BRAIN CONNECTIVITY ANALYSIS USING INFORMATION THEORY AND STATISTICAL SIGNAL PROCESSING

By

Zhe Wang

A DISSERTATION

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

Electrical Engineering — Doctor of Philosophy

2017

ABSTRACT

BRAIN CONNECTIVITY ANALYSIS USING INFORMATION THEORY AND STATISTICAL SIGNAL PROCESSING

By

Zhe Wang

Connectivity between different brain regions generates our minds. Existing work on brain network analysis has mainly been focused on the characterization of connections between the regions in terms of connectivity and causality. Connectivity measures the dependence between regional brain activities, and causality analysis aims to determine the directionality of information flow among the functionally connected brain regions, and find the relationship between causes and effects.

Traditionally, the study on connectivity and causality has largely been limited to linear relationships. In this dissertation, as an effort to achieve more accurate characterization of connections between brain regions, we aim to go beyond the linear model, and develop innovative techniques for both non-directional and directional connectivity analysis. Note that due to variability in the brain connectivity of each individual, the connectivity between two brain regions alone may not be sufficient for brain function analysis, in this research, we also conduct network connectivity pattern analysis, so as to reveal more in-depth information.

First, we characterize non-directional connectivity using mutual information (MI). In recent years, MI has gradually appeared as an alternative metric for brain connectivity, since it measures both linear and non-linear dependence between two brain regions, while the traditional Pearson correlation only measures the linear dependence. We develop an innovative approach to estimate the MI between two functionally connected brain regions and apply it to brain functional magnetic resonance imaging (fMRI) data. It is shown that:

on average, cognitively normal subjects show larger mutual information between critical regions than Alzheimer's disease (AD) patients.

Second, we develop new methodologies for brain causality analysis based on directed information (DI). Traditionally, brain causality is based on the well-known Granger Causality (GC) analysis. The validity of GC has been widely recognized. However, it has also been noticed that GC relies heavily on the linear prediction method. When there exists strong nonlinear interactions between two regions, GC analysis may lead to invalid results. In this research, (i) we develop an innovative framework for causality analysis based on directed information (DI), which reflects the information flow from one region to another, and has no modeling constraints on the data. It is shown that DI based causality analysis is effective in capturing both linear and non-linear causal relationships. (ii) We show the conditional equivalence between the DI Framework and Friston's dynamic causal modeling (DCM), and reveal the relationship between directional information transfer and cognitive state change within the brain.

Finally, based on brain network connectivity pattern analysis, we develop a robust method for the AD, mild cognitive impairment (MCI) and normal control (NC) subject classification under size limited fMRI data samples. First, we calculate the Pearson correlation coefficients between all possible ROI pairs in the selected sub-network and use them to form a feature vector for each subject. Second, we develop a regularized linear discriminant analysis (LDA) approach to reduce the noise effect. The feature vectors are then projected onto a subspace using the proposed regularized LDA, where the differences between AD, MCI and NC subjects are maximized. Finally, a multi-class AdaBoost Classifier is applied to carry out the classification task. Numerical analysis demonstrates that the combination of regularized LDA and the AdaBoost classifier can increase the classification accuracy significantly.

Dedicated to my family and friends.

ACKNOWLEDGMENTS

I would like to take this opportunity to express my sincere appreciation for all the support and encouragement that have led to the completion of this dissertation. I am greatly indebted to my advisor, Prof. Tongtong Li, for her continuous support, instruction, and encouragement throughout my PhD studies at Michigan State University. It is a great honor to have worked with Dr.Li. This work would not have been possible without her guidance.

I would also like to thank Prof. Hassan Khalil, Prof. David C. Zhu, Prof. Taosheng Liu, and Prof. Dean Aslam for serving on my committee. I am deeply grateful to them for their valuable experience, experimental support and insightful discussions throughout my PhD program.

Special thanks go to my colleagues in the Broadband Access and Wireless Communication (BAWC) lab for their insightful academic inputs: Mai Abdelhakim, Tianlong Song, Ahmed Alahmadi, Zhaoxi Fang, Yuan Liang, Yu Zheng, and Run Tian. Their generous help made my study more enjoyable.

I am greatly grateful to my family and friends for their tremendous support and encouragement. This dissertation would not have been accomplished without their love, care, patience, and support.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

In this chapter, first, a brief overview of fMRI based functional brain connectivity analysis is provided, illustrating the backgrounds of connectivity analysis. Second, different topics of functional connectivity analysis, including non-directional connectivity analysis, directional connectivity (or causality) analysis, and network connectivity pattern analysis are presented. Finally, the major contributions of this dissertation are highlighted.

1.1 Overview of Brain Connectivity Analysis

The brain is a communication network. At the neuron level, information exchanges are achieved through communications between synapses. At the system level, different brain regions formulate a dynamic communication network, and connectivity between the brain regions generates our minds.

In neuroscience, connectivity analysis plays a critical role since it can provide insightful information in understanding brain functions and dysfunctions. Brain researchers are increasingly looking for advanced computational analysis tools to assist them in understanding the functions and dysfunctions of specific brain connectivities. At the same time, driven by the revolution in information theory, the communications area has accumulated rich methodologies for system modeling, design, signal processing and extraction, and network characterization and evaluation. To this end: can we develop innovative computational

analysis tools for brain connectivity analysis by exploiting advanced methodologies in communications? How can these tools help us understand patterns of functional connectivities?

As an effort to address these problems, this research aims to take the advantages of functional magnetic resonance imaging (fMRI) to develop innovative modeling and analysis methodologies for brain connectivity analysis by exploiting advanced techniques in communications, especially tools in information theory and network characterization; apply these methodologies to brain network analysis, and explore how the proposed techniques can help us identify connectivity problems and understand why the brain fulfills or fails a cognitive task.

1.2 Revisit of the Existing Work

In brain connectivity analysis, there are three closely related components: non-directional connectivity analysis, directional connectivity (or causality) analysis, and network connectivity pattern analysis. In this section, we revisit the existing work on these three topics.

1.2.1 Non-directional Functional Connectivity Analysis of Brain Regions

Connectivity among the brain regions during a cognitive activity helps us understand brain functions and perform disease analysis. For decades, Pearson correlation [1] has widely been used as a quantitative metric to characterize the functional connectivity between two different brain regions. In recent years, mutual information (MI) has gradually appeared as an alternative metric for brain connectivity. The underlying argument is that: the Pearson correlation coefficient only measures the linear dependence, while MI measures both linear

and non-linear dependence between two brain regions. Moreover, MI has a clear physical meaning: it represents the information successfully transmitted over the two brain regions under consideration.

In [1], Tsai et al. used MI to build the brain activation map. They showed that MI was robust in quantifying the relationship between any two fMRI time series. An outstanding merit of the MI approach was that, it does not depend on the a priori assumptions about the relationship between the protocol time line and the fMRI voxel temporal response, and yet could be as effective as Pearson correlation for calculating activation maps. In [2], Michiel et al. applied MI to the decoding algorithm in feature selection from high dimensional data. Their results showed that, comparing with analysis of variance (ANOVA) based method, MI is efficient in selecting very few but strongly informative voxels, and meanwhile can achieve the same or even better generalization or overall performance. In [3], Chai et al. used multivariate mutual information to select voxels in decoding natural scene categories from the human brain. Their experiments showed that, comparing with the classical variance-based "most active selection" method [4], MI based voxel selection could improve the decoding accuracy significantly. In [5], Gomez-Verdejo et al. used MI to identify regionally specific effects produced by a particular cognitive task, and showed that MI could confirm known functional connections identified by Pearson correlation, and can also discover new connections.

1.2.2 Directional Functional Connectivity Analysis of Brain Regions

Causality analysis aims to find the relationship between causes and effects. It provides insightful information on how brain regions interact with each other during a cognitive task [6]. In general, causality analysis tries to determine whether the values of one time series are useful in predicting the future values of another time series. Since 1990s, a number of frameworks have been applied to fMRI based causality analysis, including *Granger Causality* (GC), *Bayesian Network*, *Dynamic Causal Modeling* (DCM), *Transfer Entropy* (TE) and *Directed Information* (DI).

Granger Causality The first practical causality analysis framework was proposed by Granger in 1969 [7]. The main idea is, if two signals X_1 and X_2 form a causal relationship, then, instead of using the past values of X_2 alone, the information contained in the past values of X_1 will help to predict X_2. More specifically, the calculation of Granger Causality is based on the linear prediction models. Suppose $\mathbf{X}_1^n = [X_1(1), X_1(2), ..., X_1(n)]$ and $\mathbf{X}_2^n = [X_2(1), X_2(2), ..., X_2(n)]$ are two time series observed from two brain regions, respectively. Granger Causality compares the prediction errors e_r and $\tilde{e}_r$ in the following equations:

$$X_2(k+1) = \sum_{l=0}^{L-1} a_l X_2(k-l) + e_r, \tag{1.1}$$

$$X_2(k+1) = \sum_{l=0}^{L-1} [b_l X_1(k-l) + c_l X_2(k-l)] + \tilde{e}_r, \tag{1.2}$$

for $k = 1, 2, ..., n$. Here, e_r is the error of predicting X_2 based only on the previous values

of X_2, and $\tilde{e}_r$ is the error of predicting X_2 based on both the previous values of X_2 and the previous values of X_1. If $\tilde{e}_r$ is much smaller than e_r, that is, the introduction of the previous values of X_1 can improve the prediction accuracy, then we say there is a Granger causal relationship between X_1 and X_2.

In literature, there have been growing interests in the use of Granger Causality analysis to identify causal interactions in neuroscience [8]. GC has been successfully applied to fMRI data, EEG measurements, as well as neural level signals [9–11]. In these pioneering work, the validity and computational simplicity of Granger Causality have been widely recognized. At the same time, it has also been noticed that GC relies on the linear prediction method. When there exist instantaneous and/or strong nonlinear interactions between two regions, GC analysis may lead to invalid results [11].

Bayesian Network In [12], J. Pearl summarized the framework of Bayesian Network for causal inference. The argument behind it is that: if a causal relationship exists between two factors X and Y, the introduction of factor X may change the distribution of another factor Y. That is, $P(Y|X) \neq P(Y)$.

Since 2000, the analyses based on Bayesian Networks have demonstrated successful applications [13, 14]. Modified Bayesian Network has been applied to fMRI data by incorporating the vector autoregressive model used in GC [14]. From a general perspective, the vector autoregressive model based Bayesian Network framework can be regarded as a variation of the Granger Causality analysis.

Dynamic Causal Modeling In 2003, Friston proposed the framework of *Dynamic Causal Modeling* to describe the general interactions among a group of brain regions [15]. DCM assumes that the invisible neurostate X, the (external) input U, the observed BOLD signal Y, the parameter θ that characterizes the connectivities between two brain regions, and

the independent noise Ω form a dynamic system that could be described by the following equations:

$$\dot{X} = f(X, U, \theta) \text{ and } Y = \Lambda(X) + \Omega, \tag{1.3}$$

where Λ represents a cascade of differential equations which map the neurostate X to the observed BOLD signal Y. Relying on the EM algorithm, DCM has been implemented on both fMRI and EEG data [16]. In practical applications, due to the computational complexity, DCM is usually used as a confirmatory approach. That is, the users need to put forward different connectivity models and then compare them based on their likelihood evaluated under DCM [17].

Transfer Entropy Another widely applied causal measurement in neuroscience is *Transfer Entropy* (TE). TE was introduced in 2000 by Schreiber [18]. It measures the decrease of entropy in one signal Y after another signal X has been observed:

$$T_{X \to Y} \stackrel{\triangle}{=} H\left(Y_t \mid Y_{t-1:t-L}\right) - H\left(Y_t \mid Y_{t-1:t-L}, X_{t-1:t-L}\right) \tag{1.4}$$

in which H denotes the entropy operator, $Y_{t-1:t-L} = [Y_{t-L}, ..., Y_{t-1}]$, $X_{t-1:t-L} = [X_{t-L}, ..., X_{t-1}]$.

The first exploration of applying transfer entropy in causality description was conducted by Sporns et al. on the sensorimotor network in 2006 [19]. TE has also been applied in MEG data to evaluate non-linear connectivity [20], and used for fMRI data [21] to detect the directed flow of information between brain regions.

As an information theoretic framework, transfer entropy does not rely on any model

assumptions of the signals. However, current algorithms on TE estimation have not been proved to be convergent [22]. Also, in [18, 23], it was shown that the amplitude of transfer entropy could not accurately quantify the strength of influence between brain regions.

Directed Information Directed Information is an information theoretic metric, which was first introduced by Massey when studying communication channels with feedback [24]. It measures the directed information flow from one time series X to another time series Y, denoted as $I(X \to Y)$. If $I(X \to Y) > I(Y \to X)$, then we say X has more influence on Y, or X is the causal side in the connectivity.

DI is a universal method. Unlike GC, which mainly relies on the linear prediction theory, or linear modeling for the involved parameters, the DI-based causality analysis does not have any modeling constraint on the sequences to be evaluated, hence, can be used to characterize more general relationships. This advantage of DI has been reported in recent advances in causality analysis [22, 25, 26]. In [27], it was pointed out that GC analysis is effective in detecting linear or nearly linear causal relationship, but may have difficulty in capturing nonlinear causal relationships. On the other hand, DI-based causality analysis is more effective in capturing both linear and nonlinear causal relationships. In [28], Liu et al. applied DI to the EEG data and compared the result with that of GC. Their conclusion was that DI based approach could be superior to GC in capturing the instantaneous and nonlinear causal relationship in EEG data. Moreover, in [29], it was shown that the Granger Causality graphs of stochastic processes can be generated from the DI framework, and the authors indicated that the DI theory provides an adequate framework for the connectivity inference problems in neuroscience applications. A comprehensive investigation of DI can be found in [30].

