed information does not discriminate between direct and indirect interactions among multivariate time series. However, this is not a shortcoming of DI since DI does not assume any particular signal interaction model: bivariate or multivariate. Similar to other information theoretic measures, such as mutual information, whether the particular measure can identify interactions between multiple processes depends on how the measure is applied. For example, in the case of mutual information, though the original definition is for two random processes $\mathbf{X}$ and $\mathbf{Y}$, it is possible to extend it to multiple processes [60]. Similarly, we can apply DI over multiple processes using conditional directed information such as the definition given by Kramer. We address this issue in the next chapter by using conditional directed information and develop algorithms to infer the actual network. Similarly, GC originally is defined for two time series that a stochastic process $\mathbf{X}$ causing another process $\mathbf{Y}$ if the prediction of $\mathbf{Y}$ at the current time point, Y_n, is improved when taking into account the past samples of $\mathbf{X}$. However, in application it has been extended to multiple processes through the use of multivariate AR models, such as PDC. We also compare the performance of our algorithm based on conditional directed information with PDC in the next chapter.

Chapter 3
DIRECTED NETWORK INFERENCE BASED ON DIRECTED INFORMATION

3.1 Introduction

In many complex systems, such as the brain network, another interesting problem is to reveal the actual causal structure of the network, i.e. effective network inference, rather than quantifying pairwise causal relationships which is not sufficient to discriminate direct interactions from indirect interactions. Effective network inference algorithms can be categorized into three groups: pair-wise algorithms, equation-based algorithms and network-based algorithms [81], which are based on different measures to quantify the causality. The pair-wise algorithms try to find pairs of variables that are correlated and influence the behavior of each other. Cross-correlation and its extension in the frequency plane, i.e. coherence, are the most traditional measures to capture the causal relationships in neural networks, brain networks and so on [82, 83]. However, these approaches have two drawbacks: (1) they assume the linearity of the relationship between the variables, which is not always true, e.g. EEG signals are known to have nonlinear dependencies [17]; (2) they quantify the relationship between two variables without considering the effect of the third variable, which may generate false positive connections in a network. Equation-based algorithms use a model, such as multivariate autoregressive model (MVAR) and dynamic causal model, to relate the values of variables [44, 84]. For example, partial directed coherence (PDC) and direct transfer function (DTF), derived from MVAR, are widely used to determine the neural networks from multivariate recordings [44, 85]. However, these equation-based methods require a *priori* knowledge of the dynamics of the data generating systems (models) and sufficient time samples to build the model. The network-based algorithms are dedicated to finding the best network, such as Boolean networks or Bayesian network, to describe the observation-al data [81]. One of the representative algorithms, i.e., Dynamic Bayesian network (DBN) inference algorithm [86], was first developed to infer nonlinear transcriptional regulatory networks and was later used successfully at reconstructing nonlinear neural information flow networks [81, 87]. One major issue with this kind of algorithms is that the computational complexity to learn the structure of a network proves to be NP-hard [88]. Therefore, it can only be applied to relatively small networks [89, 90].

Recently, information-theoretic approaches have been widely used for the inference of large networks in the bioinformatics community [91]. Most of these methods rely on estimating the mutual information between variables from data to quantify the dependency. Unlike correlation-based algorithms, information-theoretic approaches can quantify the nonlinear dependencies [92, 60]. The Chow-Liu tree algorithm is the first to adopt the mutual information in probabilistic model design with minimum spanning tree, which has a low number of edges even for non-sparse target networks [91]. The Relevance network (RELNET) extends these ideas by determining the relevant connections such that a pair of nodes X and Y is connected when the mutual information is above a threshold. However, this method may infer false connections when two nodes are indirectly connected through an intermediate node [93]. The algorithm for the reconstruction of accurate cellular networks (ARACNE) addresses this problem by using the data processing inequality for mutual information [92]. Zhao *et al.* also address the same problem using conditional mutual information [60]. However, since mutual information is a symmetric quantity, all of these methods are limited to inferring undirected networks. Quinn *et al.* extend the Chow-Liu tree for random variables to the causal dependence tree for random processes, and show that the best causal dependence tree approximation is the one which maximizes the sum of directed information on its edges [94]. Similar to the Chow-Liu tree, this method also has a low number of edges since not all real networks follow the tree structure, and requires *a priori* knowledge of the root in some implementations. In this thesis, we propose a directed acyclic network inference algorithm based on estimating directed information between signals over time in order to quantify both the connectivity and causality in networks. Since we addressed the problem of computation and estimation of DI in the previous chapter, one major problem remains with the application of the directed information to infer the directed network, i.e., discriminating between direct and indirect connections in the network [95]. We propose time-lagged conditional directed information and modified conditional directed information based inference algorithms to address this problem.

The organization of this chapter is as follows. First, we introduce the concept of conditional directed information. Then we propose time-lagged conditional directed information and modified conditional directed information to reduce the computational complexity of this measure. Finally, three different network inference algorithms are given and the comparison with existing methods are shown through simulated network models.

3.2 Conditional directed information

3.2.1 Background

Based on the definition of directed information, if $\mathbf{X}$ is the direct cause of $\mathbf{Y}$, then the value of $DI(\mathbf{X} \to \mathbf{Y})$ is significant indicating causality between the two processes. However, a significantly large DI value does not guarantee that $\mathbf{X}$ directly causes $\mathbf{Y}$; $\mathbf{X}$ and $\mathbf{Y}$ may interact indirectly through other nodes [31]. Kramer extended the directed information to three variables using causal conditional directed information (CDI) $DI(X^N \to Y^N \parallel Z^N)$, which measures the information flow from X^N to Y^N when causally conditioned on sequence Z^N [27]:

$$\begin{aligned} DI(X^N \to Y^N \parallel Z^N) &= H(Y^N||Z^N) - H(Y^N||X^N Z^N) \\ &= \sum_{n=1}^{N}[H(Y_n|Y^{n-1}Z^n) - H(Y_n|Y^{n-1}X^n Z^n)] \\ &= \sum_{n=1}^{N} I(X^n; Y_n|Y^{n-1}Z^n), \end{aligned} \tag{3.1}$$

where $I(X;Y|Z)$ is the conditional mutual information between two random variables X and Y given Z. This definition differs from the conditional mutual information $I(X^N;Y^N|Z^N)$ only in that X^n and Z^n replace X^N and Z^N in each term on the right of equation (3.1). When $H(Y_n|Y^{n-1}Z^n) = H(Y_n|Y^{n-1}X^nZ^n)$, $DI(X^N \to Y^N \parallel Z^N) = 0$, i.e., given the observation of the third time series $\mathbf{Z}$ up to the current time point n, $\mathbf{X}$ does not provide any information about $\mathbf{Y}$.

3.2.2 Motivational example

Directed information can reveal the causal dependencies between two variables, but it can not distinguish between direct and indirect causality. Indirect interactions will cause false positive connections for network reconstruction. To address the problem, we first specify how conditional directed information can help in network inference problems using two simple motivational models. Then, time-lagged conditional directed information and modified conditional directed information are introduced to solve the computational complexity problem while still being able to remove the indirect connections.

First, consider the following two trivariate models: a hub and chain model. A trivariate autoregressive model following a hub pattern is as follows:

$$\begin{aligned} X_i &= u_i, \\ Y_i &= bX_{i-m} + v_i, \\ Z_i &= cX_{i-n} + w_i, \end{aligned} \tag{3.2}$$

where u_i, v_i and w_i are i.i.d processes following a Gaussian distribution $N(0, \sigma^2)$. In this model, nodes $\mathbf{Y}$ and $\mathbf{Z}$ are indirectly connected through a hub (parent) node $\mathbf{X}$, and $\mathbf{X}$ interacts with them with different time delays (Figure. 3.1). The trivariate model following a chain pattern is given as:

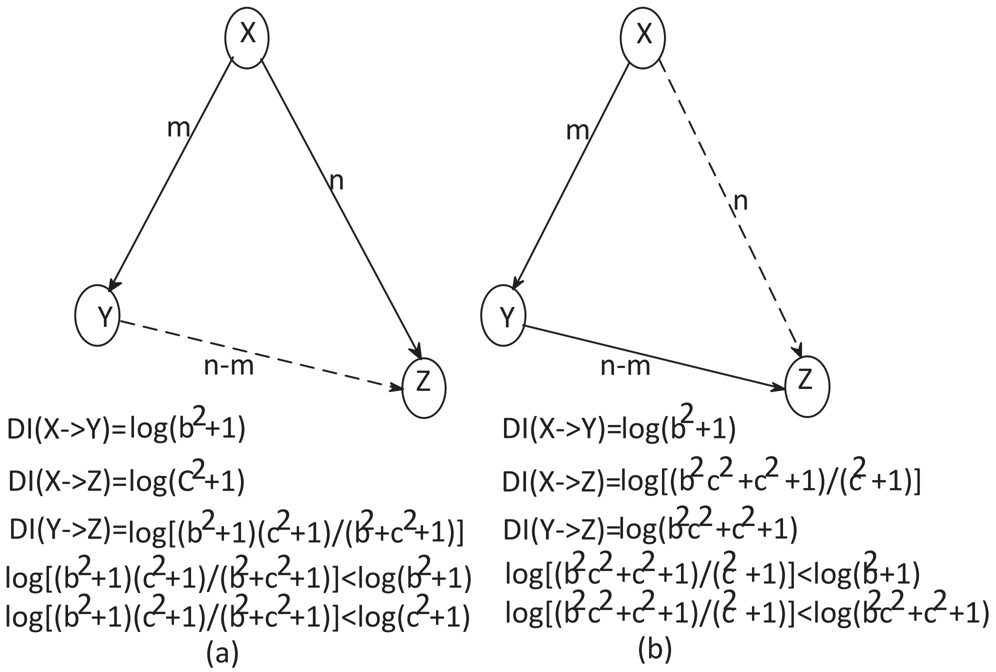
$$\begin{aligned} X_i &= u_i, \\ Y_i &= bX_{i-m} + v_i, \\ Z_i &= cY_{i-(n-m)} + w_i, \end{aligned} \tag{3.3}$$

where $n > m$, u_i, v_i and w_i are i.i.d processes following a Gaussian distribution $N(0, \sigma^2)$. In this model, **X** interacts with **Z** through intermediate (proxy) node **Y** .

For both models, the time lags between $\mathbf{X} \to \mathbf{Y}$, $\mathbf{X} \to \mathbf{Z}$ and $\mathbf{Y} \to \mathbf{Z}$ are m, n and $n - m$, respectively. The directed information rate between any two variables is larger than 0. The DI values and time lags between all of the pairs are shown in Figure 3.1. For network inference, if we determine there is a connection between two nodes when the DI value is larger than 0, then the inferred network for both hub and chain models will be the same. Therefore, only employing DI, it is hard to discriminate between the direct and indirect connections which may cause incorrect network inference. We introduce conditional directed information to remove the indirect connections. For the hub case, there is no information flow from **Y** to **Z** directly and $DI(Y^N \to Z^N || X^N) = 0$. While for the chain case, **X** interacts with **Z** through intermediate (proxy) node Y and $DI(X^N \to Z^N || Y^N) = 0$ (See Appendix A for details). Thus, CDI helps to discriminate between the two connection patterns.

3.2.3 Computation of conditional directed information

Similar to directed information, the computational complexity of computing conditional directed information also increases with the signal length. Thus, we proposed a simplified approach to compute CDI, while still reflecting the actual information flow from **X** to **Y** influenced by **Z**. In order to reduce the computational complexity, in practice, estimation of CDI is limited to every two time samples of three processes **X**, **Y** and **Z**. We can define the



X
m
n
Y
n-m
Z
$DI(X->Y)=\log(b^2+1)$
$DI(X->Z)=\log(C^2+1)$
$DI(Y->Z)=\log[(b^2+1)(c^2+1)/(b^2+c^2+1)]$
$\log[(b^2+1)(c^2+1)/(b^2+c^2+1)]<\log(b^2+1)$
$\log[(b^2+1)(c^2+1)/(b^2+c^2+1)]<\log(c^2+1)$
(a)
X
m
n
Y
n-m
Z
$DI(X->Y)=\log(b^2+1)$
$DI(X->Z)=\log[(b^2c^2+c^2+1)/(c^2+1)]$
$DI(Y->Z)=\log(b^2c^2+c^2+1)$
$\log[(b^2c^2+c^2+1)/(c^2+1)]<\log(b^2+1)$
$\log[(b^2c^2+c^2+1)/(c^2+1)]<\log(b^2c^2+c^2+1)$
(b)

Figure 3.1: The DI value and time lag for both hub and chain models. (a) **X** drives **Y** and **Z** with different time delays; (b) **X** drives **Y**, **Y** drives **Z**. Solid line indicates an actual connection in the real network and dash line represents no connection. The values shown inside and outside the triangle are the DI value and time lag between two processes, respectively.

time-lagged CDI resembling the time-lagged DI as follows:

$$\begin{aligned} &DI_k(X_{k-d}X_{k-d+1} \to Y_kY_{k+1}||Z_{k-l}Z_{k-l+1}) = H(X_{k-d}Z_{k-l}) + H(Y_kZ_{k-l}) \\ &- H(X_{k-d}Y_kZ_{k-l}) - H(Z_{k-l}) + H(X_{k-d}X_{k-d+1}Y_kZ_{k-l}Z_{k-l+1}) \\ &+ H(Y_kY_{k+1}Z_{k-l}Z_{k-l+1}) - H(Y_kZ_{k-l}Z_{k-l+1}) - H(X_{k-d}X_{k-d+1}Y_kY_{k+1}Z_{k-l}Z_{k-l+1}). \end{aligned} \tag{3.4}$$

where X_{k-d} and Z_{k-l} refer to the $(k-d)$th and $(k-l)$th time samples of random processes **X** and **Z**, respectively, $k = \max[d, l] + 1, \cdots, N-1$, $d, l = 0, \cdots, L-1$, which are the time lag variables, and N is the length of the signal.

When **X**, **Y** and **Z** are not normally distributed, the conditional directed information can be expressed and estimated in terms of mutual information as follows:

$$\begin{aligned} &DI_k(X_{k-d}X_{k-d+1} \to Y_kY_{k+1}||Z_{k-l}Z_{k-l+1}) = I(X_{k-d};(Y_kZ_{k-l})) - I(X_{k-d};Z_{k-l}) \\ &+ I((X_{k-d}X_{k-d+1});(Y_kY_{k+1}Z_{k-l}Z_{k-l+1})) - I((X_{k-d}X_{k-d+1});(Y_kZ_{k-l}Z_{k-l+1})). \end{aligned} \tag{3.5}$$

To quantify the same amount of causally conditional dependency and be able to distinguish the indirect connections from direct connections, proper values of d and l should be chosen for the computation of time-lagged conditional directed information. Actually, when the time delay between any two variables is known (time delay between $\mathbf{X} \to \mathbf{Y}$ is m, $\mathbf{X} \to \mathbf{Z}$ is n and $\mathbf{Y} \to \mathbf{Z}$ is $n-m$), then for the hub model, $DI_k(Y_{k-(n-m)}Y_{k-(n-m)+1} \to Z_kZ_{k+1}||X_{k-n}X_{k-n+1}) = 0$. While for the chain model, $DI_k(X_{k-n}X_{k-n+1} \to Z_kZ_{k+1}|| Y_{k-(n-m)}Y_{k-(n-m)+1}) = 0$ (See Appendix B for details). Thus the time-lagged conditional directed information can detect the indirect connections with the CDI value being equal to 0, which is aligned with what we expect from the model. Therefore, to remove the indirect connections in a network inference problem, we need to calculate the time-lagged conditional DI for both chain and hub cases respectively. For three connected nodes with significant DI values, if we do not know whether it is a hub or chain pattern but know the time lags between any two nodes, e.g. the time lag from $\mathbf{X}$ to $\mathbf{Y}$ is m, $\mathbf{X}$ to $\mathbf{Z}$ is l_c, and $\mathbf{Y}$ to $\mathbf{Z}$ is l_h, then we can calculate the time-lagged CDI (TLCDI) for both chain and hub cases along the whole time series as follows:

- For the hub case:

$$TLCDI_{Y,Z||X} = \frac{1}{2} \sum_{k=\max[l_c,l_h]+1}^{N-1} DI(Y_{k-l_h}Y_{k-l_h+1} \to Z_kZ_{k+1}|X_{k-l_c}X_{k-l_c+1}), \tag{3.6}$$

  where N is the length of the signal.

- For the chain case:

$$TLCDI_{X,Z||Y} = \frac{1}{2} \sum_{k=\max[l_c,l_h]+1}^{N-1} DI(X_{k-l_c} X_{k-l_c+1} \to Z_k Z_{k+1} | Y_{k-l_h} Y_{k-l_h+1}), \tag{3.7}$$

  where N is the length of the signal.

Once the time-lagged conditional directed information is obtained for both cases, it can be applied to the network inference problem.

3.2.4 Modified conditional directed information

The TLCDI is less computationally complex than the original definition of CDI, but is limited to trivariate autoregressive models with single order. Similar to the modified directed information in the previous chapter, we propose the modified conditional directed information. Consider a general Markov model, where random process Y^N is influenced by X^N and Z^N such that $p(Y_n|X_{1:n}, Y_{1:n-1}Z_{1:n}) = p(Y_n|X_{n-p_2:n-p_1}, Z_{n-p_4:n-p_3}Y_{n-p_5:n-1})$. In this model, it is assumed that $\mathbf{X}$ ($\mathbf{Z}$) starts to influence $\mathbf{Y}$ with a delay of p_1 (p_3) samples and this influence lasts for $p_2 - p_1 + 1$ ($p_4 - p_3 + 1$) time samples, where $p_2 \geq p_1$ and $p_4 \geq p_3$. p_5 is the order of Y. The upper bound of each term of conditional directed information is as follows:

$$\begin{aligned}
&I(X^n; Y_n|Y^{n-1}Z^n) = H(Y_n|Y^{n-1}Z^n) - H(Y_n|Y^{n-1}X^nZ^n)\\
&= H(Y_n|Y^{n-1}Z^n) - H(Y_n|X_{n-p_2:n-p_1}, Z_{n-p_4:n-p_3}Y_{n-p_5:n-1})\\
&\leq H(Y_n|Z_{n-p_4:n-p_3}Y_{n-p_5:n-1}) - H(Y_n|X_{n-p_2:n-p_1}, Z_{n-p_4:n-p_3}Y_{n-p_5:n-1})\\
&= I(X_{n-p_2:n-p_1}; Y_n|Z_{n-p_4:n-p_3}Y_{n-p_5:n-1}),
\end{aligned} \tag{3.8}$$

where the second equality comes from the Markov property and the inequality is true since conditioning reduces entropy. Therefore, similar to modified directed information, we define the upper bound of the conditional directed information as the modified conditional directed information (MCDI) with $p_1 = p_3 = 1$ and $p = \max(p_2, p_4, p_5)$ to reduce the number of parameters,

$$MCDI_{X,Y||Z} = \sum_{n=p+1}^{N} I(X_{n-p:n-1}; Y_n | Y_{n-p:n-1} Z_{n-p:n-1}). \tag{3.9}$$

Note that letting $p_1 = p_3 = 1$ does not lose any of the information flow compared to using the actual time delay, $p_1 > 1$ and $p_3 > 1$. The only drawback of letting $p_1 = p_3 = 1$ is that the computational complexity of estimating the joint entropies increases since the length of the window to compute MCDI increases and the dimensionality increases. The main reason why we let $p_1 = p_3 = 1$ is because estimating the actual value for the delay accurately is not practical when the amount of data is limited.

For the hub case, when $p(Z_n|Z_{n-p:n-1}X_{n-p:n-1}Y_{n-p:n-1}) = p(Z_n|X_{n-p:n-1}, Z_{n-p:n-1})$, then

$$\begin{aligned} MCDI_{Y,Z||X} &= \sum_{n=p+1}^{N} I(Y_{n-p:n-1}; Z_n | Z_{n-p:n-1} X_{n-p:n-1}) \\ &= \sum_{n=p+1}^{N} [H(Z_n|Z_{n-p:n-1}X_{n-p:n-1}) - H(Z_n|Z_{n-p:n-1}X_{n-p:n-1}Y_{n-p:n-1})] \\ &= \sum_{n=p+1}^{N} [H(Z_n|Z_{n-p:n-1}X_{n-p:n-1}) - H(Z_n|Z_{n-p:n-1}X_{n-p:n-1})] \\ &= 0. \end{aligned} \tag{3.10}$$

For the chain case, when $p(Z_n|Z_{n-p:n-1}X_{n-p:n-1}Y_{n-p:n-1}) = p(Z_n|Y_{n-p:n-1}, Z_{n-p:n-1})$, then

$$\begin{aligned}
MCDI_{X,Z||Y} &= \sum_{n=p+1}^{N} I(X_{n-p:n-1}; Z_n | Z_{n-p:n-1} Y_{n-p:n-1}) \\
&= \sum_{n=p+1}^{N} [H(Z_n | Z_{n-p:n-1} Y_{n-p:n-1}) - H(Z_n | Z_{n-p:n-1} X_{n-p:n-1} Y_{n-p:n-1})] \\
&= \sum_{n=p+1}^{N} [H(Z_n | Z_{n-p:n-1} Y_{n-p:n-1}) - H(Z_n | Z_{n-p:n-1} Y_{n-p:n-1})] \\
&= 0.
\end{aligned} \tag{3.11}$$

Therefore, the modified CDI can discriminate the direct connection from the indirect connection. MCDI is a general extension of TLCDI, which takes the influence of multiple time samples into account.

3.3 Network inference algorithms

In the previous section, we proposed simplified version of directed information and conditional directed information with minimal computational complexity to quantify the causal dependencies. In the following section, we propose three algorithms for directed network inference based on DI and CDI. As discussed before, using directed information alone will cause false connections. Any three nodes in a network can interact with each other through two possible scenarios as equations (3.2) and (3.3) show: three nodes interacting through a hub node $\mathbf{Y} \leftarrow \mathbf{X} \rightarrow \mathbf{Z}$ or interacting as a chain $\mathbf{X} \rightarrow \mathbf{Y} \rightarrow \mathbf{Z}$. As discussed in the previous section, in both cases, if three pairs $(\mathbf{X}, \mathbf{Y})$, $(\mathbf{Y}, \mathbf{Z})$, $(\mathbf{X}, \mathbf{Z})$ all have large DI values, they will infer the same directed network. Therefore, directed information fails to discriminate between direct and indirect dependencies. This leads to two questions to address the problem: (1) If we can determine the time lag between any two nodes, can we infer the network based on the time lag information? (2) Since the conditional directed information can evaluate the amount of information flow from $\mathbf{X}$ to $\mathbf{Y}$ given $\mathbf{Z}$, can we use this information for network inference? We introduce three approaches to remove indirect causal connections in a network based on the time lag and conditional directed information.

3.3.1 Algorithm based on directed information and conditional directed information

The first inference algorithm uses directed information and conditional directed information with no time lag to infer the direct interactions between nodes [95]. In the first step (lines $3-14$) of Algorithm 1 in Figure 3.2, we calculate the TLDI between two nodes i and j with $d=0$ in equation (2.21), which corresponds to applying DI to every two time samples of i and j without time delay. To test the significance level of the obtained value of DI, the p-value is determined by using the distribution of DI values computed from trial shuffling data sets. If the p-value of $DI(i \to j)$ is less than the threshold of p-value, i.e. α, which is determined by controlling the FDR bound q as 0.05, the information flow from node i to j is assumed to be significant. In order to identify the direct vs. indirect connections (lines $15-25$), for each node pair i, j, we evaluate the conditional directed information given any other node k for all $k \neq i, j$. If node k does not interact with nodes i and j, then $DI(X_i \to X_j|X_k)$ should be close to $DI(X_i \to X_j)$. Otherwise, if k is an intermediate or hub node between i and j, $DI(X_i \to X_j|X_k)$ should be close to 0. Therefore, we compute $DI(X_i \to X_j|X_k)/DI(X_i \to X_j)$, which is close to 1 when there are no interactions between k and i, j. We rank order this quantity from high to low for all k. The connections with the highest ranks are kept until a desired number of connections, ec, is achieved.

3.3.2 Algorithm based on time lag

From equation (2.18), we can see that in some cases when $d=0$, applying DI to every two time samples of $\mathbf{X}$ and $\mathbf{Y}$ can not capture the same amount of information as the actual causal information. Therefore, the proposed Algorithm 1 (Figure 3.2) based on DI and CDI ($d=0$) without considering time lag will encounter the problem of not being able to capture

```
Input time series for M nodes, $ec$ is the expected number of connections;
Initialize $D \in R^{n \times n}, C \in [0,1]^{(n \times n)}, cD \in R^{n \times n}$ as zero matrices;
for $i = 1$ to $M$ do
    for $j = 1$ to $M$ do
        $D_{i,j} \Leftarrow TLDI_0(i \to j)$;
        Shuffle the trials of the time series of node $i$;
        Calculate the $p$-value of each node pairs $P_{i,j}$;
        if $P_{i,j} < \alpha$ then
            $C_{i,j} = 1$;
        else
            $C_{i,j} = 0$;
        end if
    end for
end for
for $i = 1$ to $M$ do
    for $j = 1$ to $M$ do
        if $C_{i,j} == 1$ then
            $cD_{i,j} \Leftarrow \min\left(\frac{DI(x_i \to x_j | x_k)}{D_{i,j}}\right), k \neq i, j;$
        else
            $cD_{i,j} \Leftarrow 0$;
        end if
    end for
end for
$c_a = cD(:)$;
$c_b = sort(c_a)$ in descending order and keep the top $ec$ connections
```

Figure 3.2: Algorithm 1: directed network inference based on DI and CDI

the causal dependencies between two random processes when there is a nonzero time lag. Therefore, we propose an inference algorithm which utilizes time-lagged directed information and time lag to infer the direct interactions between nodes. We first show how time lag can be used to infer connections and then introduce the algorithm. In a complex system, if we assume that any node pair exchanges information through the fastest available path, then the indirect causal relationships can be detected by determining the time lag between two variables in the system. According to equation (2.18), for the single order linear model, the time lag between two time series can be found as follows:

$$TL(\mathbf{X} \to \mathbf{Y}) = \arg\max_{d} \sum_{k=d+1}^{N-1} DI(X_{k-d}X_{k-d+1} \to Y_kY_{k+1}). \tag{3.12}$$

where $d = 1, \cdots, L$ and N is the length of the signal. In order to illustrate the inference of causality using time lag, we will consider two trivariate autoregressive models considered in equations (3.2) and (3.3). For both cases, the time lag and the corresponding DI values of direct and indirect connections are shown in Figure 3.1. We can observe that although the DI value of any two of the random variables is significant, the value for the indirect connections ($\mathbf{Y} \to \mathbf{Z}$ in Figure 3.1(a), $\mathbf{X} \to \mathbf{Z}$ in Figure 3.1(b)) are smaller than the direct connections. Moreover, it is obvious that in these two models, the three time lags show a linear relationship. Therefore, in a real system with the assumption that any node pairs exchange their information through the fastest available route, for the chain case ($\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$), if $TL(\mathbf{X} \to \mathbf{Z}) \geq TL(\mathbf{X} \to \mathbf{Y}) + TL(\mathbf{Y} \to \mathbf{Z})$, then $\mathbf{Y}$ is the intermediate node and there is no direct causal relationship between $\mathbf{X}$ and $\mathbf{Z}$. While for the hub case ($\mathbf{Z} \leftarrow \mathbf{X} \to \mathbf{Y}$), if $TL(\mathbf{Y} \to \mathbf{Z}) \geq TL(\mathbf{X} \to \mathbf{Z}) - TL(\mathbf{X} \to \mathbf{Y}) > 0$, there is no direct causal relationship between $\mathbf{Y}$ and $\mathbf{Z}$.

Based on this observation, we introduce an algorithm based on time-lagged DI and time lag to infer the directed network which is shown in Figure 3.3. In practice, for the model with maximum order $p > 1$, the time-lagged DI is obtained by averaging across different lags, i.e. $\frac{1}{p+1}\sum_{d=0}^{p} TLDI_d(X^N \to Y^N)$. In the first step (lines $3 - 15$), we calculate the time-lagged directed information $DI_{i,j}$ between two nodes i and j and the corresponding time delay $TL_{i,j}$ based on equations (2.21) and (3.12), respectively. p-value of $DI_{i,j}$ is determined for each pair of nodes. Connections with p-values less than the threshold α are kept with α determined by controlling the FDR bound q. In order to identify the direct vs. indirect connections (lines $16 - 27$), for each node i, we consider connected triplets. For node i, if $TL_{j,k} \geq TL_{j,i} + TL_{i,k}$ and $DI_{j,k}$ is smaller than both $DI_{j,i}$ and $DI_{i,k}$, then i is the intermediate node ($j \to i$, $i \to k$) and there is no connection between j and k. Otherwise, if

```
Input time series for M nodes;
Initialize $D \in R^{n\times n}, TL \in (0, L-1)^{n\times n}, P \in (0,1)^{n\times n}, C \in [0,1]^{(n\times n)}$, as zero matrices;
for $i = 1$ to $M$ do
    for $j = 1$ to $M$ do
        $DI_{i,j} \Leftarrow TLDI(i \to j)$;
        $TL_{i,j} \Leftarrow TL(i \to j)$;
        Shuffle the trials of the time series of node $i$;
        Calculate the $p$-value of each node pairs $P_{i,j}$;
        if $P_{i,j} < \alpha$ then
            $C_{i,j} = 1$;
        else
            $C_{i,j} = 0$;
        end if
    end for
end for
for $i = 1$ to $M$ do
    for $j = 1$ to $M$ do
        for $k = 1$ to $M$ do
            if $C_{j,i} = 1$ and $C_{i,k} = 1$ and $C_{j,k} == 1$ and $TL_{j,k} \geq TL_{j,i} + TL_{i,k}$ and $DI_{j,k} < \min(DI_{j,i}, DI_{i,k})$ then
                $C_{j,k} = 0$;
            end if
            if $C_{i,j} = 1$ and $C_{i,k} = 1$ and $C_{j,k} == 1$ and $TL_{j,k} \geq TL_{i,k} - TL_{i,j}$ and $DI_{j,k} < \min(DI_{i,j}, DI_{i,k})$ then
                $C_{j,k} = 0$;
            end if
        end for
    end for
end for
```

Figure 3.3: Algorithm 2: directed network inference based on TLDI and TL

$TL_{j,k} \geq TL_{i,k} - TL_{i,j} > 0$ and $DI_{j,k}$ is smaller than both $DI_{i,j}$ and $DI_{i,k}$, then node i is the hub node ($i \to j$, $i \to k$) and the connection between j and k should be removed.

3.3.3 Algorithm based on modified time-lagged directed information and conditional directed information

Algorithm 2 (Figure 3.3)based on time-lagged directed information and time lag information can capture more causal dependencies than Algorithm 1. However, when the relationship between two variables is complex, e.g. nonlinear or multi-order, it is hard to detect the exact time lag because of the complexity of the underlying model and limited sample sizes. Therefore, we propose an inference algorithm of using MDI and MCDI. To explain the algorithm clearly, we consider the hub and chain cases again. For triple connected nodes i, j, k ($i \to j$, $i \to k$, $j \to k$), if j is the intermediate node, $DI(i \to k|j) = 0$ for the ideal case. However, because of noise and the bias of the estimator, it will not be exactly equal to 0. Therefore, in order to confirm j is the intermediate node and remove the indirect connection $i \to k$, two conditions should be satisfied: (1) $DI(i \to k|j) < DI(j \to k|i)$; (2) $DI(i \to k) < DI(j \to k)$. Similarly, if i is the hub node, then $DI(i \to k|j) > DI(j \to k|i)$ and $DI(i \to k) > DI(j \to k)$. If the two conditions contradict each other, we remove the connection with the smallest value of DI.

Based on this analysis, our Algorithm 3 (Figure 3.4) is described as follows: First (lines $3-14$), we calculate the modified time-lagged directed information $D_{i,j}$ from i to j according to equation (2.25). If the p-value of $D_{i,j}$ is larger than α, then there is no directed path from i to j. To remove the indirect causality (lines $15 - 39$), for each connected triplet nodes i, j and k without a loop ($i \to j$, $i \to k$, $j \to k$), which include both chain and hub connection patterns, we calculate the modified time-lagged conditional directed information (hub: $cD_{jk|i}$, chain: $cD_{ik|j}$). If $cD_{jk|i}$ and $D_{j,k}$ are less (greater) than $cD_{ik|j}$ and $D_{i,k}$, respectively, we keep connection $i \to k$ ($j \to k$). Otherwise, if $D_{i,j}$ is the largest value, we remove both connections $i \to k$ and $j \to k$; if not, we remove the connection with the smallest DI value.

3.3.4 Validation

In order to evaluate the performance of the different network inference algorithms, F-score is adopted [91]. The F-score can be interpreted as a weighted average of the precision and recall and is defined as:

$$F = \frac{2pr}{p+r}, p = \frac{TP}{TP+FP}, r = \frac{TP}{TP+FN}. \tag{3.13}$$

```
Input time series for M nodes;
Initialize $D \in R^{n\times n}, P \in (0,1)^{n\times n}, C \in [0,1]^{(n\times n)}$ as zero matrices;
for $i = 1$ to $M$ do
    for $j = 1$ to $M$ do
        $D_{i,j} \Leftarrow MDI(i \to j)$;
        Shuffle the trials of the time series of node $i$;
        Calculate the $p$-value of each node pairs $P_{i,j}$;
        if $P_{i,j} < \alpha$ then
            $C_{i,j} = 1$;
        else
            $C_{i,j} = 0$;
        end if
    end for
end for
for $i = 1$ to $M$ do
    for $j = 1$ to $M$ do
        for $k = 1$ to $M$ do
            if $C_{i,j} = 1$ and $C_{i,k} = 1$ and $C_{j,k} == 1$ then
                Hub: $cD_{j,k|i} = MCDI_{YZ|X}$;
                Chain: $cD_{i,k|j} = MCDI_{XZ|Y}$;
                if $cD_{jk|i} < cD_{ik|j}$ AND $D_{j,k} < D_{i,k}$ then
                    $C_{j,k} = 0$
                else if $cD_{jk|i} > cD_{ik|j}$ AND $D_{j,k} > D_{i,k}$ then
                    $C_{i,k} = 0$
                else
                    if $D_{i,j} > \max(D_{i,k}, D_{j,k})$ then
                        $C_{i,k} = 0; C_{j,k} = 0$;
                    else if $D_{i,j} < \min(D_{i,k}, D_{j,k})$ then
                        $C_{i,j} = 0$
                    else if $D_{i,k} < D_{j,k}$ then
                        $C_{i,k} = 0$
                    else
                        $C_{j,k}$
                    end if
                end if
            end if
        end for
    end for
end for
```

Figure 3.4: Algorithm 3: directed network inference based on MDI and MCDI

A positive label predicted by the algorithm is considered as true positive (TP) or false positive (FP) depending on whether there exists a corresponding edge in the true network or not. Similarly, a negative label can be a true negative (TN) or false negative (FN) depending on the true network. In this dissertation, we will compute the F-score before and after introducing CDI or time lag for each algorithm.

3.4 Results

3.4.1 Synthetic data: Linear network

In order to test the effectiveness of the proposed algorithms, we first consider a linear multivariate autoregressive model to reduce the impact of DI estimation on the accuracy of network inference. The following linear autoregressive model is considered [96],

$$\begin{pmatrix} Y_1(n) \\ Y_2(n) \\ \vdots \\ Y_r(n) \end{pmatrix} = \sum_{i=1}^{p} \mathbf{A}_i \begin{pmatrix} Y_1(n-i) \\ Y_2(n-i) \\ \vdots \\ Y_r(n-i) \end{pmatrix} + \begin{pmatrix} e_1(n) \\ e_2(n) \\ \vdots \\ e_r(n) \end{pmatrix} \tag{3.14}$$

where $\mathbf{Y}(n) = [Y_1(n), Y_2(n), \cdots, Y_r(n)]$ is the nth sample of a r-dimensional time series generated by r variables, with each $\mathbf{A}_i$ being a r-by-r matrix of coefficients (weights), coefficients are in the range of 0 to 1 and also keep the system stable. $\mathbf{e}_n = [e_1(n), e_2(n), \cdots, e_r(n)]$ is additive white Gaussian noise with zero mean and unit variance. In this dissertation, we generate a synthetic network of 18 nodes ($r = 18$) with the maximum time lag $p = 4$. The network contains only linear dependencies as shown in Figure 3.5(a). We generate 1024 realizations of the 18 different time series for each node and the results are averaged across 10 simulations. The order between any two time series is determined by the order selection criterion. We then compute the time lag and modified time-lagged DI over 12 time samples according to equations (3.12) and (2.25) and compute the p-value for each DI value under the distribution of the null hypothesis. If the p-value of DI is less than the threshold $\alpha = 0.01$ for each pair of nodes with α found by controlling the FDR bound q as 0.05, we keep the connection. Both time lag and time-lagged conditional directed information are used to eliminate the indirect connections. The F-scores using three different algorithms before and after applying CDI, time lag and MCDI are shown in Figure 3.5(b). We can observe that, Algorithm 3 (Figure 3.4) reaches the highest F-score 0.9351 and it improves the network reconstruction when introducing MCDI to remove the indirect causality relationship. The introducing of time lag information to remove the indirect connection leads to slight change in F-score. It is due to the difficulty of estimating the time lag accurately for a multi-order model in particular when the number of realizations is limited. Algorithm 1 has lower F-score compared to the other two algorithms, because it can not capture the whole causal information between two variables or among multiple variables. However, introducing CDI, though without considering the time lag information can still remove parts of the indirect causality relationship.