1.2.3 Network Connectivity Pattern Analysis of Brain Regions

Both directional connectivity and non-directional connectivity analysis study the connection between two brain regions. As will be shown in *Chapter 4*, although analysis based on pair-wise connection could provide insights into the brain network, the information from merely two regions are far from enough to reveal the general functional connectivity patterns. As a result, both of the two aforementioned analyses are inadequate in solving problems that involve multiple brain regions, such as the classification of Alzheimer's Disease (AD) patients, Mild Cognitive Impairment (MCI) patients and normal control (NC) subjects based on fMRI data. The underlying argument is that: due to variability in the brain connectivity of each individual, the connectivity between two brain regions alone may not provide comprehensive information for brain analysis; network connectivity pattern analysis, which looks for subtle changes in the pattern of connectivity among multiple or all regions in the sub-network, may reveal more in-depth information.

1.2.4 Motivations and Problem Identification

After revisiting the existing work, we identify the major problems in today's brain connectivity analysis as follows.

- **Connections between brain regions: going beyond the linear model**. Existing work on brain connectivity analysis has mainly been focused on the characterization of connections between the regions in terms of connectivity and causality [31]. Connectivity measures the dependence between regional brain activities, and tells us which brain regions are functionally connected during a cognitive task [32]. Moving one step further, causality analysis aims to determine the directionality of information flow among

the functionally connected brain regions [33, 34].

Traditional connectivity analysis mainly relies on the Pearson correlation, which only measures the linear dependence between the regions [35–37]. Traditional causality analysis largely relies on the Granger Causality (GC) approach [38, 39]. Again, GC is based on the linear prediction technique [40]. It is effective in detecting linear or nearly linear causal relationship, but may have difficulty in capturing nonlinear causal relationships [40]. As can be seen, the study on connectivity and causality has largely been limited to linear relationships. For more accurate characterization of connections between brain regions, we need to go beyond the linear model.

- **Relationships between representative causality analysis frameworks**. GC, DI and DCM are three representative causality analysis frameworks in brain connectivity analysis. In literature, the relationships between GC and DCM, and between GC and DI have been investigated. A missing link here is: what is the relationship between DCM and DI? To fill the missing link, and reveal the connection between DCM and DI, we need both theoretical and numerical analyses of the equivalence between DCM and DI in characterizing the causal relationship between two brain regions.

- **Reliable classifications of brain diseases based on size limited fMRI data**. Accurate distinction of AD and MCI patients from normal subjects is critical for early diagnosis and treatment of brain disorders. However, the size of fMRI data samples is generally quite limited, which has become a major bottleneck in fMRI based AD, MCI and NC classification. The underlying reason is that, when the sample size is small, most existing classifiers suffer from severe noise effects, due to both biological variability and measurement noise. New methodologies have to be developed for robust

AD, MCI and NC classification.

1.3 Summary of Dissertation Contributions

The main contributions of this dissertation are summarized in the following.

In Chapter 2, we explore the estimation of mutual information to measure the non-directional connectivities between brain regions. Traditional non-directional connectivity analysis mainly relies on the Pearson correlation, which only measures the linear dependence between the regions. Motivated by this observation, in this chapter, we propose to measure the non-directional connectivities with mutual information, which measures both linear and nonlinear relationships between brain regions. An innovative MI estimator is developed. The major steps include: de-correlation of data segments, kernel-based estimation of the probability density function, and Monte Carlo Integration. The proposed MI estimator is applied to experimental fMRI data obtained from Alzheimer disease patients and normal subjects. The numerical result is consistent with clinical observations.

In Chapter 3, we conduct fMRI based causality analysis in brain connectivity by exploiting the directed information based framework. First, we introduce the core concepts in the directed information framework. Second, we present how to conduct causality analysis using directed information measures between two time series. We provide the detailed procedure on how to calculate the DI for two finite time series. The two major steps involved here are optimal bin size selection for data digitization, and probability estimation. Finally, we demonstrate the applicability of DI based causality analysis using both the simulated data and experimental fMRI data, and compare the results with that of the Granger Causality analysis. Our analysis indicates that GC analysis is effective in detecting linear or nearly

linear causal relationship, but may have difficulty in capturing nonlinear causal relationships. On the other hand, DI based causality analysis is more effective in capturing both linear and non-linear causal relationships. Moreover, it is observed that brain connectivity among different regions generally involves dynamic two-way information transmissions between them. Our results show that when bidirectional information flow is present, DI is more effective than GC to quantify the overall causal relationship.

In Chapter 4, we explore the discrete Dynamic Causal Modeling (DDCM) and its relationship with DI and GC. First, we revisit DDCM, and demonstrate the relationship between DDCM and the conventional continuous time DCM. Second, we show that under certain conditions, DDCM and DI are equivalent in characterizing the causal relationship between two brain regions. Recall that traditionally, the accuracy of DI estimation is based on the accuracy of probability or statistic estimation, and hence requires the data length be sufficiently long. This equivalence between DDCM and DI, in fact, also provides a simple but effective method for DI estimation under limited data length. Finally, we illustrate the similarities and differences between DDCM and GC.

In Chapter 5, we develop a reliable method for AD, MCI and NC classification that is robust with respect to size limited *fMRI data samples, by exploiting brain network connectivity pattern analysis.* First, we propose a regularized LDA approach, which aims to reduce the noise effect by using two shrinkage methods. The first shrinkage method moves the estimated mean of each class towards the overall mean, and the second one shifts the estimated covariance matrix for each class towards the identity matrix. Second, we investigate the relationship between LDA-based and Maximum Likelihood (ML) based classification or decision making methods. Finally, we conduct the connectivity pattern classification of AD, MCI and NC subjects by applying the regularized LDA and AdaBoost classifier based ap-

proach. Numerical analysis shows that: in comparison with the original LDA approach [41], the regularized LDA can reduce the noise effect and increase the classification accuracy significantly. Our analysis also confirms the previous findings that the hippocampus and the isthmus of the cingulate cortex are closely involved in the development of AD and MCI.

In *Chapter 6*, we summarize the conclusions and present some potential directions for future research.

Chapter 2

Brain Functional Connectivity Analysis Using Mutual Information

This chapter considers measuring brain functional connectivity using mutual information (MI). First, we explain the advantage of MI based analysis over the conventional correlation based analysis. Second, we propose a novel approach for MI estimation by exploiting kernel-based probability density function (pdf) estimation and optimization under the maximum likelihood criteria. Finally, the proposed estimator is applied to true fMRI data obtained from Alzheimer's Disease (AD) patients and normal control (NC) subjects. The numerical analysis demonstrates the effectiveness of the proposed approach and shows that the MI based analysis result is consistent with clinical observations.

2.1 Introduction

Connectivity analysis based on functional Magnetic Resonance Imaging (fMRI) data helps to reveal insights on brain functioning and disease analysis [42]. For decades, an important metric in measuring functional connectivities has been the Pearson correlation coefficient. In recent years, mutual information (MI) has been applied as an alternative metric for the reason that it measures not only linear dependence between two time series but also non-linear relationships, and meanwhile has a clear physical meaning. It represents the

information successfully transmitted on the brain links. It was pointed out in [3, 5] that MI is more informative compared to traditional metrics because it could confirm known functional connections as well as discovering new connections.

Since the last decade, there has been a growing interest in applying MI on fMRI data analysis. In [1], Tsai et al. used MI to build the brain activation map. They showed that MI was robust in quantifying the relationship between any two fMRI temporal response waveforms. An outstanding merit of the MI approach was that, it does not depend on the priori assumptions about the relationship between the protocol time line and the fMRI voxel temporal response, and yet could be as effective as other methods for calculating activation maps. In [2], Michiel et al. applied MI in the decoding algorithm in selecting features from high dimensional data. Their results showed that, compared to analysis of variance (ANOVA) based method, MI was efficient in selecting very few but strongly informative voxels and meanwhile achieved the same or even better generalization performance. In [43], Afshin-Pour et al. applied MI in the activation detection. They carried out experiment in real datasets for group analyses using the general linear model, and showed that MI is a more sensitive metric than the Jaccard overlap metric.

In literature, a dominant approach for MI calculation has been the k nearest neighbor (kNN) estimator. This approach has been discussed throughly in [44]. The main idea is: when two sets of high dimensional data points are independently and identically distributed (iid), the estimation bias in the kNN density estimator will be demolished when calculating the KL distance, and the final estimator will be asymptotically unbiased. There are two limitations with this approach: first, the choice of k is highly empirical. The choice involves tradeoffs between the estimation bias and variance, with smaller k leads to lower bias and higher variance [44]. An empirical choice of k is $\sqrt{n}$. In fact, previous works on MI estimation

did not provide any discussions on how to choose k to ensure accurate estimation. Second, applying kNN estimator on two time series could face the problem of data independence. When data points are not independently distributed, it is not guaranteed the final result would converge to the true value.

In this chapter, we propose to estimate MI using a novel approach to evaluate the MI in the region level to analyze the functional connectivity of Default Mode Network (DMN) in the brain. Note that the data segments from the fMRI data are generally correlated with each other, which introduces skewness in the distribution [44]. In this chapter, first, we apply a linear transformation to the fMRI data such that the covariance matrix of the transformed data is close to the identity matrix. Second, we present an effective approach for MI estimation by exploiting kernel-based probability density function estimation and optimization under the maximum likelihood criteria. Finally, the proposed estimator is applied to true fMRI data obtained from Alzheimer's Disease (AD) patients and normal control (NC) subjects. The numerical analysis demonstrates the effectiveness of the proposed approach and shows that the MI based analysis result is consistent with clinical observations.

The rest of this chapter is organized as follows. In Section II, we outline the existing approaches used in brain connectivity analysis. In Section III, we present the proposed kernel based estimator, and the corresponding algorithms. Numerical results are provided in Section IV, and we conclude in Section V.

Notation: The uppercase letters (X,Y,...) denote random variables, and the lowercase letters (x,y,...) denote the possible values they can acquire. $\overrightarrow{X} = [X_1, X_2, ..., X_d]$ denotes a time series vector, where X_i is the ith sample. For any x, $f_X(x)$ denotes the probability density function (pdf) of X, and $f_{XY}(x, y)$ the joint probability density function for (X, Y). The log function $log(*)$ denotes the base 2 logarithm.

2.2 Existing Approaches

In this section, we revisit some representative approaches in the measurement of functional connectivities, and discuss some of their limitations.

2.2.1 Pearson Correlation versus Mutual Information

In fMRI studies, a widely used statistical metric in brain functional connectivity analysis is the Pearson correlation coefficient, which measures the linear dependence between two time series. It is defined as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{2.1}$$

where $\bar{X}$ and $\bar{Y}$ represent the mean of the two time series $\{X\}$ and $\{Y\}$, respectively. For further statistical hypotheses testing, the Fisher's Z-transformation is generally applied to regulate distribution:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r). \tag{2.2}$$

The introduction of Fisher's transform could also be used to stabilize the variance of data points for regression based or ANOVA techniques, which has been explored in [42].

While the Pearson correlation coefficient is efficient in capturing linear correlations, it is an inaccurate measure when the regional activities in brain show non-linear characteristics. Moreover, as statistical metrics, the correlation coefficient r, as well as its Fisher transform z, do not really reflect the connection strength comprehensively. The mutual information,

on the other hand, denoted the information successfully transmitted through a channel, has a more clear physical meaning in measuring the strength of a connection. This motivates applying MI analysis in measuring the brain functional connectivity.

The mutual information for two random variables X and Y is defined as:

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y) log \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)} dxdy. \tag{2.3}$$

For fMRI data, which is composed of time series, the mutual information is carried out on segments rather than single points of the data. In this chapter, we choose the length of one segment, d, to be the rough duration of the Hemodynamic Response in terms of sampling periods. That is,

$$I(\overrightarrow{X};\overrightarrow{Y}) = \int_{R^d}\int_{R^d} f_{XY}(\overrightarrow{x},\overrightarrow{y}) log \frac{f_{XY}(\overrightarrow{x},\overrightarrow{y})}{f_X(\overrightarrow{x})f_Y(\overrightarrow{y})} d\overrightarrow{x}\, d\overrightarrow{y}. \tag{2.4}$$

2.2.2 Limitations with the Existing Work on MI estimation

Although MI has considerable advantages over the Pearson correlation coefficient, current work on estimating MI has its own limitations.

A widely exploited algorithm to estimate mutual information of high dimensional data is built upon the k nearest neighbor (kNN) estimator. Given n samples with dimension d, this estimator calculates the pdf as:

$$f(\overrightarrow{x_i}) = \frac{1}{2}\frac{\Gamma(d/2+2)}{\pi^{d/2}}\frac{1}{r(\overrightarrow{x_i})^d}, \tag{2.5}$$

in which $r(\overrightarrow{x_i})$ is the Euclidean distance from $\overrightarrow{x_i}$ to its kth nearest neighboring points, and $\Gamma(*)$ the Gamma function [3].

There are several concerns related to the kNN estimator. First, the choice of k is not a well-defined problem and usually solved by heuristic techniques [45]. Second, the underlying assumption of the kNN estimator is that: the data points are identically and independently distributed (i.i.d.). Otherwise, the algorithm can not be guaranteed to converge to the true value. In fMRI studies, however, even under resting state, the time series points can not be simplified as i.i.d processes. Third, the MI calculation in [3] is based on individual single data points. However, information contained in the time changing Hemodynamic waveforms are certainly more informative than that in single data points.