Without loss of generality, we permuted the matrices A_1 and A_2 50 times to compare the performance of different algorithms further, and we let A_3 and A_4 equal to all zero matrices to reduce the order and computational complexity. Only the location of the connections were changed with the additional constraint that there are no connected triplets in the permuted networks. More important, in practice the realizations of each random processes is limited. Therefore, we only generate 256 realizations of 18 different time series. The average F-score for each algorithm is shown in Table 3.1. We also compare the proposed algorithms based on DI with model based methods, such as PDC. Matlab toolbox for PDC developed by Baccala is used [44]. In practice, the implementation of PDC depends on the length of the signal and the performance of DI based algorithms relies on the number of realizations of time series. Therefore, to compare the two measures, the length of the generated signal for PDC is chosen as 256, which is the same with the number of realizations for DI. The results

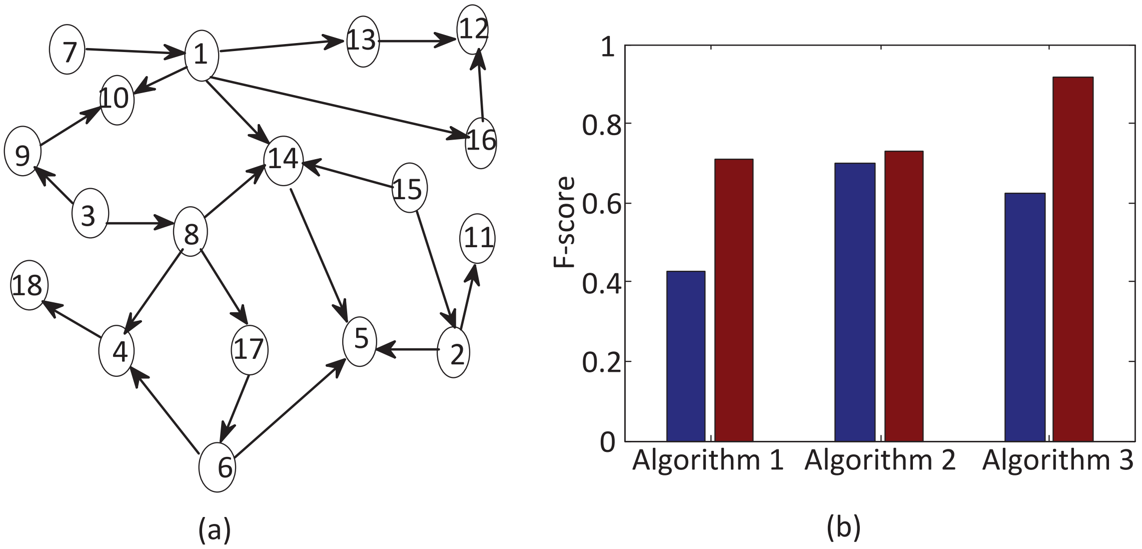


Figure 3.5: The performance of proposed algorithms for linear network inference. (a) The synthetic linear network. (b) Average F-score of Algorithm 1 before and after applying causally conditioned directed information without considering time lag information; average F-score of Algorithm 2 before and after applying time lag; average F-score of Algorithm 3 before and after applying MCDI.

Table 3.1: Average F-score for three proposed algorithms and PDC for linear network

	Algorithm	Mean
1	DI+CDI	0.8270
2	TLDI+TL	0.7903
3	MDI+MCDI	0.8301
4	PDC	0.7351

are shown in Table 3.1. We observe that Algorithm 3 outperforms the other algorithms. The performance of Algorithm 1 is slightly worse than Algorithm 3, because DI without considering time lag information will not lose all causal information when the order of the model is low. The failure of Algorithm 2 is due to the inaccurate estimation of time lag. PDC is based on MVAR model and is expected to have better performance, but since the length of the signal is limited and there is not enough time samples to reconstruct the model, PDC fails to reveal the network structure compared to other algorithms.

3.4.2 Synthetic data: Nonlinear network

In this subsection, we test our algorithm on a synthetic nonlinear network of 14 nodes in [95]. The network contains both linear and nonlinear causality as shown in Figure 3.6(a):

$$
\begin{aligned}
X_1(n) &= 0.7X_1(n-1) + w_1(n);\\
X_2(n) &= 0.29X_2(n-1) + 0.65X_1(n-1) + w_2(n);\\
X_3(n) &= 0.15X_3(n-1) + 0.79X_2(n-1) + \frac{0.9X_{14}(n-1)}{1+e^{-6X_{14}(n-1)}} + w_3(n);\\
X_4(n) &= 0.17X_4(n-1) + 0.7X_3(n-1) + 0.7X_6(n-1) + w_4(n);\\
X_5(n) &= 0.6X_5(n-1) + 0.8X_4(n-1) + w_5(n);\\
X_6(n) &= 0.12X_6(n-1) + 0.8X_7(n-1) + \sqrt{0.7X_8(n-1)} + \frac{0.6+0.8X_9(n-1)}{1+e^{-2X_7(n-1)}} + w_6(n);\\
X_7(n) &= 0.8X_7(n-1) + w_7(n);\\
X_8(n) &= 0.7X_8(n-1) + w_8(n);\\
X_9(n) &= 0.77X_9(n-1) + w_9(n);\\
X_{10}(n) &= 0.4X_{10}(n-1) + 0.4[X_{11}(n-1)]^2 + w_{10}(n);\\
X_{11}(n) &= 0.7X_{11}(n-1) + \frac{0.7X_{14}(n-1)}{1+e^{-6X_8(n-1)}} + w_{11}(n);\\
X_{12}(n) &= 0.4X_{12}(n-1) + 0.8X_{11}(n-1) + w_{12}(n);\\
X_{13}(n) &= 0.4X_{13}(n-1) + 0.9X_{11}(n-1) + w_{13}(n);\\
X_{14}(n) &= 0.65X_{14}(n-1) + w_{14}(n);
\end{aligned}
\tag{3.15}
$$

where $w_1, \cdots, w_{14}$ are white Gaussian random processes. We generate 512 realizations of the 14 different time series for each node and compute the time-lagged DI over 12 time samples. The p-values of each DI value were computed and the connections with p-values less than α are kept. To remove the indirect connections, algorithms based on time lag and MCDI were used, respectively. We generate this simulation 10 times and the averaged F-score before and after applying CDI without lag information, time lag and MCDI for each algorithm are shown. From Figure 3.6(b), we see that the proposed algorithm using both MDI and MCDI has better performance compared to other two algorithms, and obtains a F-score of 0.9455, which effectively detects all linear and nonlinear dependencies. Algorithm 1 using DI and CDI without considering any time lag information performs worst since it may lose important causal information to reveal the structure of the network. Algorithm 2 performs better than Algorithm 1 but worse than Algorithm 3, which is due to the inaccurate estimation of the time lag in a complex network with nonlinear interactions. Therefore, we observe that time-lagged directed information and modified directed information can effectively detect most of the causal interactions, and time-lagged conditional directed information and modified conditional directed information are able to effectively removes most of the false positive connections.

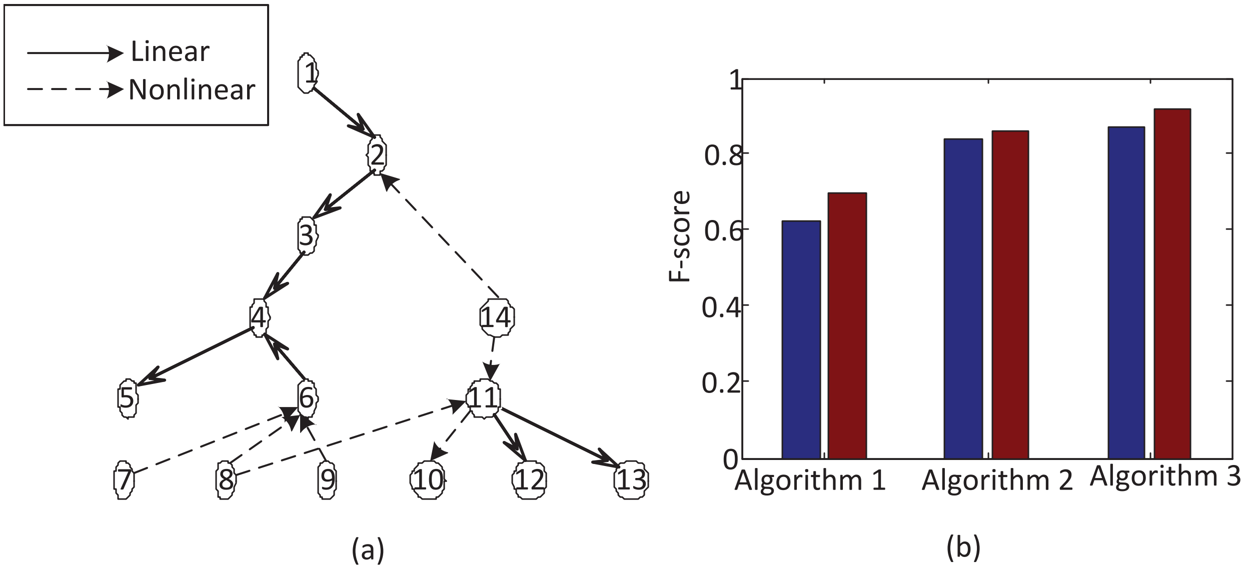


Figure 3.6: The performance of proposed algorithms for nonlinear network inference. (a) The synthetic nonlinear network. (b) Average F-score of Algorithm 1 before and after applying causally conditioned directed information without considering time lag information; average F-score of Algorithm 2 before and after applying time lag; average F-score of Algorithm 3 before and after applying MCDI.

Similarly, to verify the effectiveness of our algorithms further, we permuted the matrices A ten times, which only changes the location of the connections with the additional constraint that there are no connected triplets. The average F-score for each algorithm is shown in

Table 3.2: Average F-score for three algorithms of nonlinear network

	Algorithm	Mean
1	DI+CDI	0.6580
2	TLDI+TL	0.7798
3	MDI+MCDI	0.8175
4	PDC	0.7087

Table 3.2. Similarly, we also apply PDC to the randomized networks and the length of the generated signal for PDC is set to 512. The performance of PDC is also shown in Table. 3.2. Our proposed algorithm has the highest F score among all the algorithms, indicating more efficient and stable performance for nonlinear networks. The performance of PDC is worse than Algorithm 2 and 3, because PDC is implemented under a linear AR frame work, which is not aligned with the actual model of the signal and fails to reconstruct the network correctly.

3.5 Discussions

3.5.1 Problems with current algorithms

The proposed algorithm based on modified directed information and modified conditional directed information can effectively infer the directed network. However, there are still some issues remaining with the current algorithms. First, the bias and variance of DI estimator depends on the sample size, which may affect the significance testing of the DI value and in turn may increase false negative connections. Second, the algorithm is based on the assumption that the network is acyclic and has no triple connected nodes. The algorithm considers connections in a connected triplet and break at least one of the connections. Therefore, if a particular connection does not connect any triplets, the algorithm will keep it without applying conditional directed information on it. In fact, algorithms in this thesis can be extended to cyclic networks by stopping before the step of removing the connection with the smallest DI value. Third, the detection of time lags is limited to linear models with single order. When the model is complex, the effectiveness of determining the time lag according to the maximum time-lagged directed information needs to be further proven. Finally, the proposed algorithms just consider the scenario that there is only one intermediate node between two nodes. If there are two or more intermediate nodes between $\mathbf{X}$ and $\mathbf{Z}$, i.e., $\mathbf{X} \rightarrow \mathbf{Y} \rightarrow \mathbf{W} \rightarrow \mathbf{Z}$, the algorithm may fail to remove the indirect connection by computing the CDI from $\mathbf{X}$ to $\mathbf{Z}$ only conditioned on either $\mathbf{Y}$ or $\mathbf{W}$. In this case, $DI(X^N \rightarrow Z^N||Y^N W^N)$ needs to be considered to remove the false positive connection $\mathbf{X} \rightarrow \mathbf{Z}$. However, the computational complexity of CDI increases with the number of nodes involved, which is hard to implement.

3.5.2 Comparison with existing algorithms

The major contributions of this chapter are using the directed information and conditional directed information to infer a directed network. We propose modified conditional directed information for both chain and hub cases to remove the indirect causality and at the same time to reduce the computational complexity.

Model-dependent methods such as PDC have been widely used for effective network inference. Though they have good performance when the system is linear, there are two main issues with model dependent measures. First, model-dependent measures fail to detect the strong time-lagged nonlinear couplings, which might be the case for most complex systems. Second, the number of parameters to be estimated increases with the order of the model, which requires a large sample size. When the sample size is limited, model-dependent measures may fail to model the system accurately and detect the directed connectivity. On the other hand, because DI is a model free measure to capture the causal information in a system, our algorithms based on DI are easier to implement for complex systems with strong nonlinear relationships and high orders. Though the performance of our measure is also influenced by the sample size, the modified version of DI and CDI address this problem.

3.6 Conclusions

In this chapter, we pointed out the drawback of only using directed information for network inference and introduced the conditional directed information to address this problem. Moreover, time-lagged conditional directed information and the modified conditional directed information are proposed to reduce the computational complexity. Three algorithms are proposed for network inference and the results for both linear and nonlinear networks are shown. The simulation results show that our methods outperform the existing measures.

Chapter 4

COMMUNITY DETECTION FOR DIRECTIONAL NEURAL NETWORKS

4.1 Introduction

A remaining issue with the analysis of the brain network is to reveal the underlying community structure. The complex network theory has been used to show that both the functional and structural brain networks follow a small world topology characterized by a short minimum path length between all pairs of nodes in the network together with a high clustering coefficient [97, 34]. Although small-worldness summarizes key aspects of complex networks at both global (the whole network) and local (each node together in relationship with its most immediate neighbors) levels, it does not provide any information about the intermediate scale of network organization which is more completely described by the community structure or modularity of the network [35, 98]. The modules of a complex network are subsets of nodes that are densely connected with each other but sparsely connected to nodes in other modules. Module detection also allows one to obtain simplified reduced representations of complex networks in terms of subgraphs or communities. In this chapter, we address the issue of community detection for effective brain networks.

In recent years, a lot of work has been done on applying community detection algorithms from graph theory to the study of functional brain networks [99, 35, 36, 37]. Functional brain networks are usually described by undirected graphs with corresponding symmetric association matrices, where each entry indicates the pairwise functional connectivity between two regions. Therefore, most of the work on community detection for the study of brain networks has focused on undirected networks. For example, Fair *et al.* showed that young children and young adults have different community structures in functional brain networks from the study of resting state fMRI data [35]. Similarly, Ferrarini *et al.* showed that the resting-state human brain has a hierarchical functional module structure [36] and Meunier *et* *al.* revealed age-related changes in the modular structure of human brain functional networks from fMRI [37]. Chavez *et al.* pointed out that the modular structure of the human brain provides important information on the functional organization of the brain during normal and pathological neural activities [38]. However, in the study of brain networks, it is important to quantify both the dependency between different nodes or neuronal populations in the brain as well as the causality between these nodes, i.e. effective connectivity as we mentioned in the previous chapters. Friston pointed out that functional integration of the brain can be better understood through effective connectivity since it reflects the dynamic (activity dependent and time dependent) characteristics of a system [12]. In this sense, the brain network can be better described by an effective network where the edges of the graph have direction and the corresponding association matrix is no longer symmetric. As Kim *et al.* claimed approaches that ignore the direction of links may fail to understand the dynamics of the system, and similarly any community detection approach on these networks may fail to reveal the actual community structure [100]. In addition, Leicht *et al.* also states that abundant useful information about a network's structure will be lost if we ignore the directions of the edges [101]. Therefore, we expect that using effective connectivity would reveal new topological characteristics of the brain network [102]. In this chapter, we propose a multi-subject hierarchical community detection algorithm for weighted directed networks in order to reduce effective brain networks involved in cognitive control into a small number of functional modules.

The approach outlined in this chapter advances the current study of brain networks in several key ways. First, the pairwise relationship between two processes is quantified by effective connectivity, which can reflect both the interaction and the direction of information flow of the network. Second, we employ recent work in the area of community detection in directed networks for infering functional modules. Although most of the literature on community detection focuses on undirected networks, a significant amount of information about a network's structure will be lost if we ignore the directions of the edges [101]. In order to discover the underlying organization of the network, traditional clustering algorithms such as Kernighan-Lin algorithm, agglomerative (or divisive) algorithm and k-means clustering, have been widely used. However, these algorithms need to pre-determine the number of clusters [103, 104, 105, 106]. Therefore, modularity based algorithms are widely used to choose the best partition of a network by maximizing the modularity, which include greedy techniques, spectral optimization, etc. [39]. Recently, Blondel introduced a greedy approach for the modularity optimization for weighted graphs, which is proven to be efficient, multi-level and close to the optimal value obtained from slower methods [1]. In this chapter, we extend this algorithm to weighted directed networks to find the functional modules. Third, we extend community detection algorithms developed for single networks to a group of networks in order to find a modular structure that best describes all of the subjects in a group. Traditionally, two broadly used approaches are employed for group analysis, i.e., 'virtual-typical-subject' (VTS) approach and 'individual structure' (IS) approach. The former approach pools or averages the group data to obtain one community structure for the whole group. The latter one applies the proposed algorithm to each individual subject and finds the common structure for all of them by employing different strategies, such as averaging the community structures across subjects, voting for the community structure that the majority of the subjects agree on, or finding the most representative subject whose structure is the most similar to the other subjects in the group. However, both VTS and IS approaches ignore the between-subject variability and the results may be influenced by outliers [34]. In this chapter, instead of dealing with individual results from each subject, we focus on modifying the community detection algorithm itself by combining all of the information from the data. To be specific, instead of maximizing the modularity and identifying the communities of each subject separately, we detect the community structure of a group by maximizing the total modularity of the group. Our proposed method has the advantage of being more computationally efficient than IS, because IS spends a large amount of time in detecting the community structure for each subject. Finally, the algorithm proposed in this chapter is applied to multivariate EEG recordings which offer a better view of the dynamic change of community structure with high temporal resolution compared to fMRI.