Hence, unlike previous approaches in [3], the proposed estimation of mutual information will be carried out on segments rather than single points (see Section III). The segment length is roughly the same as that of one Hemodynamic Response. Moreover, we resort to the non-parameterized kernel method to estimate the probability density function, and try to reduce inter-dependence of data points by data preprocessing.

2.3 The proposed approach for MI estimation

In this section, we present the proposed MI estimator. The major steps in the estimator include: de-correlation of data segments, kernel-based estimation of the probability density function, and Monte Carlo Integration for MI estimation.

2.3.1 De-correlation of Data Segments

The data segments from the fMRI data are generally correlated with each other, which introduces skewness in the distribution [44]. To fix this problem, we propose to adopt the whitening transform on the data segments before estimating the probability density function such that the covariance matrix of the transformed data is close to the identity matrix:

Given a set of d-dimensional data $\overrightarrow{x}$, with mean $\overrightarrow{\mu}$ and covariance matrix $\Sigma = E(\overrightarrow{x} - \overrightarrow{\mu})(\overrightarrow{x} - \overrightarrow{\mu})^T$. Its eigenvalue decomposition is $\Sigma = \phi\Lambda\phi^{-1}$, in which Λ is a diagonal matrix, with the eigenvalues as its diagonal elements, and ϕ the eigenvectors of the covariance matrix. Define the whitening transform as:

$$\overrightarrow{x}' = A^T\overrightarrow{x}, \tag{2.6}$$

where $A = \phi\Lambda^{-1/2}$. After the transform, the new data sets will have mean $A^T\overrightarrow{\mu}$ and covariance I. In fact:

$$\begin{aligned} & E(A^T\overrightarrow{x} - A^T\overrightarrow{\mu})(A^T\overrightarrow{x} - A^T\overrightarrow{\mu})^T \\ &= (\phi\Lambda^{-1/2})^T \bullet \phi\Lambda\phi^{-1} \bullet \phi\Lambda^{-1/2} \qquad (2.7) \\ &= I, \qquad (2.8) \end{aligned}$$

where we used the fact that $\phi^T = \phi^{-1}$. It can be seen from (2.8) that, after the linear transformation, the covariance is now an identity matrix, which means the transformed data segments are linearly uncorrelated with each other.

In [44], Wang et al. also proposed a similar approach to deal with skewness problem of the distribution. The difference between their approach and the proposed approach is that they treat two sets of data $\{\overrightarrow{X}\}$ and $\{\overrightarrow{Y}\}$ jointly, with the assumption that they follow the same distribution. More specifically, instead of calculating the overall covariance matrix Σ for $\{\overrightarrow{X}\}$ and $\{\overrightarrow{Y}\}$ in (2.6) separately, they calculated the covariance matrix as:

$$\begin{aligned}\Sigma &= \frac{1}{2n-1}[\sum_{i=1}^{n}(\overrightarrow{X_i}-\overrightarrow{\hat{\mu}})(\overrightarrow{X_i}-\overrightarrow{\hat{\mu}})^T \\ &+ \sum_{i=1}^{n}(\overrightarrow{Y_i}-\overrightarrow{\hat{\mu}})(\overrightarrow{Y_i}-\overrightarrow{\hat{\mu}})^T], \end{aligned} \tag{2.9}$$

where $\overrightarrow{\hat{\mu}} = \frac{1}{2n}[\sum_{i=1}^{n}\overrightarrow{X_i} + \sum_{i=1}^{n}\overrightarrow{Y_i}]$.

We believe that this approach is not suitable in our application scenario because it requires that two sets of data share the same statistical property, which may not be practical in fMRI data as Hemodynamic Responses diverse considerably among different brain regions. For this reason, in our analysis here, we choose to analyze $\{\overrightarrow{X}\}$ and $\{\overrightarrow{Y}\}$ separately.

2.3.2 Kernel-Based Estimation of Probability Density Function

The kernel-based estimation is an alternative framework to the assumption based parametric estimation. Originally, the kernel-based approach was introduced to address the problem of the phase uncertainty or origin uncertainty [46]. The basic idea is to calculate the average of kernel functions K on each point that falls into a pre-specified kernel window. More specifically, given a set of data points $\{\epsilon_i | i \in [1, m]\}$, the kernel estimation for the probability

density function at any point x is given by:

$$\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} K\left(\frac{x-\epsilon_i}{h}\right), \tag{2.10}$$

in which h is the bandwidth for the kernel function. For any d-dimensional data vector $\overrightarrow{x} = [x_1, ..., x_j, ..., x_d]$, the estimation function can be extended as:

$$\hat{f}(\overrightarrow{x}) = \frac{1}{mh_1h_2...h_d}\sum_{i=1}^{m} K_h(\overrightarrow{x}, \overrightarrow{\epsilon_i}), \tag{2.11}$$

where

$$K_h(\overrightarrow{x}, \overrightarrow{\epsilon_i}) = \prod_{j=1}^{d} K\left(\frac{x_j-\epsilon_{ij}}{h_j}\right). \tag{2.12}$$

The kernel function K for a continuous variable x is often chosen as the Gaussian function [46]:

$$K(h, \epsilon, x) = \frac{1}{2\sqrt{\pi}} e^{-\frac{(x-\epsilon)^2}{4h^2}}. \tag{2.13}$$

2.3.3 Optimal Kernel Bandwidth Estimation

The bandwidth h for the kernel function has a significant influence on the estimation accuracy. Unlike the choice of k in the kNN estimator, here h can be chosen theoretically according to the Cross Validation Maximum Likelihood (CV-ML) criteria. More specifically,

the bandwidth h is chosen to maximize the *leave-one-out* log likelihood function given by:

$$L = \sum_{i=1}^{m} log\hat{f}_{-i}(\overrightarrow{\epsilon_i}), \tag{2.14}$$

in which the leave-one-out function $\hat{f}_{-i}(\overrightarrow{\epsilon_i})$ is defined as:

$$\hat{f}_{-i}(\overrightarrow{\epsilon_i}) = \frac{1}{(m-1)h_1h_2...h_d} \sum_{j \neq i} K_h(\overrightarrow{\epsilon_i}, \overrightarrow{\epsilon_j}). \tag{2.15}$$

It can be shown that pdf estimation under this criteria could yield a result that will approach the real density in a *Kullback-Leibler entropy* sense [46].

The maximization problem for the likelihood function L in (2.14) belongs to a category of optimization problems, and can be solved reliably using various algorithms [46,47]. In this chapter, we choose the downhill simplex method, because the complexity of the log likelihood makes the differentiate operation quite computationally infeasible, and classical second-order optimization method like Quasi-Newton algorithm is difficult to be implemented. More details about downhill simplex method could be found in [47]. The algorithm is implemented as follows:

At the initial step, randomly construct a simplex of $d+1$ vertices in a d dimensional definition domain of the function $L(h)$. Order the vertices according to the function values, i.e., $L(h_1) \geq L(h_2) \geq ... \geq L(h_{d+1})$. Then, calculate the centroid point h_0 and iteratively execute following steps: Reflection, Expansion, Contraction and Reduction.

- Reflection: A reflected point h_r is computed as $h_0 + \alpha(h_0 - h_{d+1})$. If the reflected point is better than h_d but worse than h_1, then reconstruct the simplex by replacing h_{d+1} with h_r.

- Expansion: If h_r outperforms h_1, compute the expanded point $h_e = h_0 + \gamma(h_0 - h_{d+1})$. If h_e is better than h_r, construct the new simplex by replacing h_{d+1} with h_e; else, replace h_{d+1} with h_r.

- Contraction: If h_r is inferior to h_d, compute the contracted point $h_c = h_0 + \rho(h_0 - h_{d+1})$. Check if h_c is better than h_{d+1}. If yes, reconstruction the simplex by replacing h_{d+1} with h_c; else, execute the Reduction step.

- Reduction: Replace all but the best point with $h_i = h_1 + \sigma(h_i - h_1)$.

The values of α, γ, ρ and σ are set to be 1, 2, $-1/2$ and $1/2$, respectively. The algorithm ends after a predefined number of iterations. It can be guaranteed that when the function is locally smooth, an optimal solution could be reached [47].

2.3.4 MI Estimation Through Monte Carlo Integration

The Monte Carlo integration method is used here to calculate the MI after the probability distribution function has been obtained. For notation simplicity, let $i(\overrightarrow{x}, \overrightarrow{y}) = f_{XY}(\overrightarrow{x}, \overrightarrow{y}) log \frac{f_{XY}(\overrightarrow{x}, \overrightarrow{y})}{f_X(\overrightarrow{x}) f_Y(\overrightarrow{y})}$. Then, to calculate the MI $I = \int\limits_{R^d} \int\limits_{R^d} i(\overrightarrow{x}, \overrightarrow{y}) d\overrightarrow{x}\, d\overrightarrow{y}$, the algorithm uniformly samples a finite space with a volume of V, and generates P samples $\{(\overrightarrow{x_i}, \overrightarrow{y_i}), i \in [1, P]\}$. The mutual information, then, can be estimated as:

$$I_P \approx \frac{V}{N} \sum_{i=1}^{P} i(\overrightarrow{x_i}, \overrightarrow{y_i}). \tag{2.16}$$

Since the definition domain of a Gaussian function is infinite, sampling on the whole space is impossible. Therefore, we limit the sampling space within the interval $[\mu - 3\sigma, \mu + 3\sigma]$, where μ and σ denote the mean and standard deviation of the Gaussian function, respectively.

Convergence: It can be shown by the Law of Large Numbers (LLN) [48]: as P goes to infinity, the approximation in (2.16) will converge to the real value of I, i.e., $\lim_{P\to\infty} I_P = I$.

2.4 Numerical Results

In this section, we will apply the proposed approach to resting state fMRI data collected from both Alzheimer's Disease patients and normal control subjects.

Brain networks operate in a cohesive manner of connections between nodes. A progressive weakening trend of functional connectivities has been observed in the default mode network (DMN) in AD patients [42]. In the following, we will evaluate MI between two regions of DMN, the posterior cingulate cortex (PCC) and superior frontal gyrus (SFG). In the data collection process, eleven patients with mild-to-moderate probable AD and twelve age- and education-matched healthy normal control subjects were recruited to participate in this study. The MRI experiment was conducted on a GE 3T *Signa*® HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil. To study resting-state brain function, echo-planar images, starting from the most inferior regions of the brain, were acquired for 7 minutes with the following parameters: 38 contiguous 3-mm axial slices in an interleaved order, time of echo = 27.7 ms, time of repetition = 2500 ms, flip angle = 80°, field of view = 22 cm, matrix size = 64×64, ramp sampling, and with the first four data points discarded. Each volume of slices was acquired 164 times. Common pre-processing procedures on resting state fMRI data were carried as detailed in [42]. Then, we carried out the proposed approach on the pre-processes fMRI data.

Figure 1 and figure 2 show the calculated mutual information for connections between PCC and SFG. As expected, the connections experienced a decrease in AD patients compared

to NC subjects. Figure 1 shows the probability distribution function of two groups. Because of the existence of outliers in each group, two pdf curves can not be separated completely. However, it is clear that in group level, normal subjects have shown stronger connectivities over the AD patients.

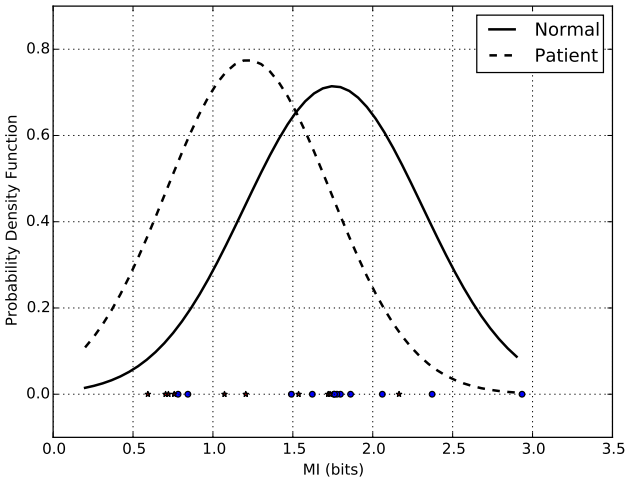


Figure 2.1: Comparison of probability density function between AD patients and NC subjects using MI analysis.

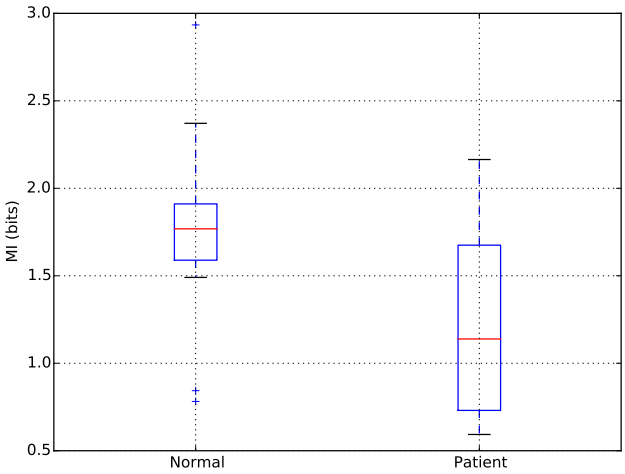


Figure 2.2: Box plots between AD patients and NC subjects using MI analysis.

Figure 2 shows the boxplots for the two groups. It can be seen that the median suffered a 35.6%'s decrease in AD patients compared to NC subjects. This result is consistent with

previous clinical research finding. Significant difference on the MI between PCC and SFG was found between the two groups based on independent two-sample t test ($p = 0.04$).