4.2 Background

4.2.1 Modularity

The concept of modularity is motivated by the idea that nodes in the same module have very dense connections and nodes in different modules have sparse inter-module connections [101]. Modularity was proposed as a quality function to choose the best partition of a network such as in the Girvan-Newman algorithm [98] and is widely used as an optimization criterion in many graph clustering methods [107, 34, 1, 101, 108]. A good partition of a network has high modularity Q with Q=(fraction of edges within communities)-(expected fraction of such edges) [98], where the expected fraction of edges is evaluated for a random graph. The original expression of modularity for undirected binary networks is given as,

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta_{c_i,c_j}, \tag{4.1}$$

where A is the adjacency matrix with $A_{ij} = 1$, if i and j are connected, and $A_{ij} = 0$ otherwise, k_i is the degree of vertex i, δ_{c_i,c_j} is equal to 1 when i and j are in the same community and is equal to 0 otherwise. For a directed network, the probability of a directed edge relies on the in-degree and out-degree of the vertex. Leicht *et al.* extended the definition of modularity to directed binary networks as [101],

$$Q_d = \frac{1}{m} \sum_{i,j} \left[A_{ij} - \frac{k_i^{out} k_j^{in}}{m} \right] \delta_{c_i,c_j}, \tag{4.2}$$

where k_i^{in} (k_i^{out}) is the in-degree (out-degree) of vertex i. Arenas *et al.* gave a more general expression of modularity for directed weighted networks [108],

$$Q_{gen} = \frac{1}{W} \sum_{i,j} \left[A_{ij} - \frac{s_i^{out} s_j^{in}}{W} \right] \delta_{c_i,c_j}, \tag{4.3}$$

where $A_{i,j}$ is the weight of edge $e_{i,j}$, s_i^{in} (s_i^{out}) is the in-degree (out-degree) of vertex i, and $W = \sum_{i,j} A_{i,j}$. In this chapter, $A_{i,j}$ is quantified by the directed information measure introduced in Chapter 1.

4.2.2 Overview of community detection methods

Uncovering community structure is a key step to the understanding of a complex network. The idea of community detection is closely related to graph partitioning in computer science and hierarchical clustering in sociology. Therefore, traditional approaches in these two areas can be employed for community detection [109, 39]. Kernighan-Lin algorithm is one of the representative heuristic algorithms in graph partitioning [103]. It starts with an initial partitioning of the graph into two clusters with pre-defined sizes and tries to maximize the differences between the number of edges inside the modules and the number of edges lying between modules. However, the pre-defined size of a cluster is usually unknown and the performance of this algorithm is highly dependent on the initial partition of the graph [39]. On the other hand, hierarchical clustering algorithms do not require any *a priori* knowledge of clusters and try to group nodes into the same community (agglomerative) or split nodes into different clusters (divisive) by quantifying the similarity between pairs of nodes. The results are presented in a dendrogram or tree, but the algorithms fail to express the best level of partitioning during the processes. To find the optimal partitioning, Girvan-Newman algorithm proposed to not only use the 'edge betweenness' measure to remove edges from the network, but also to use the modularity function Q as a stopping criterion to choose the best level of partitioning [98]. In addition, the proposition of modularity function Q leads to the development of another category of community detection algorithms which is based on the optimization of modularity. To optimize modularity, greedy techniques, simulated annealing, extremal optimization and spectral optimization are widely employed. Unlike simulated annealing and extremal optimization, which may encounter the problems of high computational complexity and not being applicable to large networks [110, 39, 111], the greedy strategies and spectral optimization are more commonly used for finding the community structure of a network with the maximum modularity. Newman was the first to propose a greedy strategy to maximize the modularity [112] and since then several improvements of the algorithm were proposed to improve the detection accuracy [39, 113, 114]. Later, Newman proposed a more efficient approach based on the bipartition of the modularity matrix, which uses the eigenvalues and eigenvectors of the modularity matrix to find a solution for the community detection problem (spectral optimization) [107].

Although most of the modularity-based community detection algorithms have focused on binary and undirected networks, in recent years there have been some extensions to weighted and directed networks. Because a weighted and directed link can reflect both the strength and direction of the interaction between two nodes, which can reveal more characteristics of a system than the binary and undirected network. For example, Leicht *et al.* extended the definition of modularity to binary directed networks and employed spectral optimization to find the community structure for directed networks. This technique works well for bipartites, but is less applicable to networks with a large number of communities [39] and weighted links. An alternative approach is to represent the directed network as an undirected bipartite network. To do this, each node in the network is split into two nodes, with one node only receiving information and the other node only sending information. In this way, community detection algorithms for bipartite networks can be employed for directed networks [115]. However, this approach is limited to networks with a small number of clusters. Recently, Blondel *et al.* introduced an alternative greedy algorithm to find the hierarchical structure of undirected weighted graphs [1], the computation time of which is comparable to spectral optimization [39]. In this chapter, we extend this algorithm to directed weighted graphs for community detection.

4.2.3 Group analysis approaches

In many neuroimaging studies, extracting a common set of features or a representation for a group of subjects such as the common community structure is more important than extracting features for individual subjects. This common structure usually gives us an overall understanding of the group, while individual subject level representations show the subject-specific features. There are three major group analysis strategies that can be employed for community detection, i.e., the 'virtual-typical-subject' (VTS) approach, the 'individual structure' (IS) approach, and the algorithm-based approach [116, 117]. The VTS approach assumes that data from each subject performs the same function or follows the same distribution. It reconstructs a virtual subject by pooling or averaging the group data or connectivity matrices and obtains one community structure for the whole group. However, this approach does not consider the inter-subject variability and may fail when the behavior from subject to subject is not consistent [116, 117]. Therefore, the results may not reflect any of the features seen in individual analysis [118]. The IS approach applies a community detection algorithm to every individual subject and finds a unanimous community structure from these individual structures. Since this subject-specific strategy considers diversity across subject-s, various strategies are employed to integrate inconsistent results. A common community structure can be obtained through either averaging individual structures or voting/consensus algorithms [119]. For example, replicator dynamics proposed by Neumann *et al.* can be used to capture nodes that are jointly presented in the same cluster across subjects by analyzing the structures obtained from individual subjects [120]. An alternative strategy to combine the results in IS is to find the most representative subject, whose structure is the most similar to all other subjects in the group [34] and can be used to represent the structure of the whole group. However, this approach may lose important information provided by other subjects, and may be even worse than just averaging the community structure over subjects. The IS approach is usually computationally expensive since it requires the extraction of the community structure for each subject before obtaining a common structure.

Both the VTS and IS approaches focus on either preprocessing the data or post-processing the community structures obtained from each subject. However, none of these approaches reveals the community structure of multiple subjects by directly extending the community detection algorithm to multiple subjects. Recently, some work has been done to address the data-fusion or group inference problems at the algorithm level. Mechelli *et al.* constructed a network or covariance matrix that comprises of m nodes from n subjects and assumed the same model with different model parameters for all subjects [117]. However, this strategy is more suitable for model-based approaches rather than clustering problems. For data-driven problems, Correa *et al.* extended the canonical correlation analysis (CCA) to multi-set CCA by optimizing an objective function of the correlation matrix of multiple canonical variates instead of two such that the correlation among multiple variates is maximized [121]. Similar to the multi-set CCA, which redefines the optimization problem for multiple subjects, in this chapter, we propose a group analysis method by optimizing a common modularity function of directed networks from multiple subjects.

4.3 Algorithm for community detection

In this section, we first extend the method proposed by Blondel *et al.* to weighted directed networks [1] and then propose an extension for group analysis. The algorithm proposed by Blondel *et al.* is originally a modularity based community detection algorithm for undirected weighted networks. To reveal the community structure of an undirected weighted network, it maximizes the modularity through greedy search. Initially, all nodes of the network are in different communities. The algorithm is divided into two phases. In the first step, for each node i, the gain in the modularity when node i is assigned to the community of its neighbor j ($A_{i,j} \neq 0$) is computed and node i is assigned to the community which has the largest positive increase in modularity. This procedure is repeated for all the nodes until the modularity does not increase any more. In the second step, a new network is built by aggregating those nodes in the same community at the first step and forming meta-nodes.

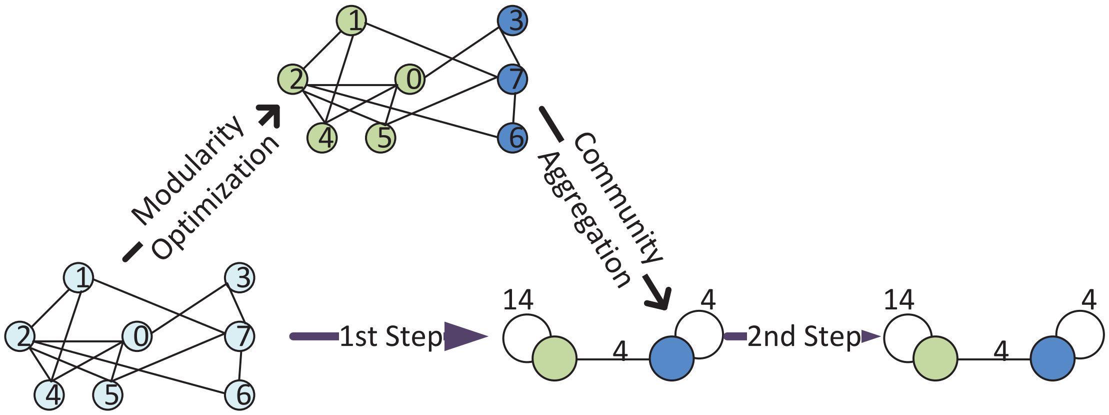


Figure 4.1: Hierarchical optimization of modularity by Blondel *et al.* [1]. The algorithm is divided into two phases. First, each node is assigned to a community and the algorithm tries to combine small communities by optimizing modularity locally. Second, it builds a new network by aggregating those nodes in the same community at the first step. These two steps are repeated iteratively until a maximum of modularity is reached.

The weight of the edge between two meta-nodes is the sum of the weights of edges between nodes in the two corresponding communities from the previous step. These two steps are repeated iteratively until a maximum modularity is reached. Compared to the existing modularity based methods, this approach is fast and can reveal the hierarchical structure of a network [39, 1]. The change of modularity is always computed with respect to the initial graph to guarantee the convergence of the algorithm [39]. An illustration of this algorithm is shown in Figure. 4.1.

4.3.1 Algorithm for community detection in weighted directed networks

In this subsection, in order to reveal the community structure of the functional human brain network which is known to have a hierarchical structure [37], we propose to extend Blondel's approach to directed weighted networks. Initially, all vertices of the graph are put in different communities. The algorithm for uncovering the community structure of a directed weighted network consists of two steps. First, for each node i, the gain in the modularity ΔQ_{gen_j} is computed when the node is assigned to the communities of all other nodes j, where $j = 1, \cdots, N$, $j \neq i$. The original algorithm only evaluates the change of modularity when node i is assigned to the communities of its neighbors j, where j is defined as the neighbor of i when $A_{i,j} \neq 0$, which may be inaccurate and yield spurious partitions in practical cases [39]. For this reason, we consider the change of modularity with respect to all other nodes. Once ΔQ_{gen_j} is obtained, where $j = 1, \cdots, N$, $j \neq i$, the community for which ΔQ_{gen_j} is positive and highest is chosen as the new community for node i. ΔQ_{gen_j}, which partly determines the efficiency of the algorithm, can be computed as follows,

$$\begin{aligned} \Delta Q_{gen_j} = & \frac{1}{W} \sum_{p=1}^{N_j} \left(A_{i,jp} - \frac{s_i^{out} s_{jp}^{in}}{W} + A_{jp,i} - \frac{s_i^{in} s_{jp}^{out}}{W} \right) \\ & - \frac{1}{W} \sum_{p=1}^{N_i - 1} \left(A_{i,ip} - \frac{s_i^{out} s_{ip}^{in}}{W} + A_{ip,i} - \frac{s_i^{in} s_{ip}^{out}}{W} \right), \end{aligned} \tag{4.4}$$

where $jp \in C_j$, $ip \in C_i$ and $ip \neq i$, N_j (N_i) is the number of nodes in community C_j (C_i) to which node j (i) belongs to, s_i^{in} (s_i^{out}) is the in-degree (out-degree) of vertex i, and $W = \sum_{i,j} A_{i,j}$. The first term in the right hand side of the above equation is the gain of modularity when node i moves to the community of node j, C_j, while the second term is the modularity gained when node i stays in its original community C_i. This process is sequentially and repeatedly applied to all nodes until there is no gain in modularity. At this stage, the partition of the network at the first level is obtained. Next, nodes in the same community after the first step are used to form several meta-nodes. The number of meta-nodes is equal to the number of current communities and the weights between any two meta-nodes are given by the sum of the weights of edges between nodes in the corresponding communities [1].

$$A_{new}(i_{new}, j_{new}) = \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} A_{i,j}, \tag{4.5}$$

where $i_{new}, j_{new} = 1, \cdots, tN$ with tN being the current number of meta-nodes, and N_i, N_j are the number of nodes in the two clusters. Note that since the network is directed, $A_{new}(i_{new}, j_{new}) \neq A_{new}(j_{new}, i_{new})$. The two steps are iterated until the modularity cannot increase anymore, and several levels of partitions are obtained at different resolutions.

```
Require: Weighted adjacency matrix $A \in (0,1)^{N \times N}$, nodes $1, \cdots, N$, initial community structure $C = \{\{1\}, \cdots, \{N\}\}$, $tN = N$.
Ensure: $M$ Communities.
Compute the modularity $Q$ of the network;
repeat
    $\Delta Q_{total} = 0$, $nchange = 0$;
    repeat
        for $h = 1$ to $tN$ do
            for $j = 1$ to $tN$ do
                The change of modularity $\Delta Q_{gen_j}$ when node $h$ is assigned to $C_j$;
            end for
            $j* = \arg\max_j \Delta Q_{gen_j}$;
            if $\Delta Q_{gen_{j*}} > 0$ then
                $C_{j*} = C_{j*} \cup v_h$;
            end if
        end for
        Compute the change of the modularity $\Delta Q_{total} = Q_{new} - Q$;
        if $\Delta Q_{total} > 0$ then
            $Q = Q_{new}$, $nchange = nchange + 1$;
        end if
    until $\Delta Q_{total} \leq 0$
    Nodes in the same community form new meta-nodes;
    $tN$ is equal to the current number of communities;
    Recompute the weighted matrix $A \in (0,1)^{tN \times tN}$;
until $nchange = 0$.
```

Figure 4.2: Algorithm 4: community detection of weighted networks

Low resolution indicates the number of nodes (meta-nodes) is small. The first level of partitioning (before the formation of meta-nodes) has the highest resolution. The modularity is always computed with respect to the initial graph topology, in this way, the two-step iterative procedure will not be trapped at a local maximum. This algorithm is summarized in Figure 4.2.

4.3.2 Algorithm for community detection for multiple subjects

One of the remaining challenging problems for the application of this algorithm is the group analysis when information from multiple subjects needs to be merged. As we mentioned in the background part, the standard approach to group analysis is based on either averaging the data or averaging the detected community structure from each subject. In this subsection, we propose a community detection algorithm for multiple subjects by integrating the information from each subject at the algorithm level, which can take both the inter-subject variability and commonality into account. To be specific, for the algorithm proposed above, we only take into account the change of modularity for moving one node to the community of another node for each subject. However, this activity leads to the change of modularity in all subjects. Therefore, we compute the change in modularity $\Delta Q^k_{gen_j}$ for subject k when assigning node i to the communities of all other nodes j, where $j = 1, \cdots, N$, $j \neq i$, and $k = 1, \cdots, L$ with L being the number of subjects, and try to maximize a common modularity function, i.e, the sum of the gain in modularity for all subjects $\Delta Q_{gen_j} = \sum_{k=1}^{L} \Delta Q^k_{gen_j}$. In this way, the effect of outliers is directly decreased at the algorithm level. The details are shown in Figure 4.3.

4.4 Results

In this section, we first illustrate the importance of edge direction information for revealing the real structure of a directed network by a simulated synthetic network. We then test the effectiveness of the proposed community detection algorithm for group analysis on both synthetic networks and real EEG data.

4.4.1 Directed vs. undirected networks

Most of the existing community detection algorithms are intended for the analysis of undirected and binary networks. However, many networks of interest, such as biological networks, are directed. One approach that has been commonly employed for community detection in directed networks is to directly apply the algorithms designed for undirected networks without considering the edge direction information [101]. To illustrate the importance of edge

```
Require: Weighted adjacency matrix $A^i \in (0,1)^{N \times N}$, $i = 1, \cdots, L$ with $L$ being the number of subjects, nodes $1, \cdots, N$, initial community structure $C = \{\{1\}, \cdots, \{N\}\}$, $tN = N$.
Ensure: $M$ Communities.
Compute the modularity $Q_i$ of subject $i$, $i = 1, \cdots, L$;
The modularity of the group is $Q = \sum_{i=1}^{L} Q_i$;
repeat
  $\Delta Q_{total} = 0$, $nchange = 0$;
  repeat
    for $h = 1$ to $tN$ do
      for $j = 1$ to $tN$ do
        The change of modularity $\Delta Q^i_{gen_j}$ when node $h$ is assigned to $C_j$ for subject $i$, $i = 1, \cdots, L$;
        The change of the whole group is $\Delta Q_{gen_j} = \sum_{i=1}^{L} \Delta Q^i_{gen_j}$;
      end for
      $j* = \arg\max_j \Delta Q_{gen_j}$;
      if $\Delta Q_{gen_{j*}} > 0$ then
        $C_{j*} = C_{j*} \cup v_h$;
      end if
    end for
    Compute the modularity of the whole group $Q_{new} = \sum_{i=1}^{L} Q^i_{new}$, $Q^i_{new}$ is the modularity of subject $i$;
    Compute the change of the modularity $\Delta Q_{total} = Q_{new} - Q$;
    if $\Delta Q_{total} > 0$ then
      $Q = Q_{new}$, $nchange = nchange + 1$;
    end if
  until $\Delta Q_{total} \leq 0$
  Nodes in the same community form new meta-nodes;
  $tN$ is equal to the current number of communities;
  Recompute the weighted matrix $A_i \in (0,1)^{tN \times tN}$ of subject $i$, $i = 1, \cdots, L$;
until $nchange = 0$.
```

Figure 4.3: Algorithm 5: community detection of multiple weighted networks

direction information for community detection, we generate a simulated directed network and employ two strategies to detect the community structure of this network. The first approach is to apply our proposed algorithm to the association matrix of the directed network A directly and the second approach is to ignore the edge direction information and apply the original Blondel's algorithm to the association matrix of the undirected network $\frac{1}{2}(A+A^T)$, where A^T is the transpose of matrix A. The simulated network consists of 24 nodes and 2 clusters, with each cluster having 12 nodes. Nodes 1 to 12 are in the same cluster, while the rest are in the other cluster. Each entry of the association matrix A is uniformly distributed between $[0, 1]$, which resembles the normalized DI value and A is not symmetric. The mean of connectivity strength in each cluster is 0.5. The mean inter-cluster connectivity strength from cluster 1 to 2 is 0.9, and from 2 to 1 is 0.1 . The standard deviation of the connectivity strengths in each cluster is 0.2. The community detection results are evaluated by computing the percentage of false discoveries or false discovery rate (FDR) F,

$$F = \sum_{i,j}^{N} \frac{O_{i,j} - M_{i,j}}{N^2} \tag{4.6}$$

where N is the number of nodes, $O_{i,j}$ is a binary matrix with entries equal to 1 if nodes i and j are in the same cluster. If nodes i and j are identified in the same cluster by the algorithm, then $M_{i,j} = 1$ and $M_{i,j} = 0$ otherwise. The community detection results are shown in Fig. 4.4. We observe that the proposed algorithm can detect the community structure of the network if we use the edge direction information (Figure 4.4(b)), whereas the original algorithm designed for undirected networks fails to capture the actual community structure (Figure 4.4(c)). Without loss of generality, we generate the network 100 times and the average false discovery rates for both approaches are obtained. If we consider the edge direction, the false discovery rate is 0.0423 and is 0.4789 otherwise. Therefore, our algorithm has good performance for community detection of directed networks and can reveal the real structural information of the directed network that conventional clustering algorithms for undirected networks cannot. In fact, when the association matrix of a network is strongly asymmetric or the total in-degree and out-degree for each node is significantly different, the difference of the community detection results between directed and undirected representation of the network will be relatively large, and the community detection results based on representing the network with directed weighted graphs can reveal the real community structure of the network.