2.5 Summary

In this chapter, we considered the measurement of functional brain connectivities using mutual information. We proposed a novel approach for the estimation of MI, which was composed of three major components: de-correlation, kernel based estimation of probability density function and Monte Carlo Integration for MI estimation. The analysis results obtained using the proposed method were consistent with clinical observations in the AD data sets.

Chapter 3

Causality Analysis of fMRI Data Based on The Directed Information Theory Framework

This chapter aims to conduct fMRI based causality analysis in brain connectivity by exploiting the directed information (DI) theory framework. Unlike the well known Granger Causality (GC) analysis, which relies on the linear prediction technique, the directed information theory framework does not have any modeling constraints on the sequences to be evaluated and ensures estimation convergence. Moreover, it can be used to generate the Granger Causality graphs. In this chapter, first, we introduce the core concepts in the directed information framework. Second, we present how to conduct causality analysis using directed information measures between two time series. We provide the detailed procedure on how to calculate the DI for two finite time series. The two major steps involved here are optimal bin size selection for data digitization, and probability estimation. Finally, we demonstrate the applicability of DI based causality analysis using both the simulated data and experimental fMRI data, and compare the results with that of the Granger Causality analysis. Our analysis indicates that GC analysis is effective in detecting linear or nearly linear causal relationship, but may have difficulty in capturing nonlinear causal relationships.

On the other hand, DI based causality analysis is more effective in capturing both linear and non-linear causal relationships. Moreover, it is observed that brain connectivity among different regions generally involves dynamic two-way information transmissions between them. Our results show that when bidirectional information flow is present, DI is more effective than GC to quantify the overall causal relationship.

3.1 Introduction

Causality analysis provides important information on how brain regions interact with each other to accomplish a cognitive task [6]. In general, causality analysis tries to determine whether the values of one time series X is useful in predicting the future values of another time series Y. Here, we will first briefly revisit the work on causality analysis in literatures, including *Granger Causality* (GC), *Bayesian Network*, *Dynamic Causal Modeling* (DCM) and *Transfer Entropy* (TE). Then, we will introduce the Directed Information (DI) framework, explain why we adopt it, and how to apply it for causality analysis.

3.1.1 Some Representative Techniques on Causality Analysis

Granger Causality The first practical causal analysis framework was proposed by Granger in 1969 [7]. The fundamental idea is, if two signals X and Y form a causal relationship, then, instead of using the past value of Y alone, the information contained in the past values (or lagged values) of X will help to predict Y. More specifically, the calculation of Granger Causality is based on the autoregressive or linear prediction models. Suppose $\mathbf{X}^n = [X_1, X_2, ...X_n]$ and $\mathbf{Y}^n = [Y_1, Y_2, ...Y_n]$ are two time series. The most commonly used method in Granger Causality analysis is to compare the following two prediction errors e_i

and $\tilde{e_i}$:

$$Y_i = \sum_{j=1}^{L} a_j Y_{i-j} + e_i \tag{3.1}$$

$$Y_i = \sum_{j=1}^{L} [b_j Y_{i-j} + c_j X_{i-j}] + \tilde{e_i} \tag{3.2}$$

Here, e_i is the error of prediction Y_i based only on the previous value of Y, $(Y_{i-1}, ...Y_{i-L})$, and $\tilde{e_i}$ is the error of predicting Y_i based on both the previous values of Y, $(Y_{i-1}, ...Y_{i-L})$, and the previous values of X, $(X_{i-1}, ...X_{i-L})$. In practical analysis, Granger Causality can be tested using a nested model comparison based on the F statistics [49]. If $\tilde{e_i}$ is much smaller than e_i, that is, the introduction of the previous value of X can improve the prediction accuracy, then we say there is a Granger Causal relationship between X and Y.

Since 1990s, there have been growing interests in the use of Granger Causality analysis to identify causal interactions in neuroscience [8]. An early exploitation of GC in neuroscience was carried out by Bernasconi et al. in electrophysiological data [50]. Their paper verified the applicability of GC for electrophysiological data, particularly EEG measurements. Goebel et al. presented an application of GC on the fMRI data [51, 52]. They applied the GC approach to a dynamic sensorimotor mapping paradigm. Bressler et al. applied GC analysis to examine the blood oxygen level-dependent (BOLD) time series corresponding to the top-down control signals from the frontal and parietal cortex [53]. Hu et al. applied GC analyses on fMRI data to evaluate the causal relationship among specific brain regions, so as to understand the impact of amnesic mild cognitive impairment (aMCI) on brain connectivity [10]. Wen et al. carried out simulations on neural signals to examine GC in both neural level (neural GC) and fMRI level (fMRI GC) [54, 55].

David et al. applied GC (together with Dynamic Causal Modeling) in a combination of fMRI and EEG data [11]. Their experiments showed that as the hemodynamics (i.e., the blood flow or the circulation) vary from region to region, GC may not be applied directly on the fMRI signals. However, when the hemodynamic effects were explicitly removed, GC test can perform effective causality analysis in linear relationships.

As a well-known technique, the validity and computational simplicity of Granger Causality have been widely recognized. However, it has also been noticed that GC relies heavily on the linear prediction method. When there exist instantaneous and/or strong nonlinear interactions between two regions, GC analysis may lead to invalid results [11]. To address this problem, several approaches on nonlinear Granger Causality have been proposed in literature. For example, in [56], Bezruchko et al. proposed an autoregression model constructed in the form of a polynomial. More recently, Marinazzo et al. proposed a method to generalize GC to include the nonlinear case using the kernel technique [57]. The copula approach has been applied for GC assessment in [9, 58]. A comprehensive discussion on nonlinear GC could be found in [59].

Bayesian Network In [12], J. Pearl summarized the framework of Bayesian Network for causal inference. The argument behind it is that: if a causal relationship exists between two factors X and Y, the introduction of factor X may change the distribution of another factor Y. That is, $P(Y|X) \neq P(Y)$.

Since 2000, the analyses based on Bayesian Networks have demonstrated successful applications [13,14]. Luessi et al. [14] modified the Bayesian Network and applied it to fMRI data by incorporating the vector autoregressive model used in GC. Their result was in consistent with that of the GC analysis. From a general perspective, the vector autoregressive model based Bayesian Network framework can be regarded as a variation of the Granger Causality

analysis.

Dynamic Causal Modeling In 2003, Friston proposed the framework of *Dynamic Causal Modeling* (DCM) to describe the general interactions among a group of brain regions [15]. DCM assumes that the invisible neurostate x, the (external) input u, the parameter θ that characterizes the connection between two brain regions, and the independent noise ω form a complex dynamic system that could be described by the following equations:

$$\dot{x} = f(x, u, \theta) \text{ and } y = L(\theta, h(x)) + \omega, \tag{3.3}$$

where $h(x)$ represents a cascade of differential equations which connect the neurostate to changes in blood volume and deoxyhemoglobin content, and L represents a non-linear output function which relates θ and $h(x)$ to the observed BOLD signal y [16].

With the help of EM algorithm, DCM has been attempted on both fMRI and EEG data [16]. Some concerns with this framework are [17]: (1) As the observation model in DCM is non-linear, estimating the latent variable that describes the neuronal activity could be quite difficult. (2) DCM is a confirmatory approach, for which the users have to start with different connectivity describing models, then rank them based on an approximation of the model evidences.

Transfer Entropy Another widely applied causal measurement in neuroscience is *Transfer Entropy* (TE). TE was introduced in 2000 by Schreiber [18]. It measures the decrease of entropy in one signal Y after another signal X has been observed:

$$T_{X \to Y} \triangleq H\left(Y_t \mid Y_{t-1:t-L}\right) - H\left(Y_t \mid Y_{t-1:t-L}, X_{t-1:t-L}\right) \tag{3.4}$$

in which H denotes the entropy operator, $Y_{t-1:t-L} = [Y_{t-L}, ..., Y_{t-1}]$,$X_{t-1:t-L} = [X_{t-L}, ..., X_{t-1}]$.

Similar to GC, TE measures how much additional information the past values of process X contains about the future observations of Y, given that we already knew the past values of Y. The quantity measured by TE is the amount of predictive information rather than the size of causal effect or coupling strength. In [23], it was pointed out that TE can differentiate between interactions in the process of information storage and those in the process of information transfer.

The first exploration of applying transfer entropy in causality description was conducted by Sporns et al. on the sensorimotor network in 2006 [19]. Vicente et al. [18] applied TE in the magnetoencephalography (MEG) data and showed that TE was an effective metric for non-linear connectivity, especially for sensor-level MEG signals. Lizier et al. [21] developed a framework that combined multivariate mutual information and transfer entropy together. They used TE to analyze fMRI time series to detect the directed flow of information between brain regions involved in a visuo-motor tracking task.

As an information theoretic framework, a major advantage of Transfer Entropy is that it does not does not rely on any model assumptions of the signals. However, current algorithms on TE estimation have not been proved to be convergent [22]. Also, in [18, 23], it was shown that the amplitude of transfer entropy could not accurately quantify the strength of influence between brain regions.

3.1.2 Proposed Approach: DI Based Causality Analysis

In the discussions above, we revisited some representative methods on causality analysis. These methods are either limited to an existing model on the time series under investigation, or cannot guarantee convergence or validity in practical estimation. In an effort to overcome

these weaknesses, we propose to adopt the directed information theory framework.

Directed Information Given two random sequences $\mathbf{X}^n$ and $\mathbf{Y}^n$, the directed information from $\mathbf{X}^n$ to $\mathbf{Y}^n$ is defined as a sum of some conditional mutual information:

$$I(\mathbf{X}^n \to \mathbf{Y}^n) \stackrel{\triangle}{=} \sum_{i=1}^{n} I(\mathbf{X}^i; Y_i | \mathbf{Y}^{i-1}), \tag{3.5}$$

where $\mathbf{X}^i = [X_1, X_2, ..., X_i]$, $\mathbf{Y}^i = [Y_1, Y_2, ..., Y_i]$. First introduced by Massey to study communication channel with feedback [24], DI has been proved to be an effective tool for network analysis in communications [60] and neuroscience [25, 26]. As an information theoretical metric, DI shares some similarities with Transfer Entropy. Both of them do not rely on any model assumptions of the signals. Moreover, it was pointed in [29] and [23] that: as time goes to infinity, DI may approximate the rate of transfer entropy.

The DI framework is adopted here for the following reasons: (i) It is a universal method. Unlike GC, which mainly relies on the linear prediction theory, or linear modeling for the involved parameters, the DI based causality analysis does not have any modeling constraint on the sequences to be evaluated, hence can be used to characterize more general relationships. (ii) It is well defined with specific physical meaning. Recall that the amplitude of TE cannot reflect the strength of dependence between brain regions, the amplitude of DI reflects the information flow from $\mathbf{X}^n \to \mathbf{Y}^n$, hence has a clear physical meaning. (iii) It has been shown in [29] that the Granger Causality graphs could be obtained using the DI framework. As can be seen, directed information theory provides an adequate framework for the connectivity inference problems in neuroscience applications.

In literature, there has been a limited number of references on applications of directed information in neuroscience [22, 28]. Quinn et al. applied DI in studying neuron spike

recording by introducing a Markov Process model for the signal. Liu et al. applied DI to the EEG data and compared the result with that of GC. Their conclusion was that DI based approach could be superior to GC in capturing the instantaneous and nonlinear causal relationship from EEG data.

Chapter Overview In this chapter, first, we introduce the core concepts in the directed information framework. Second, we present how to conduct causality analysis using direct information measures between two time series. We provided the detailed procedure on how to calculate the DI for two finite time series. The two major steps involved here are optimal bin size selection for data digitization, and probability estimation. Finally, we demonstrate the effectiveness of directed information based causality analysis using both the simulated data and experimental fMRI data, and compare the results with that of the Granger Causality analysis. For practical evaluation, we collected both stimulation fMRI data with a well defined block-design scene-object fMRI paradigm [61, 62] and resting-state fMRI data. Our analysis indicates that GC analysis is effective in detecting linear or nearly linear causal relationship, but has difficulty in capturing nonlinear causal relationships. On the other hand, DI based causality analysis can be used to capture both linear and non-linear causal relationships. Moreover, it is observed brain connectivity among different regions generally involves dynamic two-way information transmissions between them. Our results show that when bidirectional information flow is involved, DI is a more effective than GC to quantify the overall causal relationship.

3.2 Methods

Let uppercase letters (X,Y,...) denote random variables, and lowercase letters (x,y,...) the possible values they can acquire. For $n \in N$, define $\mathbf{X}^n = [X_1, X_2, ...X_n]$, where X_i is the ith sample. Each X_i is a random variable taken from the same finite alphabet Ω, with cardinality $|\Omega|$. For any $x_i \in \Omega$, $P(x_i) = Prob\{X_i = x_i\}$ denotes the probability for X_i to take the value x_i; and $P_{X_i|X^{i-1}}(x_i|\mathbf{x}^{i-1}) = Prob\{X_i = x_i|\mathbf{x}^{i-1} = [x_1, x_2, ...x_{i-1}]\}$ the conditional probability that the current sample X_i is x_i, given that the previously observed sequence is $\mathbf{x}^{i-1} = [x_1, x_2, ..., x_{i-1}]$. Without extra explanation, the log function $log(*)$ denotes the base 2 logarithm.

3.2.1 Core Concepts in the Directed Information Framework

Entropy A fundamental concept in information theory is *entropy*, which is a measure of uncertainty. For a random variable X, the entropy of X is defined as:

$$H(X) = -\sum_{x_i \in \Omega} P(x_i) log P(x_i). \tag{3.6}$$

The entropy of a random variable X represents the minimum average number of bits needed for loseless encoding of each symbol of X.

For a random sequence $\mathbf{X}^n$, the entropy could be calculated according to the chain rule:

$$\begin{aligned} H(\mathbf{X}^n) &= H(X_1) + H(X_2|X_1) + ... + H(X_n|\mathbf{X}^{n-1}) && (3.7) \\ &= \sum_{i=1}^{n} H(X_i|\mathbf{X}^{i-1}). && (3.8) \end{aligned}$$

Mutual Information Mutual Information (MI) measures the decrease of uncertainty of one random variable after observing another one. The definition of mutual information between two random variables X and Y is:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (3.9) \\ &= \sum_{x,y} P(x,y) log \frac{P(x,y)}{P(x)P(y)}, && (3.10) \end{aligned}$$

here $H(Y)$ is the total uncertainty (or information) in Y, and $H(Y|X)$ is the uncertainty (or information) left in Y after X is observed. It's clear that mutual information is a symmetric measurement: for two random variables X and Y, $I(X;Y) = I(Y;X)$. Mutual information measures the dependence between two random variables. If $X = Y$, then $H(Y|X) = 0$, and $I(X;Y) = H(X) = H(Y)$. If X and Y are independent, then $I(X;Y) = 0$. Unlike the Pearson Correlation Coefficient which only measures the linear dependence between two random variables, mutual information includes both linear and non-linear dependence. For this reason, in recent years, there has been a growing interest in applying MI to neuroscience to measure the coupling strength among different brain regions or groups of neurons [63,64].

For two random sequences $\mathbf{X}^n$ and $\mathbf{Y}^n$, the mutual information could also be calculated using a chain rule:

$$\begin{aligned} I(\mathbf{X}^n;\mathbf{Y}^n) &= H(\mathbf{Y}^n) - H(\mathbf{Y}^n|\mathbf{X}^n) && (3.11) \\ &= \sum_{i=1}^{n} H(Y_i|\mathbf{Y}^{i-1}) - H(Y_i|\mathbf{Y}^{i-1},\mathbf{X}^n) && (3.12) \\ &= \sum_{i=1}^{n} I(\mathbf{X}^n;Y_i|\mathbf{Y}^{i-1}). && (3.13) \end{aligned}$$

Directed Information Since both correlation and mutual information are non-directional,

in 1990, Massey refined Markov's work and proposed the concept of *directed information* (DI) [24] aiming to measure the directed information flow from one random sequence to another. Given $\mathbf{X}^n$ and $\mathbf{Y}^n$, the directed information from $\mathbf{X}^n$ to $\mathbf{Y}^n$ is defined as:

$$I(\mathbf{X}^n \to \mathbf{Y}^n) = \sum_{i=1}^{n} I(\mathbf{X}^i; Y_i | \mathbf{Y}^{i-1}). \tag{3.14}$$

Recall that the *causally conditional entropy* is defined as:

$$H(\mathbf{Y}^n || \mathbf{X}^n) = \sum_{i=1}^{n} H(Y_i | \mathbf{Y}^{i-1}, \mathbf{X}^i), \tag{3.15}$$

then it follows from (3.14) and (3.15) that:

$$I(\mathbf{X}^n \to \mathbf{Y}^n) \;\; = \;\; H(\mathbf{Y}^n) - H(\mathbf{Y}^n || \mathbf{X}^n). \tag{3.16}$$

On the other hand,

$$\begin{aligned} I(\mathbf{X}^n; \mathbf{Y}^n) &= H(\mathbf{Y}^n) - H(\mathbf{Y}^n | \mathbf{X}^n) && (3.17)\\ &= I(\mathbf{X}^n \to \mathbf{Y}^n) + I(\mathbf{Y}^{n-1} \to \mathbf{X}^n) && (3.18)\\ &= I(\mathbf{X}^{n-1} \to \mathbf{Y}^n) + I(\mathbf{Y}^{n-1} \to \mathbf{X}^n) \\ &+ \sum_{i=1}^{n} I(X_i; Y_i | \mathbf{X}^{i-1}, \mathbf{Y}^{i-1}). && (3.19) \end{aligned}$$

In (3.19), the first term $I(\mathbf{X}^{n-1} \to \mathbf{Y}^n)$ specifies the directed information flow from $\mathbf{X}^n$ to $\mathbf{Y}^n$, the second term $I(\mathbf{Y}^{n-1} \to \mathbf{X}^n)$ specifies the *reverse directed information* from $\mathbf{Y}^n$ to $\mathbf{X}^n$, and the third one represents the conditional mutual information shared by both $\mathbf{X}^n$ and $\mathbf{Y}^n$. DI reflects directional and interactive influence between two random sequences, and

has recently been applied to characterize the connectivity between different brain regions [22, 28].

Directed Information and Granger Causality Granger Causality (GC) [7, 22] has long been used in identifying causal relations between two random series. The main idea behind the Granger Causality analysis is that, if one random process X causally influences another random variable Y, then the knowledge of previous values of X will help to decrease errors in predicting future values of Y. The calculation of Granger Causality is based on the autoregressive or linear prediction models.

As can be seen from (3.1), the traditional Granger Causality analysis relies on the linear prediction models. Its nonlinear extensions generally still rely on the "linear-in-the-parameter" modeling [29]. The directed information, on the other hand, contains no requirement on models, and hence provides the freedom to characterize more generalized relationships. In [29], Amblard and Olivier investigated the relationship between directed information and Granger Causality, and showed that Granger Causality graphs could be obtained using directed information. They further pointed out that, directed information theory provided an adequate information theoretical framework for the connectivity inference problems in neuroscience applications.

3.2.2 Directed Information Calculation and Causality Analysis

Developing practical estimators for directed information measures is always a challenging problem. Over the last two decades, a limited number of directed information estimators have been purposed. In [22], Quinn et al. utilized DI to infer causality based on neural spike recordings. Their estimator is built upon the assumption that the random sequences corresponding to spike recordings form stationary ergodic Markov processes and adopts the

simple General Linear Model(GLM). As a result, the causally conditional entropy can be simplified as $E[g_{JK}(Y_{l-J}^{l}, X_{l-(K-1)^{l}})]$, in which g is a log-probability function, and J and K denote the orders of the Markov Processes. Although this method is not model free, its validity has been verified in discovering causal relations among groups of neurons. Verdu et al. purposed a *K-nearest neighbor*(KNN) based estimator for KL distance [44]. This idea has been adopted in the work by Chai et al. to estimate entropy and mutual information [3]. Theoretically, directed information could be estimated after obtaining other information theoretical measures like entropy and mutual information. However, the KNN estimator is based on the assumption that samples in the random sequences are independent and identically distributed (i.i.d). Also, this approach requires a large number of data points.

In this chapter, we calculate the directed information $I(\mathbf{X}^n \rightarrow \mathbf{Y}^n)$ by exploiting the method initiated by Weissman et al. [30]. This approach is universal, and not limited to any modeling assumptions on the random sequences. There are two parts in the estimation.

Part I This part has three steps.

1. Estimate $H(\mathbf{Y}^n)$:

$$
\begin{aligned}
H(\mathbf{Y}^n) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{y_{i+1}} P(y_{i+1}|\mathbf{y}^i) log \frac{1}{P(y_{i+1}|\mathbf{y}^i)} && (3.20)\\
&\approx \frac{1}{n}\sum_{i=1}^{n}\sum_{x_{i+1},y_{i+1}} P(x_{i+1}, y_{i+1}|\mathbf{x}^i, \mathbf{y}^i) \\
&\times \; log \frac{1}{P(y_{i+1}|\mathbf{y}^i)}. && (3.21)
\end{aligned}
$$

2. Estimate the $H(\mathbf{Y}^n||\mathbf{X}^n)$:

$$\begin{aligned} H(\mathbf{Y}^n||\mathbf{X}^n) &= \frac{1}{n}\sum_{i=1}^{n} log\frac{1}{P(y_i|\mathbf{y}^{i-1},\mathbf{x}^i)} && (3.22)\\ &\approx -\frac{1}{n}\sum_{i=1}^{n}\sum_{x_{i+1},y_{i+1}} P(x_{i+1},y_{i+1}|\mathbf{x}^i,\mathbf{y}^i) \\ &\times \ log\frac{P(x_{i+1},y_{i+1}|\mathbf{x}^i,\mathbf{y}^i)}{P_{x_{i+1}|X^iY^i}(x_{i+1}|\mathbf{x}^i,\mathbf{y}^i)}. && (3.23)\end{aligned}$$

3. It then follows that $I(\mathbf{X}^n \to \mathbf{Y}^n)$ can be estimated as $H(\mathbf{Y}^n) - H(\mathbf{Y}^n||\mathbf{X}^n)$.

Part II For validity of this estimation, it has been shown in the work by Weissman et al. [30] that as $n \to \infty$, $\hat{I}(\mathbf{X}^n \to \mathbf{Y}^n)$ converges to the expected real value of $I(\mathbf{X}^n \to \mathbf{Y}^n)$. To measure the causal influence of one region on another, we resort to $D_n = I(\mathbf{X}^n \to \mathbf{Y}^n) - I(\mathbf{Y}^n \to \mathbf{X}^n)$. Using (3.19), we have:

$$\begin{aligned} D_n &= I(\mathbf{X}^n \to \mathbf{Y}^n) - I(\mathbf{Y}^n \to \mathbf{X}^n) && (3.24)\\ &= [I(\mathbf{X}^n;\mathbf{Y}^n) - I(\mathbf{Y}^n \to \mathbf{X}^n)] \\ &- [I(\mathbf{X}^n;\mathbf{Y}^n) - I(\mathbf{X}^n \to \mathbf{Y}^n)] && (3.25)\\ &= I(\mathbf{X}^{n-1} \to \mathbf{Y}^n) - I(\mathbf{Y}^{n-1} \to \mathbf{X}^n). && (3.26)\end{aligned}$$

As shown in (3.24), D_n is the difference of two directed information between $\mathbf{X}^n$ and $\mathbf{Y}^n$. If D_n is positive, that is, $I(\mathbf{X}^{n-1} \to \mathbf{Y}^n) > I(\mathbf{Y}^{n-1} \to \mathbf{X}^n)$, then we say that $\mathbf{X}^n$ shows more influence on $\mathbf{Y}^n$, and can be interpreted as the causal driver during the connectivity; otherwise we say $\mathbf{Y}^n$ shows more influence on $\mathbf{X}^n$.

To make the result more comparable, we will use $\gamma = D_n/I(\mathbf{X}^n;\mathbf{Y}^n)$ instead of D_n. Clearly, $\gamma \in [-1,1]$. When $|\gamma|$ approaches 1, it can be said with high confidence that there

does exist a causal influence between two stochastic processes; while if $|\gamma|$ is adjacent to 0, it is more likely that no clear causal relationship exists, or the samples in random sequences are subject to strong noises. Therefore, in the simulations, a threshold based method is developed in interpreting the γ metric.

The causality analysis of two brain regions helps us to understand which region is more likely to be the causal driver during a particular connectivity. However, we would like to point out that brain connectivity between two different regions generally involves dynamic two-way information transmission between them, rather than a fixed one-way source to destination relationship.

3.2.3 Practical evaluation

Based on our discussions in Section 2.2, for practical evaluation of the directed information, the main point is how to estimate the probabilities involved in (3.20) accurately from discrete time data or observations.

There are two major issues in probability estimation. *First*, how to choose the optimal bin size for digitization. *Second*, how to estimate the probability of a particular sequence after the random sequence has been mapped into a series of symbols.

3.2.3.0.1 Optimal Bin Size for Time Series Digitization The first problem in directed information estimation of fMRI signals is digitization. If the bin size is too large, then it results in considerable approximation error, and cannot reflect the true data distribution accurately. If the bin size is too small, then the number of samples falling into each bin tends to be 0 or 1 due to the very limited data length. As a result, the probability estimation become inaccurate. In fact, when the bin size is too large or too small, the estimated DI

will approach zeros, due to the limited data length. Here we choose to use the bin size that minimizes the Integrated Mean Square Error (IMSE) between the estimated probability and its true value [65].

The random sequences in those fMRI signals acquire values in the real number field, while the directed information probability estimation can be carried out only for discrete, finite-size alphabets. Hence, mapping real-valued numbers into the symbols within a finite alphabet is the first step for further procedures. In the digitization process, we adopt the traditional histogram methods to estimate the probability of the data points falling into each of the bins representing symbols in the alphabet.

Suppose the true *probability distribution function* (pdf) of a random variable X is $f(x)$,and the estimated pdf is $\hat{f}(x)$, the IMSE is defined as:

$$IMSE = \int E\{\hat{f}(x) - f(x)\}^2 dx. \tag{3.27}$$

For a random sequence of length n, the optimal bin size that minimizes the IMSE is given by:

$$h_n^* = \{\frac{n}{6}\int_{-\infty}^{\infty} f'(x)^2 dx\}^{-1/3}, \tag{3.28}$$

Assume the random sequence x^n was sampled from a white stochastic process with distribution: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Here μ is the mean and σ the standard deviation of the signal.

Then the optimal bin size for digitization can be obtained:

$$h_n^* = \left\{ \frac{n}{6} \int_{-\infty}^{\infty} [-(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}/\sqrt{2\pi}\sigma^3]^2 dx \right\}^{-1/3} \tag{3.29}$$

$$= \left\{ \frac{n}{6} \int_{0}^{\infty} u^2 e^{-u^2} du/\pi\sigma^3 \right\}^{-1/3} \quad (\text{Let } u = \tfrac{x-\mu}{\sigma}) \tag{3.30}$$

$$= \left\{ \frac{n}{6} \frac{1}{4\sqrt{\pi}\sigma^3} \right\}^{-1/3} \tag{3.31}$$

$$\approx 3.49\sigma n^{-1/3}. \tag{3.32}$$

The results in (3.31) relies on the fact that $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$. In this chapter, the fMRI data sequence is regarded as a Gaussian random process. For digitization, the bin size is chosen according to (3.32).