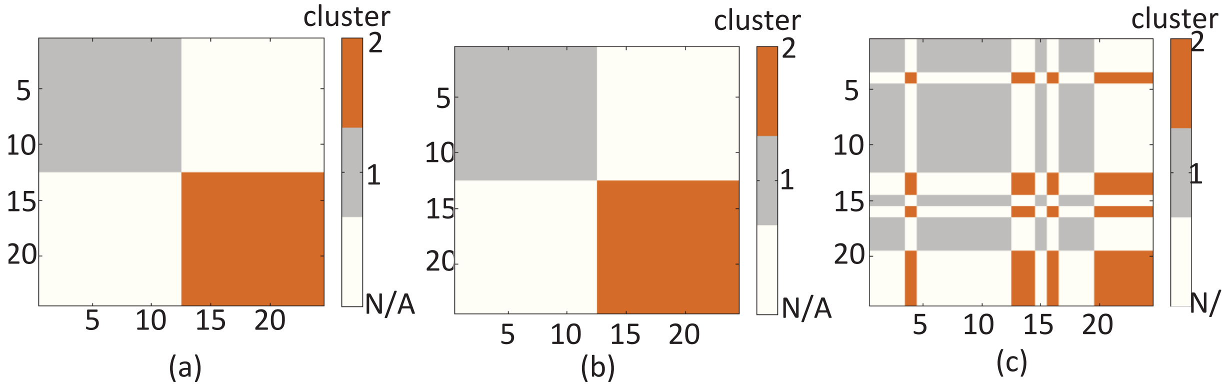


Figure 4.4: Community detection of different representation of the network. Community membership matrices for (a) Actual community structure. (b) Community structure obtained from the proposed algorithm which considers edge direction information. (c) Community structure obtained from Blondel's original algorithm which does not consider edge direction. White indicates that the corresponding node pairs are not in the same cluster (cluster N/A). Gray indicates that the corresponding node pairs are in cluster 1. Brown indicates that the corresponding node pairs are in cluster 2.

4.4.2 Group analysis on synthetic data

In this subsection, we evaluate the performance of the proposed group analysis algorithm by applying it to a single simulated network and a group of simulated networks. In addition, we compare our proposed group analysis algorithm with alternative group analysis approaches to show the effectiveness of our method. First, we test our algorithm on a directed network with 64 nodes and 4 clusters, with each cluster having 16 nodes. Each entry of the association matrix is uniformly distributed between $[0, 1]$, which resembles the normalized DI value. Means of intra-cluster connectivity strength in the four clusters are 0.3, 0.5, 0.7, and 0.9, respectively. The mean of inter-cluster connectivity is 0.15. To demonstrate the robustness

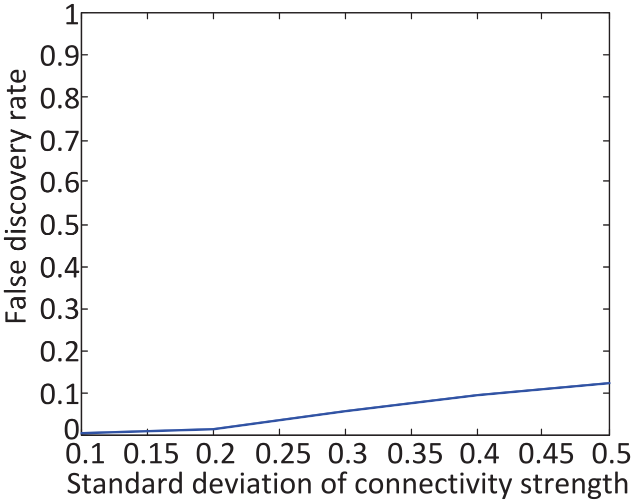


Figure 4.5: Average false discovery rate of the community detection algorithm for a simulated directed weighted network.

of the algorithm, the standard deviation of these distributions is modified from 0.1 to 0.5 with a step size of 0.1. Without loss of generality, we generate the network 50 times and the average false discovery rate F is obtained for different standard deviations. The result is shown in Figure 4.5. We observe that the false discovery rate grows with increasing standard deviation of the strength of connectivity for each edge. Even so, the maximum false discovery rate of our algorithm is 0.126, which is low and acceptable.

To test the effectiveness of the proposed group analysis method, we test Algorithm 5 (Figure 4.3) on a group of ten directed networks with the same community structure. Each network has 64 nodes and 4 clusters, with each cluster having 16 nodes. Each entry of the association matrix is uniformly distributed between $[0, 1]$. Means of intra-cluster connectivity strength in the four clusters are 0.3, 0.5, 0.7, and 0.9, respectively. The mean of inter-cluster connectivity is 0.15. For each network, the standard deviation of all the edge values are randomly chosen from $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$, which leads to the variation across the ten networks. Without loss of generality, we generate 50 simulations of networks to get the average false discovery rate. In addition, we compare our method with two standard approaches, i.e., VTS (average the association matrix) and IS (average the subject-

Table 4.1: Average false discovery rate for group analysis methods

	Proposed Algorithm	VTS	IS (average)	IS (voting)
FDR	0.0175	0.0225	0.6254	0.0231
Time (s)	141.1615	13.7438	141.9871	141.9870

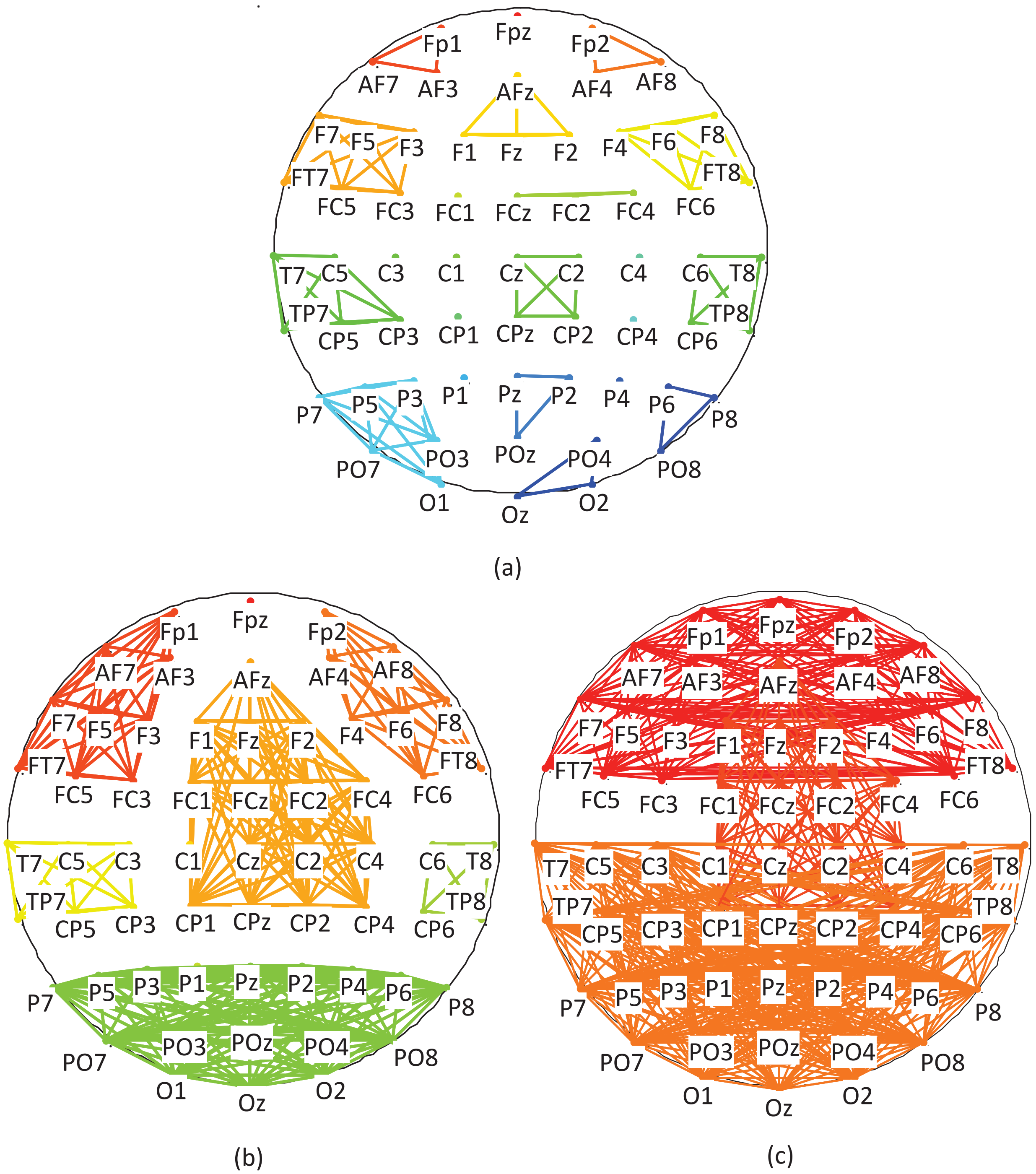
specific community structure or majority voting). The average false discovery rate F for each method is shown in Table 4.1. We observe that the proposed algorithm outperforms existing approaches with the lowest false discovery rate, which indicates our proposed method can provide promising results for group analysis problems. In addition, our proposed method is slightly computationally more efficient than IS but less efficient than VTS. Due to the fact that VTS is applied to averaged association matrix, it is only applied once instead of ten times. IS approach consumes a large amount of time in extracting the community structure for each subject before obtaining a common structure, while our proposed algorithm spends time in computing the change of modularity for all subjects when combing small clusters.

4.4.3 Group analysis on EEG Data

In this chapter, we examined EEG data from a study containing the error-related negativity (ERN) as described in Chapter 1. Previous work indicates there is increased information flow associated with ERN for the theta frequency band ($4 - 8$ Hz) and ERN time window $25 - 75$ ms for Error responses (ERN) compared to Correct responses (CRN) [122]. The EEG data collected are preprocessed by the spherical spline current source density (CSD) waveforms to sharpen event-related potential (ERP) scalp topographies and eliminate volume conduction [73]. In addition, the bandpass filter is used to obtain signals in the theta band. The effective connectivity quantified by the time-lagged DI is computed over a window corresponding to the ERN response ($0 - 100$ ms after the response), for all trials between each pair of 61 electrodes in the theta band. The time-lagged DI is averaged over the information flow within $10 - 20$ ms time delay [59], i.e. , $d = 5, \cdots, 10$ in equation (2.13).

Once the connectivity matrices for each response type and each subject are obtained, we use Algorithm 5 to identify the community structure of each response type. Since the proposed clustering approach is multi-level, we give the clustering results at all levels for each response type. The results are shown in Figure 4.6 and Figure 4.7. We observe that both the CRN and ERN show hierarchical structures. The optimal number of partition levels for the CRN is 3, while for the ERN it is 4. Since the modularity for each response type achieves its maxima at the top (final) level, we interpret the partition at the top (final) level for each response type. The third level of partition for CRN has three large clusters, i.e., frontal, parietal, and some nodes in the central regions, which indicates that the frontal and parietal regions exchange information through the central nodes. The fourth level of partition for ERN also has three large modules, i.e., left frontal-central region, right front-central-parietal region, and the parietal region, which indicates that the frontal and parietal regions work together when an error occurs, and the left and right side of the brain work differently. The differences between CRN and ERN also implies that ERN has more large-scale (across different regions) interactions compared to CRN, while the information flow in CRN is more local or less integrated. These results are aligned with previous work which indicates that there is increased information flow associated with ERN for the theta frequency band ($4-8$ Hz) and ERN time window $25-75$ ms for Error responses compared to correct responses in particular between mPFC and lPFC regions [74].

Though we have the community structures for both CRN and ERN at different resolutions, there are still two issues that need to be addressed. First, we need to determine whether the obtained clusters are significantly different from those from a random network. In order to address this issue, the modularity of the community structure from the actual data needs to be compared with modularity from random networks. Second, the modularity is known to have the problem of resolution limit for weighted graphs and it may not be the best criterion for the evaluation of the obtained community structure of a weighted network [123]. Therefore, we propose to determine the optimal level of partition for each

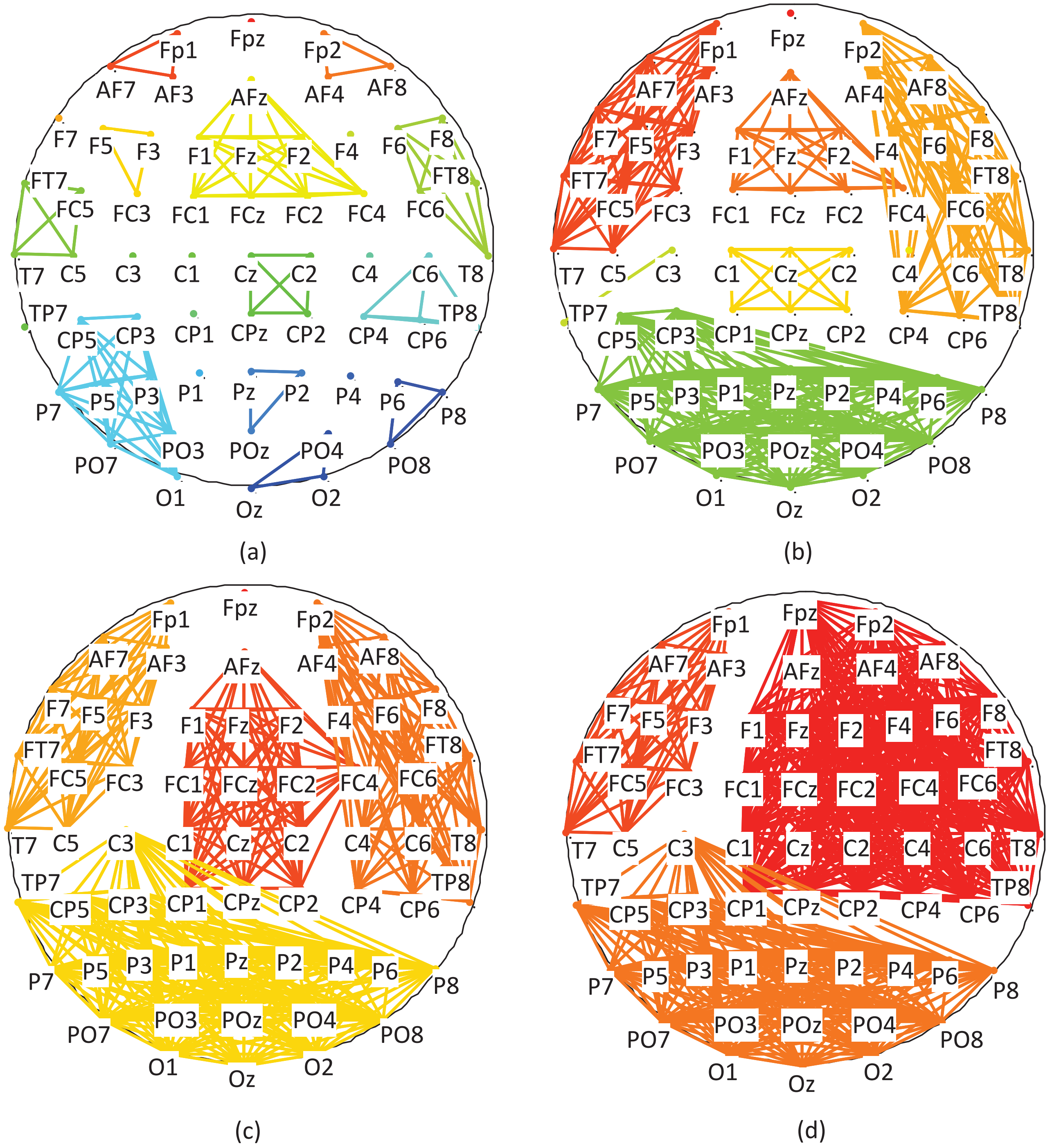

Fpz Fp1 Fp2
AF7 AF3 AFz AF4 AF8
F7 F5 F3 F1 Fz F2 F4 F6 F8
FT7 FT8
FC5 FC3 FC1 FCz FC2 FC4 FC6
T7 C5 C3 C1 Cz C2 C4 C6 T8
TP7 TP8
CP5 CP3 CP1 CPz CP2 CP4 CP6
P7 P5 P3 P1 Pz P2 P4 P6 P8
PO7 PO3 POz PO4 PO8
O1 Oz O2

(a)

(b)

(c)

Figure 4.6: Applying the multi-subject community detection algorithm to 10 subjects for CRN. (a) The first level of partition for CRN. (b) The second level of partition for CRN. (c) The third level of partition for CRN.


(a)

(b)

(c)

(d)

Figure 4.7: Applying the multi-subject community detection algorithm to 10 subjects for ERN. (a) The first level of partition for ERN. (b) The second level of partition for ERN. (c) The third level of partition for ERN. (d) The fourth level of partition for ERN.

Table 4.2: Comparison of Modularity for CRN with for random graphs

Modularity	Level 1	Level 2	Level 3
CRN	-0.0125	0.0678	0.0992
Averaged across random graphs	-0.0512	-0.0027	0.0151

Table 4.3: Comparison of Modularity for ERN with for random graphs

Modularity	Level 1	Level 2	Level 3	Level 4
ERN	-0.0161	0.0587	0.0813	0.087
Averaged across random graphs	-0.0526	-0.0053	0.0130	0.0170

response type in terms of information flow. To address the first problem, we compute the modularity from random graphs and compare them with the modularity obtained from the CRN and ERN. For comparison, random graphs that have the same number of nodes and weighted links with the original graphs are generated. To be specific, we keep the weight of each link unchanged and randomly assign the links [32]. In this way, the differences of modularity between real networks and random networks are mainly due to the structure of the network. We generate 100 sets of random graphs for each response type, which have the same level of partitions with the original one, i.e., for the CRN the number of levels is 3 and for the ERN it is 4. The modularity from the original networks of both response types and the modularity averaged across random networks at different levels of partitions are summarized in Tables 4.2 and 4.3. The modularity for the original networks is greater than that of random graphs. In addition, the null hypothesis, i.e. the modularity obtained from the CRN and ERN is not different from the modularity from random graphs at each level of partition, is rejected at $\alpha = 0.001$ significance level. Therefore, the clusters obtained for each response type reveal significant differences from the equivalent random networks, which indicates significant modular structures for both response types.

Modularity may sometimes fail to determine the best partition for a weighted network. Therefore, we propose to determine the best partition in terms of information flow. We choose the partition which has high information flow inside the module and low information flow between modules to represent the community structure of the whole group. In this way, we can also reveal the community at different resolutions in terms of information flow. Suppose we have K clusters at partition level l, the ratio between intra-cluster and inter-cluster information flow is given as follows,

$$\begin{aligned} WIFL_k &= \frac{1}{N_k^2} \sum_{i,j \in C_k} TA_{i,j}; \\ BIFL_k &= \frac{1}{K-1} \sum_{p=1, p \neq k}^{K} \sum_{i \in C_k, j \in C_p} \frac{TA_{i,j} + TA_{j,i}}{2 \times N_k \times N_p}; \\ R &= \frac{1}{K} \sum_{k=1}^{K} \frac{WIFL_k}{BIFL_k}, \end{aligned} \tag{4.7}$$

where N_k is the number of nodes in cluster C_k, $k = 1, \cdots, K$, $TA = \frac{1}{L} \sum_{m=1}^{L} A^m$ where A^m is the association matrix of the mth subject in the group and L is the number of subjects in that group (In our analysis $L = 10$), $WIFL_k$ is the information flow within cluster k, $BIFL_k$ is the averaged inter-cluster information flow between cluster k and other clusters, while R is the averaged ratio of intra-cluster information flow to inter-cluster information flow over clusters. We compute the information flow ratio for both the CRN and ERN response types and compare this ratio with the one obtained from random assignment of the cluster labels to each node for each response type. To be specific, for each level of partition, we keep the number of clusters and the number of elements in each cluster unchanged, but randomly assign each node a community label and compute the information flow ratio at each level. In this way, the differences of information flow within and among clusters between real networks and random networks at each specific level of partition are mainly due to the level-specific structure rather than the number of clusters and the number of nodes in each cluster. For each level of partition, we repeat the random assignment 100 times and the averaged information flow ratio across these random networks are shown in Figure 4.8. We observe that at each level of partition the information flow ratio obtained from original community structure is higher than the ratio averaged across random graphs. In addition, for each level of partition, the hypothesis that there is no difference between the information flow ratio of the original community structure and a randomly assigned community structure is rejected at $\alpha = 0.001$ significance level. Therefore, the obtained structures at each level of partition for brain networks are significantly different from random graphs, which is aligned with the results when using modularity for significance testing. We also observe that ERN has a slightly higher information flow ratio, i.e., stronger within cluster information exchange compared to intra-cluster information exchange, which indicates strong local activity or functional segregation compared to CRN. On the other hand, for the comparison of best representation of CRN and ERN in terms of information flow, we choose the second level of partition for both groups (Figure 4.6(b) and Figure 4.7(b)), because the information flow ratio of both response types reaches the local maxima at level 2. We observe that at this level the ERN group and the CRN group have different clustering (Figure 4.6(b) and Figure 4.7(b)) in the frontal and central-parietal regions. The frontal and central-parietal regions around the cerebral midline are not in the same cluster for ERN, which shows the functional specialization of the frontal and central-parietal regions around the cerebral midline whereas for CRN that specialization does not exist. On the other hand, the right lateral frontal and central-parietal regions are in the same cluster for ERN, which is contrary to the CRN. These results are aligned with the previous work in [75], which shows that error processing is controlled by the communication between the lateral prefrontal cortex and medial prefrontal cortex.

4.5 Conclusions

In this chapter, we propose a method to identify modules in the effective brain network. In order to achieve this goal, first, we applied the directed information measure to EEG data involving a study of ERN to obtain the association matrix of the network. Directed information can effectively detect the nonlinear causal relationship between EEG signals, which is the basis for obtaining a reliable community structure. In addition, we extended

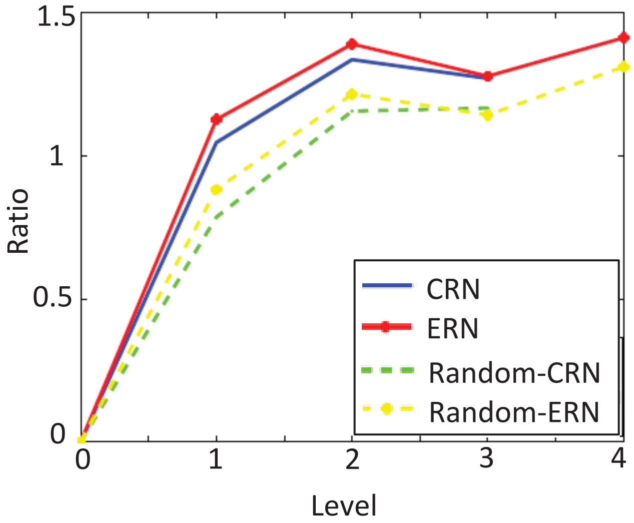


Figure 4.8: The information flow ratio of both response types and their corresponding random networks.

a modularity based community detection algorithm proposed by Blondel et al. to weighted directed networks. Compared to current community detection methods, our proposed algorithm discovers the actual structure of a directed network by employing the edge direction information. Finally, we proposed a group analysis method to obtain a common community structure across subjects to address the problem of variability across subjects. We extended the idea of modularity optimization from a single subject to a group of subjects. This strategy decreases the effect of outliers without making any assumptions about the data. The proposed group analysis method has higher accuracy in community detection than standard approaches, such as VTS and IS. It is also applied to EEG data and is shown to discriminate between error and correct responses in terms of the community structures obtained.

The proposed algorithm is based on the optimization of modularity. However, the modularity optimization encounters the problem of resolution limit, which indicates that it may miss detecting clusters whose size is comparatively small to the whole graph [39]. Therefore, it would be of interest to investigate and extend methods that do not depend on modularity optimization, e.g. random walk, to find the common communities across a group of weighted directed networks. In addition, one can consider overlapping communities by extending the current framework to consider multiple community memberships. Finally, this work can be extended to dynamic networks and detect the change of modules across time and frequency [102, 124].

Chapter 5
SUMMARY AND FUTURE WORK

This thesis discusses the problem of revealing the underlying structure of complex networks from multivariate time series. In particular, we focus on quantifying the pairwise causal relationships, inferring the topological structure, and detecting functional modules of a complex network.

5.1 Summary

The first part of this thesis focuses on finding a proper measure to quantify the interaction and causality between two random processes. To achieve this goal, we tried to answer the following questions: (1) What are some suitable measures? (2) What are the problems in the implementation and application to real data with limited sample size? Motivated by these questions, we introduced directed information and illustrated its relationship with existing measures, such as Granger causality, mutual information, and transfer entropy. The implementation of DI requires large sample sizes and is very time consuming. To reduce the computational complexity of computing DI while still quantifying the causal dependencies, we derived simplified expressions of DI, i.e., the time-lagged directed information and modified directed information. We also showed the relationship between modified directed information and transfer entropy. The proposed expressions are shown to be more efficient than the original DI in terms of computational complexity and capture more causal dependencies than short time DI without considering time lags. Moreover, we compared the performance of DI with model based measures such as Granger causality on different realistic signal models, and DI is found to be applicable to a wider range of signal types. In addition, we developed a new directed information estimation method based on multi-information and provided a quantitative comparison of various DI estimation methods.

Quantifying the pairwise causal relationships does not reflect the true topological structure of a system. Two nodes which have a high DI value may influence each other indirectly through a third node. In the second part of our proposed work, we pointed out the drawback of only using directed information for network inference, i.e., using DI alone can not distinguish direct causality from indirect causality. Therefore, the conditional directed information is applied to address this problem. Moreover, time-lagged conditional directed information and the modified conditional directed information are proposed to reduce the computational complexity. Three algorithms are proposed for network inference and the results for both linear and nonlinear networks are shown. The results of our algorithms on simulated data show that the combination of modified DI and CDI can effectively increase the accuracy of network inference.

The proposed network inference algorithms are able to detect the organization of any collection of triplets in a network, but may fail to reflect the dynamics between a large group of nodes or signals because of the limitation of computational complexity. In the last part of our proposed work, we aim to simplify the inferred networks as well as to determine the functional modules that result in the observed connectivity patterns by developing community detection algorithms for directed networks. We introduced a hierarchical community detection algorithm to discover the modules in a complex weighted directed network. In addition, we proposed a group analysis method to obtain a common community structure across subjects to address the problem of variability across subjects in neurophysiological study. We applied the proposed framework to an EEG data set collected during a study of cognitive control networks in the brain. In particular, we looked at a data set of subjects involved in error-processing. The proposed method is applied to both synthetic data and EEG data and is shown to discriminate between error and correct responses in terms of the community structures obtained.

5.2 Future work

There are still remaining challenges with the application of causality measures to real multivariate data and the inference of network structure. Some of these challenges include:

- The extension and evaluation of the proposed community detection algorithm. Although the proposed community detection algorithm has promising performance, there are some issues remaining to be addressed. First, both theoretical and practical development may be needed to explain which representation (directed or undirected graph) is more suitable for revealing the actual module structure of the network in different scenarios. For example, when the association matrix quantified by effective connectivity is approximately symmetric, the results of community detection based on functional and effective connectivity will be similar to each other. Second, one node can belong to more than one community, therefore, it might be interesting to extend the algorithm to uncover the overlapping community structure of a network. Third, the community structure of the brain network may change over time, therefore, community structure for a group of time-dependent, multi-scale networks are needed. Mucha *et al.* developed a generalized framework for network quality function that detects the community structure in time-dependent multi-scale and multiplex networks [125]. This approach may be extended to find the community structure of the effective brain network over time.

- The group analysis proposed in Chapter 4 may also be applied to the network inference problem. The proposed network inference algorithm was applied to synthetic data, but when applied to the EEG data, the results are hard to interpret because of the variability across subjects. Therefore, instead of reconstructing the brain network for each subject, it is possible to reconstruct one network for the whole group. For example, we can keep the common strong connections and remove the common indirect connections for a group. In addition, the accuracy of the inference algorithm can also be improved by considering more complex scenarios. Node X may influence Y through nodes Z and W and it is hard to remove the false positive connection between X and W when only computing $DI(\mathbf{X} \rightarrow \mathbf{Y}||\mathbf{Z})$ or $DI(\mathbf{X} \rightarrow \mathbf{Y}||\mathbf{W})$, because the conditional directed information will equal to zero when considering both nodes, i.e., $DI(\mathbf{X} \rightarrow \mathbf{Y}||\mathbf{ZW})$. However, considering more nodes means higher dimensionality and computational complexity compared to only considering one node. Therefore, more reliable estimation of conditional directed information is needed. This problem can be addressed by either using parametric density models or improving existing mutual information and entropy estimators. In addition, the performance of MCDI and TLCDI also depend on the estimated order of the model. Recently, Faes *et al.* proposed a sequential procedure to determine the embedding dimension of multivariate series [80]. This method is based on an information-theoretic technique and shows promising performance for various signal models, which may be extended to MCDI and TLCDI computation in the future.

APPENDICES

Appendix A
CONDITIONAL DIRECTED INFORMATION IN TWO GENERAL TRIVARIATE MODELS

Two general models for triplets connected in a hub and chain pattern are given respectively below:

Model 1*: Hub Model*

$$\begin{aligned} X_n &= f(X_{n-p_1}, \cdots, X_{n-1}) + u_n, \\ Y_n &= g(X_{n-p_2}, \cdots, X_{n-1}, Y_{n-p_3}, \cdots, Y_{n-1}) + v_n, \\ Z_n &= h(X_{n-p_4}, \cdots, X_{n-1}, Z_{n-p_5}, \cdots, Z_{n-1}) + w_n, \end{aligned} \tag{A.1}$$

where $\mathbf{X}$ causes $\mathbf{Y}$ and $\mathbf{Z}$ ($\mathbf{X}$ is the hub), respectively. $f(\cdot)$, $g(\cdot)$, $h(\cdot)$ are three different functions, p_2 (p_4) is the maximum time lag between $\mathbf{X}$ and $\mathbf{Y}$ ($\mathbf{X}$ and $\mathbf{Z}$), p_1, p_3, and p_5 are the order of time series $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ respectively. u_n, v_n and w_n are white Gaussian noises and are independent of the signals.

Model 2*: Chain Model*

$$\begin{aligned} X_n &= f(X_{n-p_1}, \cdots, X_{n-1}) + u_n, \\ Y_n &= g(X_{n-p_2}, \cdots, X_n, Y_{n-p_3}, \cdots, Y_{n-1}) + v_n, \\ Z_n &= h(Y_{n-p_4}, \cdots, Y_n, Z_{n-p_5}, \cdots, Z_{n-1}) + w_n, \end{aligned} \tag{A.2}$$

where $\mathbf{X}$ causes $\mathbf{Y}$, and $\mathbf{Y}$ causes $\mathbf{Z}$ ($\mathbf{Y}$ is the intermediate node). $f(\cdot)$, $g(\cdot)$, $h(\cdot)$ are three different functions, p_2 (p_4) is the maximum time lag between $\mathbf{X}$ and $\mathbf{Y}$ ($\mathbf{Y}$ and $\mathbf{Z}$), p_1, p_3, and p_5 are the order of time series $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ respectively. u_n, v_n and w_n are white Gaussian noises and are independent of the signals.

We compare two causally conditional directed information for both models, i.e., $DI(Y^N \rightarrow Z^N || X^N)$ and $DI(X^N \rightarrow Z^N || Y^N)$.

- For hub case:

$$
\begin{aligned}
DI(Y^N \to Z^N||X^N) &= \sum_{n=1}^{N} I(Y^n; Z_n|Z^{n-1}X^n) \\
&= \sum_{n=1}^{N} [H(Z_n|Z^{n-1}X^n) - H(Z_n|Y^n Z^{n-1}X^n)] \\
&= \sum_{n=1}^{N} [H(h(X_{n-p_4}, \cdots, X_n, Z_{n-p_5}, \cdots, Z_{n-1}) + w_n|Z^{n-1}X^n) \\
&- H(h(X_{n-p_4}, \cdots, X_n, Z_{n-p_5}, \cdots, Z_{n-1}) + w_n|X^n Z^{n-1}Y^n)] \\
&= \sum_{n=1}^{N} [H(w_n|Z^{n-1}X^n) - H(w_n|X^n Z^{n-1}Y^n)] \\
&= \sum_{n=1}^{N} [H(w_n) - H(w_n)] = 0.
\end{aligned}
\tag{A.3}
$$

where the forth equality comes from that $H(Y|X_1, \cdots, X_n) = H(Z|X_1, \cdots, X_n)$ when $Y = \sum_{i=1}^{n} X_i + Z$ [126], and the fifth equality is due to the fact that noise w_n is independent of signals.

$$
\begin{aligned}
DI(X^N \to Z^N||Y^N) &= \sum_{n=1}^{N} I(X^n; Z_n|Z^{n-1}Y^n) \\
&= \sum_{n=1}^{N} [H(Z_n|Z^{n-1}Y^n) - H(Z_n|X^n Z^{n-1}Y^n)], \\
&\geq 0.
\end{aligned}
\tag{A.4}
$$

where the inequality is true since conditioning reduces entropy and the equality holds only when X^n and Y^n are identical.

Therefore,

$$DI(X^N \to Z^N||Y^N) > 0 = DI(Y^N \to Z^N||X^N). \tag{A.5}$$

- For chain case:

$$
\begin{aligned}
DI(Y^N \to Z^N||X^N) &= \sum_{n=1}^{N} I(Y^n; Z_n|Z^{n-1}X^n) \\
&= \sum_{n=1}^{N} [H(Z_n|Z^{n-1}X^n) - H(Z_n|X^n Z^{n-1} Y^n)] \\
&\geq 0.
\end{aligned}
\tag{A.6}
$$

$$
\begin{aligned}
DI(X^N \to Z^N||Y^N) &= \sum_{n=1}^{N} I(X^n; Z_n|Z^{n-1}Y^n) \\
&= \sum_{n=1}^{N} [H(Z_n|Z^{n-1}Y^n) - H(Z_n|Y^n Z^{n-1} X^n)] \\
&= \sum_{n=1}^{N} [H(h(Y_{n-p_4}, \cdots, Y_n, Z_{n-p_5}, \cdots, Z_{n-1}) + w_n|Z^{n-1}Y^n) \\
&- H(h(Y_{n-p_4}, \cdots, Y_n, Z_{n-p_5}, \cdots, Z_{n-1}) + w_n|Y^n Z^{n-1} X^n)] \\
&= \sum_{n=1}^{N} [H(w_n|Z^{n-1}Y^n) - H(w_n|Y^n Z^{n-1} X^n)] \\
&= \sum_{n=1}^{N} [H(w_n) - H(w_n)] = 0.
\end{aligned}
\tag{A.7}
$$

Therefore,

$$
DI(Y^N \to Z^N||X^N) > 0 = DI(X^N \to Z^N||Y^N). \tag{A.8}
$$

Appendix B

COMPUTATION OF TIME-LAGGED CONDITIONAL DIRECTED INFORMATION

In this Appendix, we compute the time-lagged conditional directed information for both hub and chain models.