3.2.3.0.2 Probability Estimation After digitization, real-valued fMRI time courses become a sequence of symbols $\{\mathbf{x}^n\}$ within an alphabet Ω. Denote the alphabet as $\Omega = \{x | x \in \{0, 1, ...M-1\}\}$, where $M = |\Omega|$; $N_0, N_1, ...N_{M-1}$ represent the counts for each symbol in the alphabet, respectively. The next step is to estimate the sequence probabilities $P(\mathbf{x}^i)$ and $P(x_i|\mathbf{x}^{i-1})$, $i \in [1, N]$. Here we will resort to the *Krichevsky-Trofimov* (KT) estimator [66] for the probability estimation. The primary reason of using the KT estimator is that this estimator is universal and does not put any specific modeling constraints on the random sequence. It has been shown in [30, 66] that although KT estimator is not optimal, the bias it introduces will be upper bounded.

KT estimator first assigns 0 to the initial value for the sequence probability, and updates this value as the sequence goes on. In each step, the algorithm analyzes the current sequence and generates a list of count numbers for each symbol in the alphabet. Denote this list as

$\{N_0, N_1, ...N_{M-1}\}$. The algorithm goes as follows:

Initialize:

$X = \{\emptyset\}, \{N_0, N_1, ...N_{M-1}\} = \{0, 0, ..., 0\}, P(\emptyset) = 0$

Loop:

while $i \leq n$ **do**

$\mathbf{x}^i \leftarrow \{\mathbf{x}^{i-1}, X_i = j\}, j \in [0, M-1]$;

$\{N_0, N_1, ...N_{M-1}\} \leftarrow \{N_0, N_1, ..., N_j + 1, ...N_{M-1}\}$;

$P(\mathbf{x}^i) \leftarrow P(\mathbf{x}^{i-1}) \times \frac{N_j + 0.5}{N_0 + N_1 + ...N_{M-1} + M/2}$

end while

After estimating the probability $P(\mathbf{x}^n)$, the conditional probability $P_{X_{i+1}|X^i}(x_{i+1}|\mathbf{x}^i)$ can be obtained as $P(\mathbf{x}^{i+1})/P(\mathbf{x}^i)$.

3.3 Materials

3.3.1 Data Acquisition

Fourteen right-handed healthy college students (7 males, 23.4 ± 4.2 years of age) from Michigan State University volunteered to participate in this study and signed consent forms approved by the Michigan State University Institutional Review Board. The experiment was conducted on a 3T GE Signa HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil.

For each subject, fMRI datasets were collected on a visual stimulation condition with a scene-object fMRI paradigm and then on a resting-state condition. The parameters for the fMRI scan were: gradient-echo EPI, 36 contiguous 3-mm axial slices in an interleaved order,

time of echo (TE) = 27.7 *ms*, time of repetition (TR) = 2500 *ms*, flip angle = 80°, field of view (FOV) = 22 *cm*, matrix size = 64×64, ramp sampling, and with the first four data points discarded.

On the visual stimulation fMRI condition, each volume of images were acquired 192 times (8 *min*) while a subject was presented with 12 blocks of visual stimulation after an initial 10 s 'resting period. In a predefined randomized order, the scenery pictures were presented in 6 blocks and the object pictures were presented in other 6 blocks. All pictures were unique. In each block, 10 pictures were presented continuously for 25 s (2.5 s for each picture), followed with a 15 s baseline condition (a white screen with a black fixation cross at the center). The subject needed to press his/her right index finger once when the screen was switched from the baseline to picture condition. Stimuli were displayed in color in full screen on a 1024×768 32-inch LCD monitor (Salvagione Design, Sausalito, CA) placed at the back of the magnet room. The LCD subtended $10.2° \times 13.1°$ of visual angle. On the resting-state fMRI (rs-fMRI) condition, each volume of images were acquired 164 times (6 *min* and 50 *s*) after a subject was informed to relax, keep his/her eyes closed and stay awake throughout the scan. After the above functional data acquisition, high-resolution volumetric T1-weighted spoiled gradient-recalled (SPGR) images with cerebrospinal fluid suppressed were obtained to cover the whole brain with 120 1.5-*mm* sagittal slices, 8° flip angle and 24 *cm* FOV. These images were used to identify anatomical locations.

3.3.2 fMRI Data Pre-processing and Analysis

All stimulus fMRI data pre-processing and analysis for each subject were conducted with AFNI software (Cox, 1996) as described in Henderson et al. [61]. Essentially, slice-timing correction and rigid-body motion correction were carried. Spatial blurring with a full width

half maximum of 4 mm was applied to reduce random noise. Multiple linear regressions (using the "3dDeconvolve" routine in AFNI) were applied on a voxel-wise basis to find the magnitude change when each picture condition was presented, followed with general linear tests to find the statistical significances between stimulus conditions.

The regions of interest (ROI) in this study were defined in the Talairach coordinate space [67]. Regions showing preferential activation to scenes over objects (voxel-based p-value $< 10^{-4}$) in the right and left parahippocampal gyri were defined as the right and left PPA [61]. The right and left V1 ROIs were defined as the regions activated by pictures (voxel-based p-value $< 10^{-10}$) within Brodmann area 17. Because there was a high level of activation at and around V1, a highly conservative p value threshold was chosen to define relatively focal ROIs. The right and left SMC spherical ROIs with 6-mm radius were defined with the centers at (R36, P22, S54) and (L38, P26, S50) correspondingly in the Talairach coordinate space (R = Right, L = Left, P = Posterior, S = Superior). The SMC coordinate locations were defined by Witt et al. [68] and the ROIs were created as in Zhu et al. [69]. The time courses from the stimulation fMRI dataset that were already pre-processed as above were detrended and had their baselines removed also. The spatially averaged time course at each of the above ROIs was generated for the causality analyses discussed later.

The rs-fMRI pre-processing was also processed in AFNI [70] as commonly applied in the field and as described in details in Zhu et al. [69]. Essentially, slice-timing correction and rigid-body motion correction were carried. Spatial blurring with a full width half maximum of 4 mm was applied to reduce random noise. The time courses were detrended and the baselines were removed. Brain global, cerebrospinal fluid and white-matter mean signals were modeled as nuisance variables and removed from the time courses. Finally, the time courses were band-pass filtered to the range of 0.009 Hz – 0.08 Hz. The spatially averaged

time course at each of the above ROIs was generated for the causality analyses discussed later.

3.3.3 Simulated Data

The simulated data was synthesized from the fMRI data corresponding to the primary visual cortex (V1). Recall that the total number of samples in the time series at V1 is 192, as described earlier. Denote this sequence as $\mathbf{x}^n = [x_1, x_2, ..., x_n]$, where $n = 192$. Here we will use two sets of simulated data.

Set I For $i \in [1, 2, ..., 192]$, the first group of simulated data $\mathbf{y}_1^n = [y_{1,1}, y_{1,2}, ..., y_{1,n}]$ was obtained as:

$$y_{1,i} = 0.3 * x_i + 0.2 * x_{i-1}. \tag{3.33}$$

It's clear that $\mathbf{X}^n$ has a causal influence on $\mathbf{Y}^n$. The true fMRI data and the simulated data set I form a linear causal relationship.

Set II For further comparison, we introduced another group that had a nonlinear relationship with the true fMRI data. For $i \in [1, 2, ..., 192]$, the second set of simulated data $\mathbf{y}_2^n = [y_{2,1}, y_{2,2}, ..., y_{2,n}]$ was obtained as:

$$y_{2,i} = \begin{cases} 1 & if \quad x_i \geq 0; \\ 0 & if \quad x_i < 0. \end{cases} \tag{3.34}$$

Clearly, the nonlinear relationship in the second group is difficult to be captured by a linear autoregression model. It has also introduced a significant change in the power level in comparison with true fMRI data.

To make the data more realistic, we added white Gaussian noise to $\mathbf{y}_1^n$ and $\mathbf{y}_2^n$, where the

noise is of zero mean and variance σ_0^2. In the simulation, the Signal-to-Noise Ratio (SNR), which was calculated as $20log(\sigma_x/\sigma_0)$, was set in a range between 4 dB and 20 dB. Before performing fMRI causality analysis, we carry out both DI based causality analysis and the GC analysis over $\mathbf{x}^n$ and $\mathbf{y}^n$ for method validation. The analysis based on simulated data helps to set up a threshold for the metric $\gamma = [I(\mathbf{X}^{n-1} \to \mathbf{Y}^n) - I(\mathbf{Y}^{n-1} \to \mathbf{X}^n)]/I(\mathbf{X}^n; \mathbf{Y}^n)$. Note that the subtraction operation in γ may help to reduce the noise effect in fMRI data. In the Granger Causality analysis, the parameter F-test was adopted to generate the p-value. The max lag in the test was set to be 2.

3.4 Results

In this section, we demonstrate the effectiveness of DI based causality analysis using both stimulated and acquired fMRI data, and compare the results with that of the GC analysis.

3.4.1 Causality between the fMRI data and its descending simulated data

In this subsection, we validate the DI based causality analysis approach using fMRI data and the simulated data generated from it.

DI based Causality Analysis

Fig.1(a) and 1(b) present the DI based causality analysis results corresponding to the fMRI data and the simulated data set I. Fig.1(a) shows the comparison of mutual information $I(\mathbf{x}^n; \mathbf{y}_1^n)$, directed information $I(\mathbf{x}^n \to \mathbf{y}_1^n)$ and the reversed directed information $I(\mathbf{y}_1^n \to \mathbf{x}^n)$. In this example, SNR = 8 dB. Clearly, $I(\mathbf{x}^n \to \mathbf{y}_1^n) > I(\mathbf{y}_1^n \to \mathbf{x}^n)$. As can be seen, the estimated Directed Information shows a clear surplus from $\mathbf{x}^n \to \mathbf{y}_1^n$, which implies that

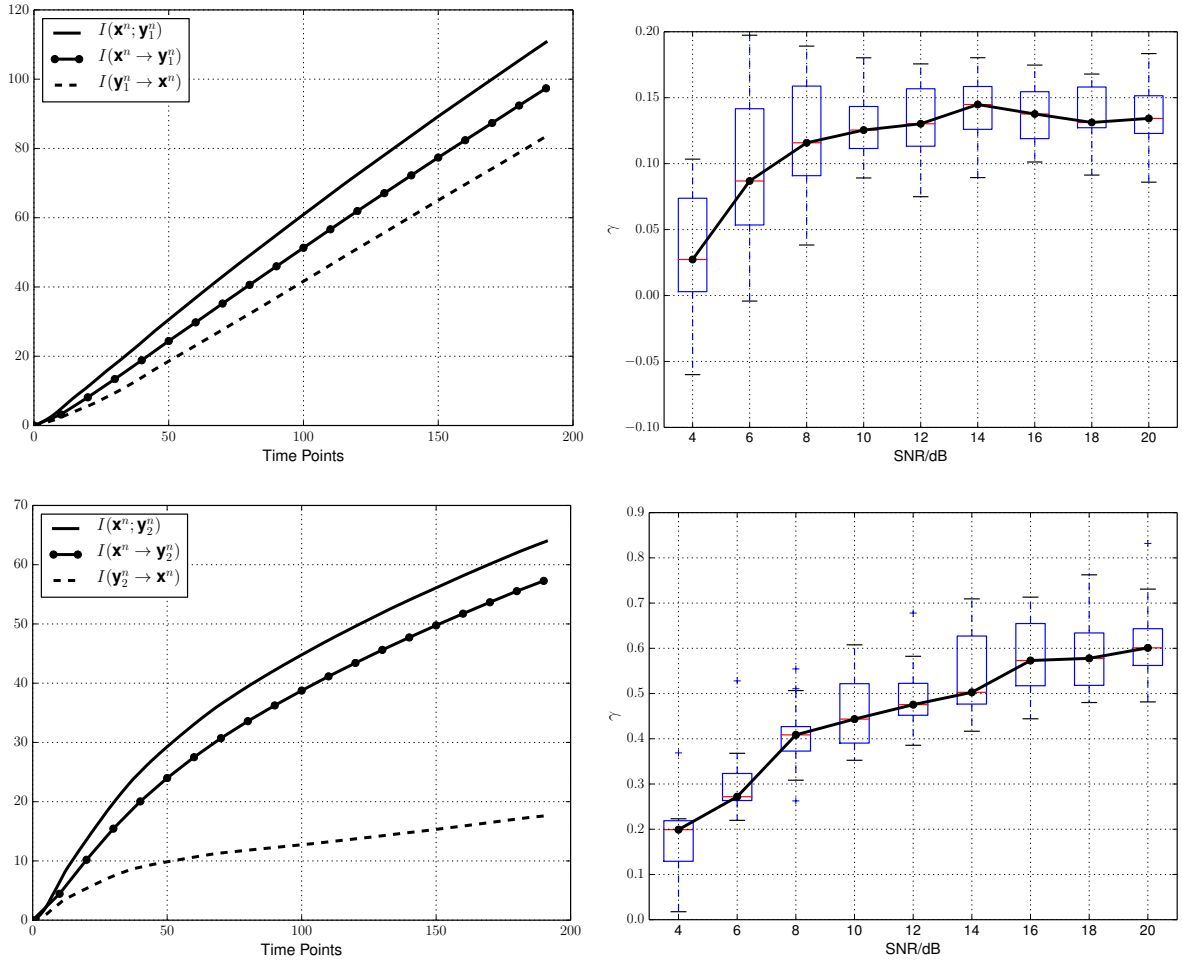


Figure 3.1: Directed Information based test results: mutual information, directed information and the γ metric. Here $\mathbf{x}^n$ denotes the fMRI data, $\mathbf{y}_1^n$ the simulation data set I, and $\mathbf{y}_2^n$ the simulation data set II. (a) MI and DI between $\mathbf{x}^n$ and $\mathbf{y}_1^n$; (b) γ versus SNR corresponding to $\mathbf{x}^n$ and $\mathbf{y}_1^n$; (c) MI and DI between $\mathbf{x}^n$ and $\mathbf{y}_2^n$; (d) γ versus SNR corresponding to $\mathbf{x}^n$ and $\mathbf{y}_2^n$.

there is a causal relationship between $\mathbf{x}^n$ and $\mathbf{y}_1^n$, and $\mathbf{x}^n$ is more likely to be the cause as expected.