Model 1: $\mathbf{X}$ interacts with $\mathbf{Y}$ and $\mathbf{Z}$ with different time delays,

$$\begin{aligned} X_i &= u_i, \\ Y_i &= bX_{i-m} + v_i, \\ Z_i &= cX_{i-n} + w_i. \end{aligned} \tag{B.1}$$

Therefore,

$$\begin{aligned} &E(X_k) = 0, Var(X_k) = \sigma^2, Cov(X_k X_{k+1}) = 0, \\ &E(Y_k) = 0, Var(Y_k) = (b^2+1)\sigma^2, Cov(Y_k Y_{k+1}) = 0, \\ &E(Z_k) = 0, Var(Z_k) = (c^2+1)\sigma^2, Cov(Z_k Z_{k+1}) = 0, \end{aligned}$$

$$Cov(X_{k-l}Y_k) = E[X_{k-l}(bX_{k-m} + v_k)] = \begin{cases} b\sigma^2, \ l = m \\ 0, \ \text{otherwise} \end{cases}$$

$$Cov(X_{k-l}Y_{k+1}) = E[X_{k-l}(bX_{k+1-m} + v_{k+1})] = \begin{cases} b\sigma^2, \ l = m-1 \\ 0, \ \text{otherwise} \end{cases}$$

$$Cov(X_{k-l}Z_{k+n-m}) = E[X_{k-l}(cX_{k+n-m-n} + w_{k+n-m})] = \begin{cases} c\sigma^2, \ l = m \\ 0, \ \text{otherwise} \end{cases}$$

$$Cov(X_{k-l}Z_{k+n-m+1}) = E[X_{k-l}(cX_{k+n-m+1-n} + w_{k+n-m+1})] = \begin{cases} c\sigma^2, \ l = m-1 \\ 0, \ \text{otherwise} \end{cases}$$

$$
\begin{aligned}
&Cov(X_{k-l+1}Y_k) = E[X_{k-l+1}(bX_{k-m}+v_k)] = \begin{cases} b\sigma^2, \ l=m+1 \\ 0, \ \text{otherwise} \end{cases} \\
&Cov(X_{k-l+1}Y_{k+1}) = E[X_{k-l}(bX_{k+1-m}+v_{k+1})] = \begin{cases} b\sigma^2, \ l=m \\ 0, \ \text{otherwise} \end{cases} \\
&Cov(X_{k-l+1}Z_{k+n-m}) = E[X_{k-l+1}(cX_{k+n-m-n}+w_{k+n-m})] = \begin{cases} c\sigma^2, \ l=m+1 \\ 0, \ \text{otherwise} \end{cases} \\
&Cov(X_{k-l+1}Z_{k+n-m+1}) = E[X_{k-l+1}(cX_{k+n-m+1-n}+w_{k+n-m+1})] = \begin{cases} c\sigma^2, \ l=m \\ 0, \ \text{otherwise} \end{cases} \\
&Cov(Y_kZ_{k+n-m}) = E[(bX_{k-m}+v_k)(cX_{k+n-m-n}+w_{k+n-m})] = bc\sigma^2, \\
&Cov(Y_kZ_{k+n-m+1}) = E[(bX_{k-m}+v_k)(cX_{k+n-m+1-n}+w_{k+n-m+1})] = 0, \\
&Cov(Y_{k+1}Z_{k+n-m}) = E[(bX_{k+1-m}+v_{k+1})(cX_{k+n-m-n}+w_{k+n-m})] = 0, \\
&Cov(Y_{k+1}Z_{k+n-m+1}) = E[(bX_{k+1-m}+v_{k+1})(cX_{k+n-m+1-n}+w_{k+n-m+1})] = bc\sigma^2.
\end{aligned}
\tag{B.2}
$$

Then we get the covariance matrix and obtain the CDI values for different values of l.

(1) For $l=m$,

$$
\begin{aligned}
&|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}Z_{k+n-m+1}| = \sigma^{12}, |X_{k-l}Y_k| = \sigma^4, \\
&|X_{k-l}Z_{k+n-m}| = \sigma^4, |X_{k-l}X_{k-l+1}Z_{k+n-m}Z_{k+n-m+1}| = \sigma^8, \\
&|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}| = \sigma^{10}, |X_{k-l}Y_kZ_{k+n-m}| = \sigma^6, \\
&|X_{k-l}X_{k-l+1}Z_{k+n-m}| = \sigma^6, \\
&\quad CDI_k(Y_kY_{k+1} \to Z_{k+n-m}Z_{k+n-m+1}|X_{k-l}X_{k-l+1}) = 0.
\end{aligned}
\tag{B.3}
$$

(2) For $l = m - 1$,

$$|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}Z_{k+n-m+1}| = (b^2+c^2+1)\sigma^{12}, |X_{k-l}Y_k| = (b^2+1)\sigma^4,$$
$$|X_{k-l}Z_{k+n-m}| = (c^2+1)\sigma^4, |X_{k-l}X_{k-l+1}Z_{k+n-m}Z_{k+n-m+1}| = (c^2+1)\sigma^8,$$
$$|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}| = (b^2+c^2+1)\sigma^{10}, |X_{k-l}Y_kZ_{k+n-m}| = (b^2+c^2+1)\sigma^6,$$
$$|X_{k-l}X_{k-l+1}Z_{k+n-m}| = (c^2+1)\sigma^6,$$
$$CDI_k(Y_kY_{k+1} \to Z_{k+n-m}Z_{k+n-m+1}|X_{k-l}X_{k-l+1}) = \frac{1}{2}\log\frac{(b^2+1)(c^2+1)}{b^2+c^2+1}.$$

(3) For $l = m + 1$,

$$|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}Z_{k+n-m+1}| = (b^2+c^2+1)\sigma^{12}, |X_{k-l}Y_k| = (b^2+1)\sigma^4,$$
$$|X_{k-l}Z_{k+n-m}| = (c^2+1)\sigma^4, |X_{k-l}X_{k-l+1}Z_{k+n-m}Z_{k+n-m+1}| = (c^2+1)\sigma^8,$$
$$|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}| = (b^2+1)\sigma^{10}, |X_{k-l}Y_kZ_{k+n-m}| = (b^2+c^2+1)\sigma^6,$$
$$|X_{k-l}X_{k-l+1}Z_{k+n-m}| = \sigma^6,$$
$$CDI_k(Y_kY_{k+1} \to Z_{k+n-m}Z_{k+n-m+1}|X_{k-l}X_{k-l+1}) = \log\frac{(b^2+1)(c^2+1)}{b^2+c^2+1}. \tag{B.4}$$

(4) For $l \neq m-1, m, m+1$,

$$|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}Z_{k+n-m+1}| = (b^2+c^2+1)^2\sigma^{12}, |X_{k-l}Y_k| = (b^2+1)\sigma^4,$$
$$|X_{k-l}Z_{k+n-m}| = (c^2+1)\sigma^4, |X_{k-l}X_{k-l+1}Z_{k+n-m}Z_{k+n-m+1}| = (c^2+1)^2\sigma^8,$$
$$|X_{k-l}X_{k-l+1}Y_kY_{k+1}Z_{k+n-m}| = (b^2+1)(b^2+c^2+1)\sigma^{10},$$
$$|X_{k-l}Y_kZ_{k+n-m}| = (b^2+c^2+1)\sigma^6, |X_{k-l}X_{k-l+1}Z_{k+n-m}| = (c^2+1)\sigma^6,$$
$$CDI_k(Y_kY_{k+1} \to Z_{k+n-m}Z_{k+n-m+1}|X_{k-l}X_{k-l+1}) = \log\frac{(b^2+1)(c^2+1)}{b^2+c^2+1}. \tag{B.5}$$

The time-lagged conditional directed information between **Y** and **Z** given **X** can be simplified as:

$$CDI_k(Y_kY_{k+1} \to Z_{k+n-m}Z_{k+n-m+1}|X_{k-l}X_{k-l+1})$$
$$= \begin{cases} \frac{1}{2}\log\frac{(b^2+1)(c^2+1)}{(b^2+c^2+1)},\ l = m-1 \\ 0,\ l = m \\ \log\frac{(b^2+1)(c^2+1)}{(b^2+c^2+1)},\ l \geq m+1 \text{ or } l < m-1 \end{cases} \tag{B.6}$$

Model 2: $\mathbf{X}$ causes $\mathbf{Y}$, and $\mathbf{Y}$ causes $\mathbf{Z}$,

$$\begin{aligned} X_i &= u_i, \\ Y_i &= bX_{i-m} + v_i, \\ Z_i &= cY_{i-(n-m)} + w_i. \end{aligned} \tag{B.7}$$

.

$$E(X_k) = 0, Var(X_k) = \sigma^2, Cov(X_kX_{k+1}) = 0,$$

$$E(Y_k) = 0, Var(Y_k) = (b^2+1)\sigma^2, Cov(Y_kY_{k+1}) = 0,$$

$$E(Z_k) = 0, Var(Z_k) = (c^2+1)\sigma^2, Cov(Z_kZ_{k+1}) = 0,$$

$$Cov(X_kY_{k+l}) = E[X_k(bX_{k+l-m} + v_{k+l})] = \begin{cases} b\sigma^2,\ l = m \\ 0,\ \text{Otherwise} \end{cases}$$

$$Cov(X_kY_{k+l+1}) = E[X_k(bX_{k+l+1-m} + v_{k+l+1})] = \begin{cases} b\sigma^2,\ l = m-1 \\ 0,\ \text{Otherwise} \end{cases}$$

$$Cov(X_kZ_{k+m+n}) = E[X_k(cY_{k+n+m-n} + w_{k+n+m})] = bc\sigma^2,$$

$$Cov(X_kZ_{k+m+n+1}) = E[X_k(cX_{k+n+m+1-n} + w_{k+n+m+1})] = 0,$$

$$Cov(X_{k+1}Y_{k+l}) = E[X_{k+1}(bX_{k+l-m} + v_{k+l})] = \begin{cases} b\sigma^2,\ l = m+1 \\ 0,\ \text{Otherwise} \end{cases}$$

$$Cov(X_{k+1}Y_{k+l+1}) = E[X_{k+1}(bX_{k+l+1-m} + v_{k+l+1})] = \begin{cases} b\sigma^2,\ l = m \\ 0,\ \text{Otherwise} \end{cases}$$

$$
\begin{aligned}
&Cov(Y_{k+l}Z_{k+n+m}) = \begin{cases} c(b^2+1)\sigma^2,\ l=m \\ 0,\ \text{Otherwise} \end{cases} \\
&Cov(Y_{k+l}Z_{k+n+m+1}) = \begin{cases} c(b^2+1)\sigma^2,\ l=m+1 \\ 0,\ \text{Otherwise} \end{cases} \\
&Cov(Y_{k+l+1}Z_{k+n+m}) = \begin{cases} c(b^2+1)\sigma^2,\ l=m-1 \\ 0,\ \text{Otherwise} \end{cases} \\
&Cov(Y_kZ_{k+n-m+1}) = E\begin{cases} c(b^2+1)\sigma^2,\ l=m \\ 0,\ \text{Otherwise} \end{cases}
\end{aligned} \tag{B.8}
$$

Then we get the covariance matrix and obtain the CDI values for different values of l.

(1) For $l=m$,

$$
\begin{aligned}
&|X_kX_{k+1}Y_{k+l}Y_{k+l+1}Z_{k+n+m}Z_{k+n+m+1}| = \sigma^{12}, \\
&|Y_{k+l}Z_{k+m+n}| = (b^2+1)\sigma^4, \\
&|X_kY_{k+l}| = \sigma^4, \\
&|Y_{k+l}Y_{k+l+1}Z_{k+n+m}Z_{k+n+m+1}| = (b^2+1)^2\sigma^8, \\
&|X_kX_{k+1}Y_{k+l}Y_{k+l+1}Z_{k+n+m}| = \sigma^{10}, |X_kY_{k+l}Z_{k+n+m}| = \sigma^6, \\
&|Y_{k+l}Y_{k+l+1}Z_{k+n+m}| = (b^2+1)\sigma^6, \\
&\quad CDI_k(Y_kY_{k+1} \to Z_{k+n-m}Z_{k+n-m+1}|X_{k-l}X_{k-l+1}) = 0.
\end{aligned} \tag{B.9}
$$

(2) For $l = m - 1$,

$$
\begin{aligned}
&|X_k X_{k+1} Y_{k+l} Y_{k+l+1} Z_{k+n+m} Z_{k+n+m+1}| = (b^2+1)(c^2+1)\sigma^{12},\\
&|Y_{k+l} Z_{k+m+n}| = (b^2+1)(b^2c^2+c^2+1)\sigma^4,\\
&|X_k Y_{k+l}| = (b^2+1)\sigma^4,\\
&|Y_{k+l} Y_{k+l+1} Z_{k+n+m} Z_{k+n+m+1}| = (b^2+1)^2(b^2c^2+c^2+1)\sigma^8,\\
&|X_k X_{k+1} Y_{k+l} Y_{k+l+1} Z_{k+n+m}| = (b^2+1)\sigma^{10},\\
&|X_k Y_{k+l} Z_{k+n+m}| = (b^2+1)(c^2+1)\sigma^6,\\
&|Y_{k+l} Y_{k+l+1} Z_{k+n+m}| = (b^2+1)^2\sigma^6,\\
&\quad CDI_k(Y_k Y_{k+1} \to Z_{k+n-m} Z_{k+n-m+1} | X_{k-l} X_{k-l+1}) = \log \frac{(b^2c^2+c^2+1)}{c^2+1}.
\end{aligned}
\tag{B.10}
$$

(3) For $l = m + 1$,

$$
\begin{aligned}
&|X_k X_{k+1} Y_{k+l} Y_{k+l+1} Z_{k+n+m} Z_{k+n+m+1}| = (b^2+1)(c^2+1)\sigma^{12},\\
&|Y_{k+l} Z_{k+m+n}| = (b^2+1)(b^2c^2+c^2+1)\sigma^4,\\
&|X_k Y_{k+l}| = (b^2+1)\sigma^4,\\
&|Y_{k+l} Y_{k+l+1} Z_{k+n+m} Z_{k+n+m+1}| = (b^2+1)^2(b^2c^2+c^2+1)\sigma^8,\\
&|X_k X_{k+1} Y_{k+l} Y_{k+l+1} Z_{k+n+m}| = (b^2+1)(c^2+1)\sigma^{10},\\
&|X_k Y_{k+l} Z_{k+n+m}| = (b^2+1)(c^2+1)\sigma^6,\\
&|Y_{k+l} Y_{k+l+1} Z_{k+n+m}| = (b^2+1)^2(b^2c^2+c^2+1)\sigma^6,\\
&\quad CDI_k(Y_k Y_{k+1} \to Z_{k+n-m} Z_{k+n-m+1} | X_{k-l} X_{k-l+1}) = \frac{1}{2}\log \frac{(b^2c^2+c^2+1)}{c^2+1}.
\end{aligned}
\tag{B.11}
$$

(4) For $l \neq m-1, m, m+1$,

$$\begin{aligned}
&|X_k X_{k+1} Y_{k+l} Y_{k+l+1} Z_{k+n+m} Z_{k+n+m+1}| = (b^2+1)^2(c^2+1)^2\sigma^{12}, \\
&|Y_{k+l} Z_{k+m+n}| = (b^2+1)(b^2c^2+c^2+1)\sigma^4, \\
&|X_k Y_{k+l}| = (b^2+1)\sigma^4, \\
&|Y_{k+l} Y_{k+l+1} Z_{k+n+m} Z_{k+n+m+1}| = (b^2+1)^2(b^2c^2+c^2+1)^2\sigma^8, \\
&|X_k X_{k+1} Y_{k+l} Y_{k+l+1} Z_{k+n+m}| = (b^2+1)^2(c^2+1)\sigma^{10}, \\
&|X_k Y_{k+l} Z_{k+n+m}| = (b^2+1)(c^2+1)\sigma^6, \\
&|Y_{k+l} Y_{k+l+1} Z_{k+n+m}| = (b^2+1)^2(b^2c^2+c^2+1)\sigma^6, \\
&\quad CDI_k(Y_k Y_{k+1} \to Z_{k+n-m} Z_{k+n-m+1} | X_{k-l} X_{k-l+1}) = \log \frac{(b^2c^2+c^2+1)}{c^2+1}.
\end{aligned} \tag{B.12}$$

The time-lagged conditional directed information between $\mathbf{X}$ and $\mathbf{Z}$ given $\mathbf{Y}$ can be simplified as:

$$DI_k(X_k X_{k+1} \to Z_{k+m+n} Z_{k+m+n+1} | Y_{k+l} Y_{k+l+1}) = \begin{cases} \log \frac{b^2c^2+c^2+1}{(c^2+1)}, \ l \leq m-1 \text{ or } l > m+1 \\ 0, \ l = m \\ \frac{1}{2} \log \frac{(b^2c^2+c^2+1)}{(c^2+1)}, \ l = m+1 \end{cases} \tag{B.13}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, 2008.

[2] P. Mathai, N. C. Martins, and B. Shapiro, "On the detection of gene network interconnections using directed mutual information," in *Information Theory and Applications Workshop*, 2007, pp. 274–283.

[3] A. Rao, A. O. Hero III, D. J. States, and J. D. Engel, "Using directed information to build biologically relevant influence networks," in *Proc. of Computational Systems Bioinformatics*, 2007, pp. 145–156.

[4] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[5] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag, "Organization, development and function of complex brain networks," *Trends in Cognitive Sciences*, vol. 8, no. 9, pp. 418–425, 2004.

[6] J. W. Duncan and H. S. Steven, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[7] R. Albert, A. L. Barabasi, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A*, vol. 272, no. 1/2, pp. 173–187, 1999.

[8] M. E. J. Newman, "The structure and function of complex networks," *Structure*, vol. 45, no. 2, pp. 167–256, 2003.

[9] Y. Halchenko, S. Hanson, and B. Pearlmutter, "Multimodal integration: fmri, mri, eeg, meg," *Advanced Image Processing in Magnetic Resonance Imaging*, pp. 223–265, 2005.

[10] M. A. Koch, D. G. Norris, and M. Hund-Georgiadis, "An investigation of functional and anatomical connectivity using magnetic resonance imaging," *NeuroImage*, vol. 16, no. 1, pp. 241–250, 2002.