Figure 1(b) shows the values of γ versus different SNR levels. As can be seen, when the SNR is above 6 dB (i.e., the noise level is relatively low), we can observe a clear causal relationship from $\mathbf{x}^n \rightarrow \mathbf{y}_1^n$. As the noise level gets higher, i.e. when $SNR < 5$ dB, the causal relationship becomes ambiguous.

The results corresponding to the fMRI data $\mathbf{x}^n$ and the simulated data set II, $\mathbf{y}_2^n$, are shown in Fig.1(c) and 1(d). As can be seen, the results are similar with that corresponding to $\mathbf{x}^n$ and $\mathbf{y}_1^n$. Again, the DI based causality analysis indicates that there is a causal relationship between $\mathbf{x}^n$ and $\mathbf{y}_2^n$, and $\mathbf{x}^n$ is more likely to be the causal part as expected.

It should be noted that the relationship between $\mathbf{y}_1^n$ and $\mathbf{x}^n$ is linear, but the relationship between $\mathbf{y}_2^n$ and $\mathbf{x}^n$ is nonlinear. It can be seen that the DI based causality analysis is effective in the nonlinear case as well. The analysis results are consistent with our priori knowledge that there is a causal relationship between $\mathbf{x}^n$ and $\mathbf{y}_1^n$, and $\mathbf{x}^n$ and $\mathbf{y}_2^n$, with $\mathbf{x}^n$ as the causal side in both cases.

Based on our simulation results on $\mathbf{x}^n$ and $\mathbf{y}_1^n$, which is more similar with true fMRI data than $\mathbf{y}_2^n$, we found that:

- $\gamma \in [0.1, 1]$ implies that $\mathbf{X}$ has more causal influence on $\mathbf{Y}$; accordingly, $\gamma \in [-1, -0.1]$ implies that $\mathbf{Y}$ has more causal influence on $\mathbf{X}$;
- $\gamma \in [-0.1, 0.1]$ implies that there is no clear dominant influence between $\mathbf{X}$ and $\mathbf{Y}$.

Granger Causality Analysis We then apply GC analysis to $\mathbf{x}^n$ and $\mathbf{y}_1^n$, and $\mathbf{x}^n$ and $\mathbf{y}_2^n$, the results are shown in Fig.3.2. Fig.3.2(a) and Fig.3.2(b) show the p-value of the GC analysis corresponding to $\mathbf{x}^n \rightarrow \mathbf{y}_1^n$ and $\mathbf{y}_1^n \rightarrow \mathbf{x}^n$, respectively. As can be seen, there is a clear

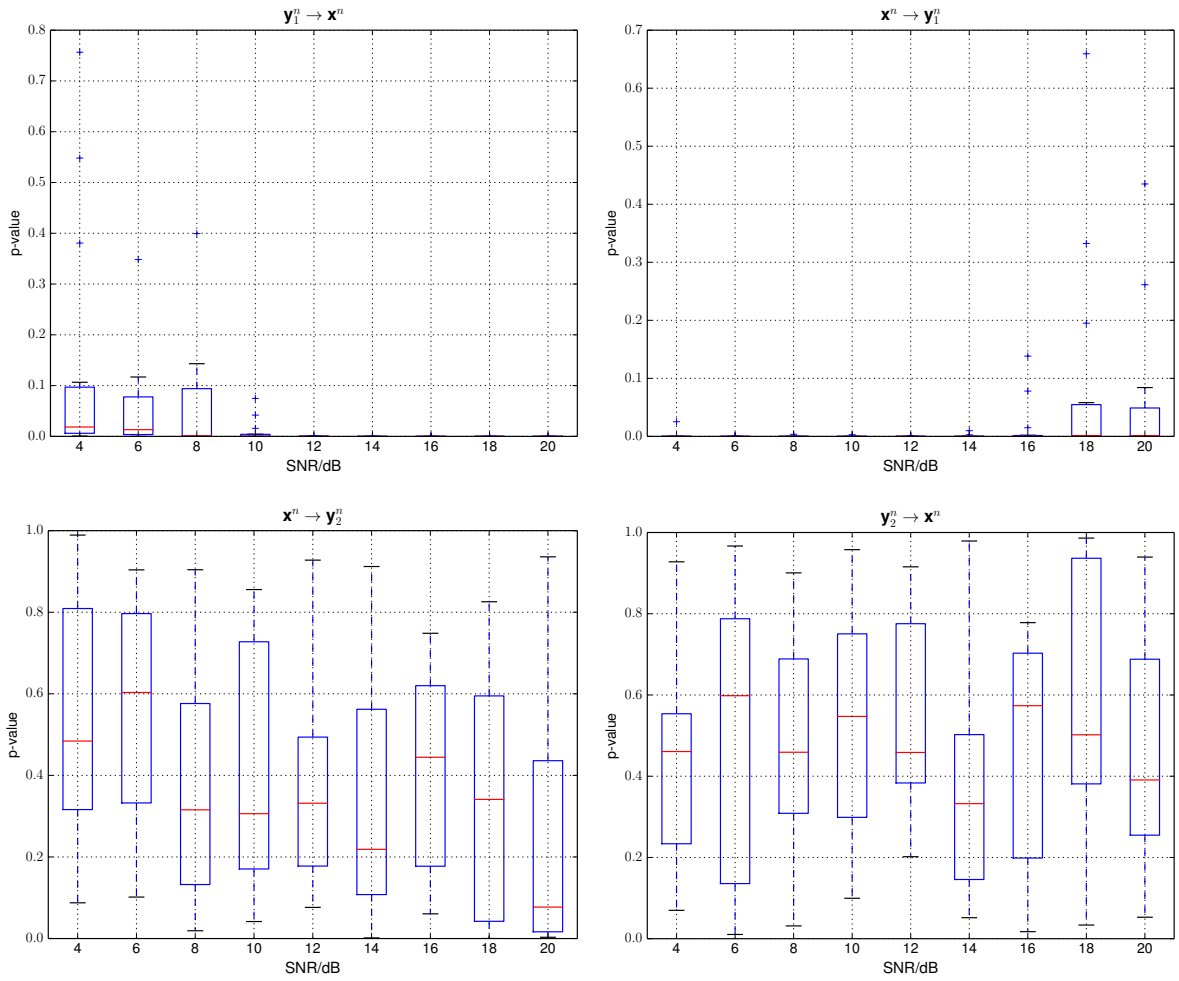
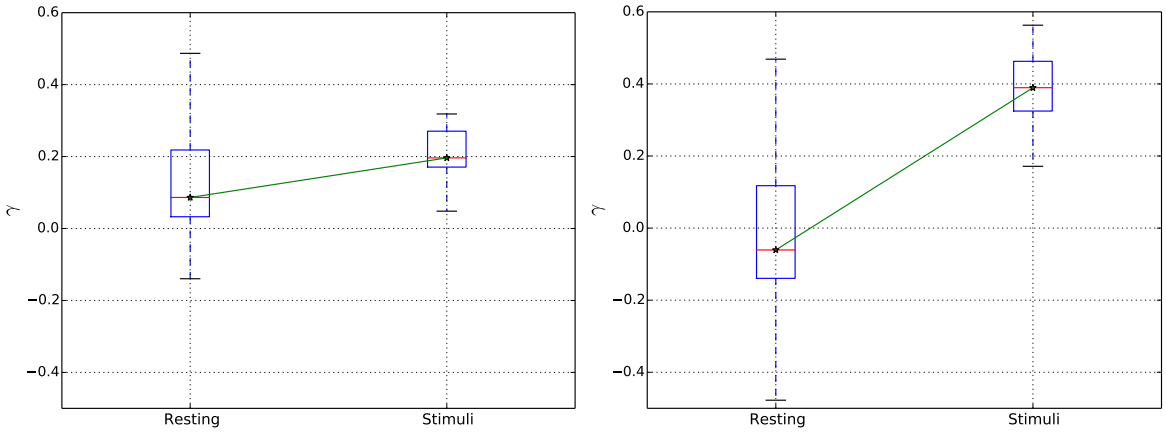


Figure 3.2: Inter-region GC test results based on fMRI data and two sets of simulation data generated from it. (a) test results for the direction $\mathbf{x}^n \to \mathbf{y}_1^n$; (b) test results for the direction $\mathbf{y}_1^n \to \mathbf{x}^n$; (c) test results for the direction $\mathbf{x}^n \to \mathbf{y}_2^n$; (d) test results for the direction $\mathbf{y}_2^n \to \mathbf{x}^n$.

causal relationship from $\mathbf{x}^n \rightarrow \mathbf{y}_1^n$; most medians in the boxes indicate a highly significant casual relationship ($p < 0.0002$, which is much smaller than the commonly accepted p-value 0.01) [71]. These results lead to the expected conclusion that two random sequences $\mathbf{x}^n$ and $\mathbf{y}_1^n$ are Granger Causally related. However, for the test on $\mathbf{x}^n$ and $\mathbf{y}_2^n$, in which the causal relationship is completely nonlinear, the Granger Causality test was not able to capture the causal relationship $\mathbf{x}^n \rightarrow \mathbf{y}_2^n$. In both 3.2(c) and 3.2(d), most medians in the boxes indicate that there is no significant causal relationship between these two time sequences ($0.2 < p < 0.6$), leading to an unexpected conclusion that $\mathbf{x}^n$ and $\mathbf{y}_2^n$ are not causally related.

3.4.2 Causality analysis based only on the Experimental fMRI Data

In this section, we apply both DI based causality analysis and GC analysis to the experimental fMRI data. We collected both stimulation based fMRI data with a well-defined block-design scene-object fMRI paradigm as discussed earlier [61,62], and resting-state fMRI data. Recall that in the scene-object paradigm, subjects viewed blocks of scenery and object pictures. They were asked to press a button once under the right index finger when they saw a block of pictures. We test the robustness of our causality analysis techniques against some expected outcomes: under the stimulation fMRI paradigm, the primary visual cortex (V1) and nearby regions are activated first, followed with activation in the parahippocampal place area (PPA) for higher level scene processing. Some but relatively small activations in the left sensorimotor cortex (SMC) is also expected following V1 activations. Overall sequential neuronal activity is not expected between the right and left homologous regions



(a) V1 and PPA (b) V1 and SMC

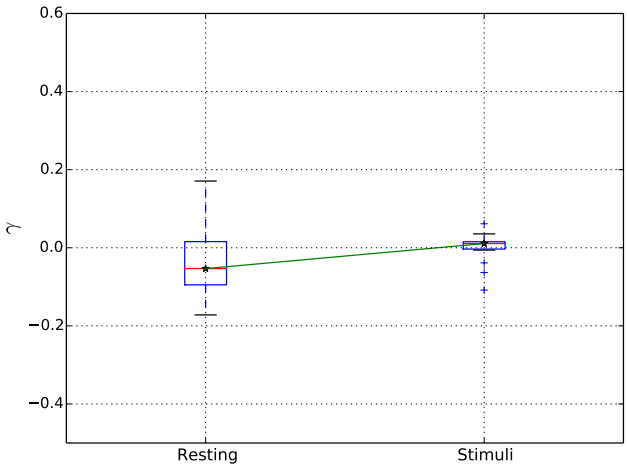
Figure 3.3: Inter-region γ values of directed information based causality analysis.

above. Under the resting-state condition, neuronal activity is not expected to occur in a sequential manner among above regions.

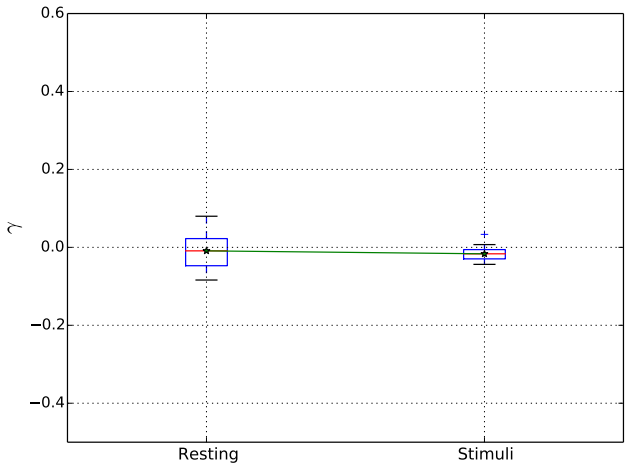
DI based Causality Analysis Here we examine the potential causal relationship between left V1 and left PPA, left V1 and left SMC under both resting state and visual stimulation conditions. The directed information based γ values are shown in Figure 3.3. In resting state condition, the medians of the γ values are within the [-0.1,0.1] region. The left V1 does not exhibit a dominating causal influence over other regions, including left PPA and left SMC. However, under the simulation paradigm, the γ values for left V1 $\rightarrow$ left PPA and left V1 $\rightarrow$ left SMC increase significantly. In other words, under the stimulation, left V1 shows stronger influences over left PPA, as well as left SMC, as expected.