[11] K. J. Friston, "Functional and effective connectivity in neuroimaging: a synthesis," *Human Brain Mapping*, vol. 2, no. 1-2, pp. 56–78, 1994.

[12] K. Friston, "Functional and effective connectivity: A review," *Brain Connectivity*, vol. 1, no. 1, pp. 13–36, 2011.

[13] K. Friston, L. Harrison, and W. Penny, "Dynamic causal modelling," *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[14] 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.

[15] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," *Progress in Neurobiology*, vol. 77, pp. 1–37, 2005.

[16] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological Cybernetics*, vol. 85, pp. 145–157, 2001.

[17] F. Lopes da Silva, J. P. Pijn, and P. Boeijinga, "Interdependence of EEG signals: Linear vs. nonlinear associations and the significance of time delays and phase shifts," *Brain Topography*, vol. 2, pp. 9–18, 1989.

[18] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, pp. 461–464, 2000.

[19] Y. Saito and H. Harashima, *Recent Advances in EEG and EMG Data Processing*, N. Yamaguchi and K. Fujisawa, Eds. Elsevier, Amsterdam, 1981.

[20] J. Massey, "Causality, feedback and directed information," in *Proc. of Intl. Symp. on ISITA*, 1990, pp. 27–30.

[21] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[22] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLoS Computational Biology*, vol. 2, no. 10, p. 144, 2006.

[23] H. Hinrichs, T. Noesselt, and H. J. Heinze, "Directed information flow model free measure to analyze causal interactions in event related EEG-MEG-experiments," *Human brain mapping*, vol. 29, no. 2, pp. 193–206, 2008.

[24] S. Sabesan, L. B. Good, K. S. Tsakalis, A. Spanias, D. M. Treiman, and L. D. Iasemidis, "Information flow and application to epileptogenic focus localization from intracranial eeg," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 17, no. 3, pp. 244–253, 2009.

[25] M. Al-khassaweneh and S. Aviyente, "The Relationship Between Two Directed Information Measures," *IEEE Signal Processing Letters*, vol. 15, pp. 801–804, 2008.

[26] H. Marko, "The bidirectional communication theory–a generalization of information theory," *IEEE Transactions on Communications [legacy, pre-1988]*, vol. 21, no. 12, pp. 1345–1351, 1973.

[27] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 4–21, 2003.

[28] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *Information Theory, IEEE Transactions on*, vol. 55, no. 1, pp. 323–349, 2009.

[29] P. O. Amblard and O. J. Michel, "On directed information theory and granger causality graphs," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 7–16, 2011.

[30] P. O. Amblard and O. J. J. Michel, "Measuring information flow in networks of s-tochastic processes," *Journal of Computational Neuroscience*, pp. 1–10, 2010.

[31] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, pp. 1–28, 2010.

[32] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, p. 824, 2002.

[33] O. Sporns, C. Honey, and R. Kotter, "Identification and classification of hubs in brain networks," *PLoS One*, vol. 2, no. 10, p. 1049, 2007.

[34] D. Meunier, R. Lambiotte, A. Fornito, K. D. Ersche, and E. T. Bullmore, "Hierarchical modularity in human brain functional networks," *Frontiers in neuroinformatics*, vol. 3, no. 37, pp. 1–12, 2009.

[35] D. A. Fair, A. L. Cohen, J. D. Power, N. U. Dosenbach, J. A. Church, F. M. Miezin, B. L. Schlaggar, and S. E. Petersen, "Functional brain networks develop from a "local to distributed" organization," *PLoS computational biology*, vol. 5, no. 5, pp. 1–14, 2009.

[36] L. Ferrarini, I. M. Veer, E. Baerends, M. J. van Tol, R. J. Renken, N. J. A. van der Wee, D. Veltman *et al.*, "Hierarchical functional modularity in the resting-state human brain," *Human brain mapping*, vol. 30, no. 7, pp. 2220–2231, 2009.

[37] D. Meunier, S. Achard, A. Morcom, and E. Bullmore, "Age-related changes in modular organization of human brain functional networks," *Neuroimage*, vol. 44, no. 3, pp. 715–723, 2009.

[38] M. Chavez, M. Valencia, V. Navarro, V. Latora, and J. Martinerie, "Functional modularity of background activities in normal and epileptic brain networks," *Physical review letters*, vol. 104, no. 11, p. 118701, 2010.

[39] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[40] T. M. Cover, J. A. Thomas, and J. Wiley, *Elements of information theory.* Wiley Online Library, 1991, vol. 1.

[41] M. Studeny and J. Vejnarova, "The multiinformation function as a tool for measuring stochastic dependence," *Learning in graphical models*, pp. 261–300, 1998.

[42] Y. Liu, S. Aviyente, and M. Al-khassaweneh, “A high dimensional directed information estimation using data-dependent partitioning,” in *Proc. of IEEE Workshop on SSP*, 2009, pp. 606–609.

[43] C. W. J. Granger, “Testing for causality:: A personal viewpoint,” *Journal of Economic Dynamics and Control*, vol. 2, pp. 329–352, 1980.

[44] L. A. Baccala and K. Sameshima, “Partial directed coherence: a new concept in neural structure determination,” *Biological Cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.

[45] J. Geweke, “Measurement of linear dependence and feedback between multiple time series,” *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.

[46] J. S. Moser, H. S. Schroder, C. Heeter, T. P. Moran, and Y.-H. Lee, “Mind your errors: Evidence for a neural mechanism linking growth mindset to adaptive post-error adjustments,” *Psychological Science*, vol. 22, no. 12, pp. 1484–1489, 2011.

[47] J. S. Moser, T. Moran, and A. Jendrusina, “Parsing relationships between dimensions of anxiety and action monitoring brain potentials in female undergraduates,” *Psychophysiology*, vol. 49, no. 1, pp. 3–10, 2012.

[48] B. A. Eriksen and C. W. Eriksen, “Effects of noise letters upon the identification of a target letter in a nonsearch task,” *Perception and psychophysics*, vol. 16, no. 1, pp. 143–149, 1974.

[49] D. M. Olvet and G. Hajcak, “The stability of error-related brain activity with increasing trials,” *Psychophysiology*, vol. 46, no. 5, pp. 957–961, 2009.

[50] G. Gratton, M. G. H. Coles, and E. Donchin, “A new method for off-line removal of ocular artifact,” *Electroencephalography and clinical Neurophysiology*, vol. 55, no. 4, pp. 468–484, 1983.

[51] 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.

[52] W. Mader, D. Feess, D. Saur, R. Lange, V. Glauchec, C. Weillerc, J. Timmer, and B. Schelter, “Investigating multivariate systems using directed partial correlation,” *International Journal of Bioelectromagnetism*, vol. 12, no. 1, pp. 21–25, 2010.

[53] B. Schelter, M. Winterhalder, M. Eichler, M. Peifer, B. Hellwig, B. Guschlbauer, C. Lücking, R. Dahlhaus, and J. Timmer, “Testing for directed influences among neural signals using partial directed coherence,” *Journal of neuroscience methods*, vol. 152, no. 1-2, pp. 210–219, 2006.

[54] D. Marinazzo, W. Liao, H. Chen, and S. Stramaglia, “Nonlinear connectivity by granger causality,” *Neuroimage*, vol. 58, no. 2, pp. 330–338, 2010.

[55] L. Leistritz, T. Weiss, J. Ionov, K. J. Bär, W. H. R. Miltner, and H. Witte, “Connectivity analysis of somatosensory evoked potentials to noxious intracutaneous stimuli in patients with major depression,” *Methods of Information in Medicine*, vol. 49, no. 5, pp. 484–91, 2010.

[56] L. Faes and G. Nollo, “Extended causal modeling to assess partial directed coherence in multiple time series with significant instantaneous interactions,” *Biological cybernetics*, vol. 103, no. 5, pp. 387–400, 2010.

[57] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, “Transfer entropya model-free measure of effective connectivity for the neurosciences,” *Journal of computational neuroscience*, vol. 30, no. 1, pp. 45–67, 2011.

[58] M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser, “Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks,” *Progress in Biophysics and Molecular Biology*, vol. 105, no. 1-2, pp. 80–97, 2010.

[59] Y. Liu and S. Aviyente, “Information theoretic approach to quantify causal neural interactions from EEG,” in *Signals, Systems and Computers (ASILOMAR), Conference Record of the Forty Fourth Asilomar Conference on.* IEEE, 2010, pp. 1380–1384.

[60] W. Zhao, E. Serpedin, and E. R. Dougherty, “Inferring connectivity of genetic regulatory networks using information-theoretic criteria,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 2, pp. 262–274, 2008.

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

[62] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering, *Numerical recipes: The art of scientific computing, third edition.* Cambridge University Press, 2007.

[63] E. G. Miller, “A new class of entropy estimators for multi-dimensional densities,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2003, pp. 297–300.

[64] J. Lizier, “The local information dynamics of distributed computation in complex systems,” Ph.D. dissertation, University of Sydney, 2010.

[65] B. Schelter, M. Winterhalder, and J. Timmer, *Handbook of time series analysis: recent theoretical developments and applications.* Vch Verlagsgesellschaft Mbh, 2006.

[66] L. Cao, “Practical method for determining the minimum embedding dimension of a scalar time series,” *Physica D: Nonlinear Phenomena*, vol. 110, no. 1, pp. 43–50, 1997.

[67] G. Pipa and S. Grün, “Non-parametric significance estimation of joint-spike events by shuffling and resampling,” *Neurocomputing*, vol. 52, pp. 31–37, 2003.

[68] A. K. Seth, “A MATLAB toolbox for granger causal connectivity analysis,” *Journal of neuroscience methods*, vol. 186, no. 2, pp. 262–273, 2010.

[69] G. Nolte, A. Ziehe, V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K. Müller, “Robustly estimating the flow direction of information in complex physical systems,” *Physical Review Letters*, vol. 100, no. 23, p. 234101, 2008.

[70] M. Breakspear, “Nonlinear phase desynchronization in human electroencephalographic data,” *Human Brain Mapping*, vol. 15, no. 3, pp. 175–198, 2002.

[71] W. Michiels and H. Nijmeijer, “Synchronization of delay-coupled nonlinear oscillators: An approach based on the stability analysis of synchronized equilibria,” *Chaos*, vol. 19, no. 3, p. 033110, 2009.

[72] J. C. Butcher and J. Wiley, *Numerical methods for ordinary differential equations.* Wiley Online Library, 2003, vol. 2.

[73] J. Kayser and C. E. Tenke, “Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks,” *Clinical Neurophysiology*, vol. 117, no. 2, pp. 348–368, 2006.

[74] S. Aviyente, E. Bernat, W. Evans, and S. Sponheim, “A phase synchrony measure for quantifying dynamic functional integration in the brain,” *Human brain mapping*, vol. 32, no. 1, pp. 80–93, 2011.

[75] J. F. Cavanagh, M. X. Cohen, and J. J. B. Allen, “Prelude to and resolution of an error: Eeg phase synchrony reveals cognitive control dynamics during action monitoring,” *The Journal of Neuroscience*, vol. 29, no. 1, pp. 98–105, 2009.

[76] C. R. Genovese, N. A. Lazar, and T. Nichols, “Thresholding of statistical maps in functional neuroimaging using the false discovery rate,” *Neuroimage*, vol. 15, no. 4, pp. 870–878, 2002.

[77] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization.* Wiley-Interscience, 1992.

[78] L. Kozachenko and N. Leonenko, “On statistical estimation of entropy of random vector,” *Problems of Information Transmission*, vol. 23, no. 2, pp. 95–101, 1987.

[79] L. Zhao, H. Permuter, Y. H. Kim, and T. Weissman, “Universal estimation of directed information,” in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on.* IEEE, 2010, pp. 1433–1437.

[80] L. Faes, G. Nollo, and A. Porta, “Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique,” *Physical Review E*, vol. 83, no. 5, p. 051112, 2011.

[81] V. A. Smith, E. D. Jarvis, and A. J. Hartemink, “Evaluating functional network inference using simulations of complex biological systems,” *Bioinformatics*, vol. 18, no. Suppl 1, p. S216, 2002.

[82] W. J. Melssen and W. J. M. Epping, "Detection and estimation of neural connectivity based on crosscorrelation analysis," *Biological cybernetics*, vol. 57, no. 6, pp. 403–414, 1987.

[83] C. Andrew and G. Pfurtscheller, "Event-related coherence as a tool for studying dynamic interaction of brain regions," *Electroencephalography and clinical Neurophysiology*, vol. 98, no. 2, pp. 144–148, 1996.

[84] O. David, S. J. Kiebel, L. M. Harrison, J. Mattout, J. M. Kilner, and K. J. Friston, "Dynamic causal modeling of evoked responses in EEG and MEG," *NeuroImage*, vol. 30, no. 4, pp. 1255–1272, 2006.

[85] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, J. Timmer, and H. Witte, "Detection of directed information flow in biosignals," *Biomedizinische Technik*, vol. 51, no. 5/6, pp. 281–287, 2006.

[86] K. P. Murphy, "Dynamic bayesian networks: representation, inference and learning," Ph.D. dissertation, Citeseer, 2002.

[87] V. A. Smith, J. Yu, T. V. Smulders, A. J. Hartemink, and E. D. Jarvis, "Computational inference of neural information flow networks," *PLoS Comput Biol*, vol. 2, no. 11, p. e161, 2006.

[88] D. M. Chickering, "Learning Bayesian networks is NP-complete," *Learning from data: Artificial intelligence and statistics v*, vol. 112, pp. 121–130, 1996.

[89] N. Friedman, I. Nachman, and D. Peer, "Learning Bayesian network structure from massive datasets: The "sparse candidate" algorithm," in *Proc. UAI*, 1999.

[90] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of computational biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[91] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, no. 1, 2007.

[92] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC bioinformatics*, vol. 7, 2006.

[93] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing*, vol. 5, 2000, pp. 418–429.

[94] C. Quinn, T. Coleman, and N. Kiyavash, "Causal dependence tree approximations of joint distributions for multiple random processes," *Arxiv preprint arXiv:1101.5108*, 2011.

[95] Y. Liu and S. Aviyente, "Directed network inference using a measure of directed information," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 513–516.

[96] L. Harrison, W. D. Penny, and K. Friston, "Multivariate autoregressive modeling of fMRI time series," *NeuroImage*, vol. 19, no. 4, pp. 1477–1491, 2003.

[97] C. J. Stam, "Functional connectivity patterns of human magnetoencephalographic recordings: a 'small-world' network?" *Neuroscience letters*, vol. 355, no. 1-2, pp. 25–28, 2004.

[98] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, p. 7821, 2002.

[99] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[100] Y. Kim, S. Son, and H. Jeong, "Finding communities in directed networks," *Physical Review E*, vol. 81, no. 1, p. 016103, 2010.

[101] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Physical Review Letters*, vol. 100, no. 11, p. 118703, 2008.

[102] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[103] B. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell System Technical Journal*, vol. 49, no. 2, pp. 291–307, 1970.

[104] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 281-297. California, USA, 1967, p. 14.

[105] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888–905, 2000.

[106] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14: Proceeding of the 2001 Conference*, 2001, pp. 849–856.

[107] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, p. 8577, 2006.

[108] A. Arenas, J. Duch, A. Fernández, and S. Gómez, "Size reduction of complex networks preserving modularity," *New Journal of Physics*, vol. 9, p. 176, 2007.

[109] M. E. J. Newman, “Detecting community structure in networks,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[110] R. Guimera and L. A. N. Amaral, “Functional cartography of complex metabolic networks,” *Nature*, vol. 433, no. 7028, pp. 895–900, 2005.

[111] J. Duch and A. Arenas, “Community detection in complex networks using extremal optimization,” *Physical Review E*, vol. 72, no. 2, p. 027104, 2005.

[112] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Physical Review E*, vol. 69, no. 6, p. 066133, 2004.

[113] K. Wakita and T. Tsurumi, “Finding community structure in mega-scale social networks:[extended abstract],” in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 1275–1276.

[114] J. M. Pujol, J. Béjar, and J. Delgado, “Clustering algorithm for determining community structure in large networks,” *Physical Review E*, vol. 74, no. 1, p. 016107, 2006.

[115] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, “Module identification in bipartite and directed networks,” *Physical Review E*, vol. 76, no. 3, p. 036102, 2007.

[116] J. Li, Z. J. Wang, S. J. Palmer, and M. J. McKeown, “Dynamic Bayesian network modeling of fMRI: A comparison of group-analysis methods,” *NeuroImage*, vol. 41, no. 2, pp. 398–407, 2008.

[117] A. Mechelli, W. D. Penny, C. J. Price, D. R. Gitelman, and K. J. Friston, “Effective connectivity and intersubject variability: using a multisubject network to test differences and commonalities,” *Neuroimage*, vol. 17, no. 3, pp. 1459–1469, 2002.

[118] M. S. Goncalves, D. A. Hall, I. S. Johnsrude, and M. P. Haggard, “Can meaningful effective connectivities be obtained between auditory cortical regions?” *NeuroImage*, vol. 14, no. 6, pp. 1353–1360, 2001.

[119] A. Strehl and J. Ghosh, “Cluster ensembles—a knowledge reuse framework for combining multiple partitions,” *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[120] J. Neumann, G. Lohmann, J. Derrfuss, and D. Y. Von Cramon, “Meta-analysis of functional imaging data using replicator dynamics,” *Human brain mapping*, vol. 25, no. 1, pp. 165–173, 2005.

[121] N. M. Correa, T. Adali, Y. O. Li, and V. D. Calhoun, “Canonical correlation analysis for data fusion and group inferences,” *Signal Processing Magazine, IEEE*, vol. 27, no. 4, pp. 39–50, 2010.

[122] S. Aviyente, E. M. Bernat, W. S. Evans, C. J. Patrick, and S. R. Sponheim, “A phase synchrony measure for quantifying dynamic functional integration in the brain,” *Human Brain Mapping*, vol. 32, no. 1, pp. 80–93, 2010.

[123] J. Berry, B. Hendrickson, R. LaViolette, and C. Phillips, "Tolerating the community detection resolution limit with edge weighting," *Physical Review E*, vol. 83, no. 5, p. 056119, 2011.

[124] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, "Detecting complex network modularity by dynamical clustering," *Physical Review E*, vol. 75, no. 4, p. 045102, 2007.

[125] P. Mucha, T. Richardson, K. Macon, M. Porter, and J. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, no. 5980, pp. 876–878, 2010.

[126] E. Ordentlich, "Maximizing the entropy of a sum of independent bounded random variables," *Information Theory, IEEE Transactions on*, vol. 52, no. 5, pp. 2176–2181, 2006.