Figure 3.4 shows the γ values between the right and left homologous brain regions in both resting state and stimulus-based state. As can be seen, the median values are well below 0.1. That is, the directed information based causality analysis indicates that there is no dominating influence between the left and right homologous brain regions.

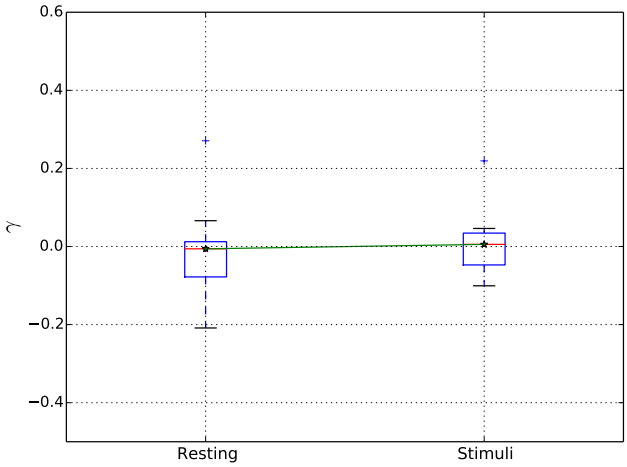
Granger Causality Analysis As in the DI based analysis, we carry parallel GC analyses



(a) V1

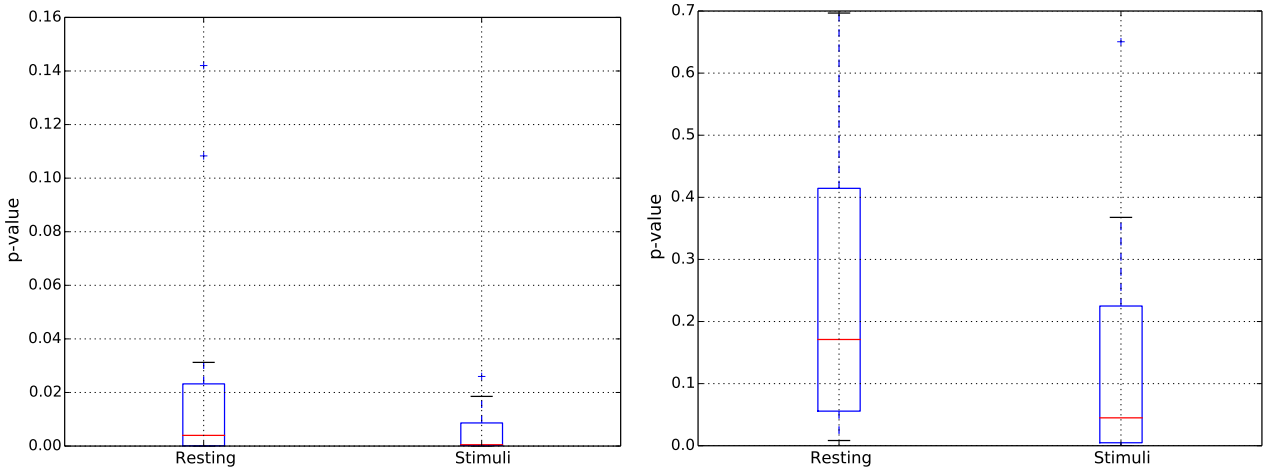


(b) PPA



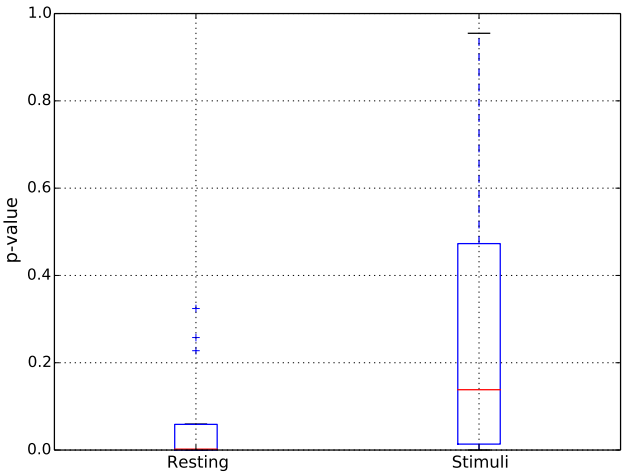
(c) SMC

Figure 3.4: Inner-region (left-right) γ values of the directed information based causality analysis.

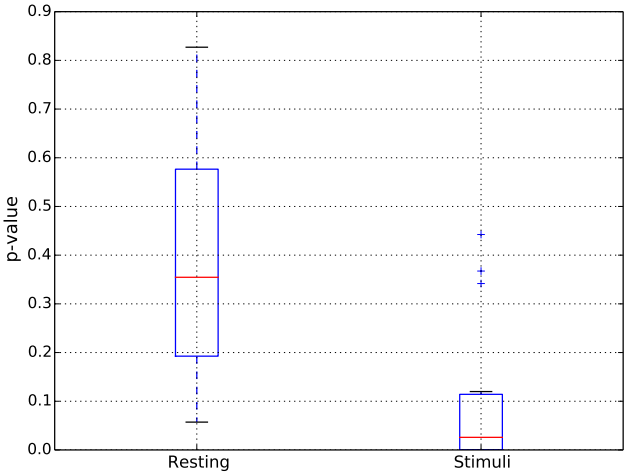


(a) V1 to PPA

(b) V1 to SMC



(c) PPA to V1



(d) SMC to V1

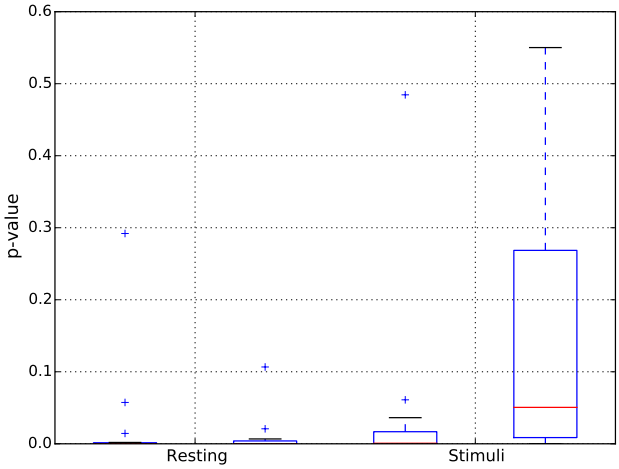
Figure 3.5: Inter-region Granger Causality test result.

between brain regions (Figure 3.5). For the left V1– left PPA pair (Figure 3.5 (a) and (c)), GC analyses indicate that there is a dominating influence of left V1 over left PPA under both resting-state (median $p = 0.004$) and stimulation (median $p = 0.0005$) conditions. However, causal relationship between left V1 and left PPA is not expected in the resting state condition. With the Granger Causality analysis, it is difficult to distinguish between the resting state and stimulus-based state, as the p-values in both states are small enough to indicate a causal relationship. For the left V1–left SMC pair (Figure 3.5 (b) and (d)), the Granger Causality analysis seems to have reversed the expected causal relationship, and indicates that SMC is more likely to be the cause.

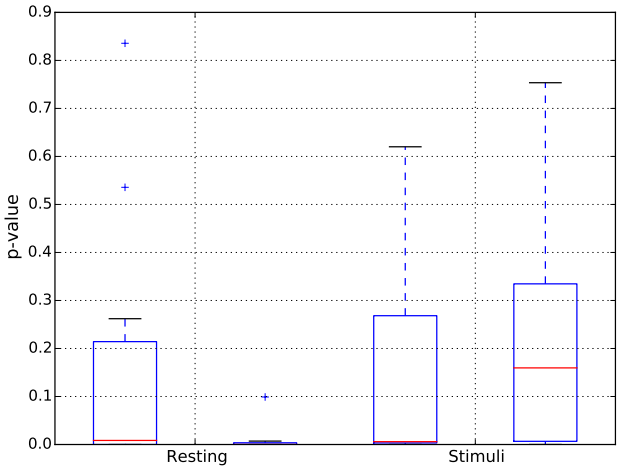
Figure 3.6 shows the results of the GC analysis for the right and left homologous brain regions, including V1, PPA and SMC. The results indicated that the information flow between each pair of homologous regions was very unbalanced, and varied significantly in most cases. This is contradicting to the expected non-sequential activation between them, as it is believed that the right and left homologous regions should not have significant sequential activation.

3.4.3 Impact of Hemodynamics on DI Based Causality Analysis

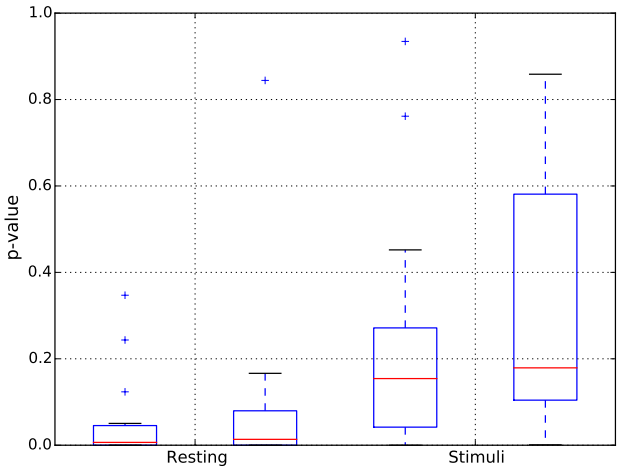
In this subsection, we evaluate the impact of hemodynamics (i.e. the blood flow or the circulation) on the performance of DI based causality analysis. We take a representative model for the hemodynamic response function $h(t) = t^{8.6}e^{-t/0.547}$ [72]. Let $x(t)$ be the sawtooth waveform (to mimic the situation under periodic stimulation), and $y(t) = 0.3x(t)+$



(a) V1



(b) PPA



(c) SMC

Figure 3.6: Inner-region (left-right) Granger Causality test results.

$0.2x(t-1)$. Clearly, $x(t)$ is the causal side. Define:

$$\hat{x}(t) = h_x(t) * x(t), \tag{3.35}$$

$$\hat{y}(t) = h_y(t) * y(t). \tag{3.36}$$

Recall that the Dirac delta function, defined as

$$\delta(t) = 0, \forall t \neq 0,$$

$$\int_{-\infty}^{\infty} \delta(t)dt = 1,$$

is generally used to model the impulse response of an ideal communication channel. We conduct DI based causal analysis for $\hat{x}$ and $\hat{y}$ under the following scenarios:

- Group 1:
  - case 1.1: $h_x(t) = \delta(t), h_y(t) = \delta(t)$;
  - case 1.2: $h_x(t) = h(t), h_y(t) = h(t)$;
  - case 1.3: $h_x(t) = \delta(t), h_y(t) = 3\delta(t)$;
- Group 2:
  - case 2.1: $h_x(t) = \delta(t), h_y(t) = \delta(t)$;
  - case 2.2: $h_x(t) = 0.3h(t) + 0.2h(t-1), h_y(t) = 0.3h(t) + 0.2h(t-1)$;
  - case 2.3: $h_x(t) = h(t), h_y(t) = 0.3h(t) + 0.2h(t-1)$;
  - case 2.4: $h_x(t) = h(t), h_y(t) = 0.7h(t) + 0.4h(t-1)$;
- Group 3:

– case 3.1: $h_x(t) = \delta(t), h_y(t) = \delta(t)$;

– case 3.2: $h_x(t) = h(t), h_y(t) = h(2t)$;

The results are shown in Figure 3.7. We look at Group 1 first. It can be observed that, when the hemodynamic response functions $h_x(t)$ and $h_y(t)$ are identical, the causal relationship between $\hat{x}(t)$ and $\hat{y}(t)$ would be the same with that of $x(t)$ and $y(t)$. That is, $\hat{x}(t)$ is the causal part. However, in case 1.3, when the power level of $\hat{y}(t)$ is much higher than that of $\hat{x}(t)$, the causal relationship is either reversed or becomes ambiguous.

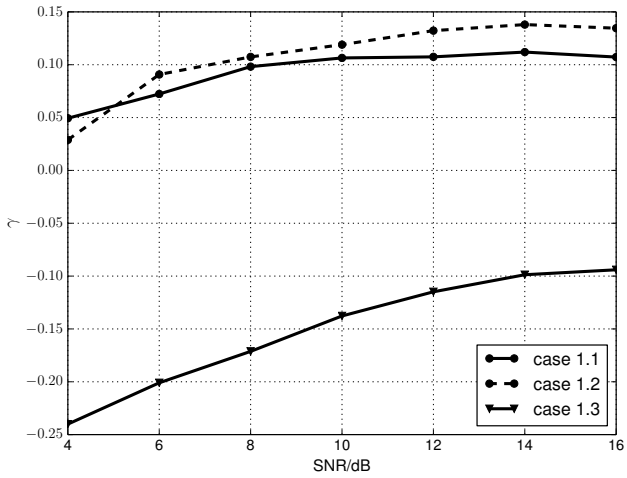
In Group 2, we consider the multipath channel model. For case 2.1 and case 2.2, we can see clearly that $\hat{x}(t)$ is the causal side, but in case 2.3, the power level of $\hat{y}(t)$ is higher than that of $\hat{x}(t)$, again, the causal relationship is reversed or becomes ambiguous.

In Group 3, we consider the case when the hemodynamic response $h_y(t)$ changes much faster than $h_x(t)$. From Figure 3.7(c), it can be seen that this did not bother the DI based method. We can see clearly that $\hat{x}(t)$ is the causal side.

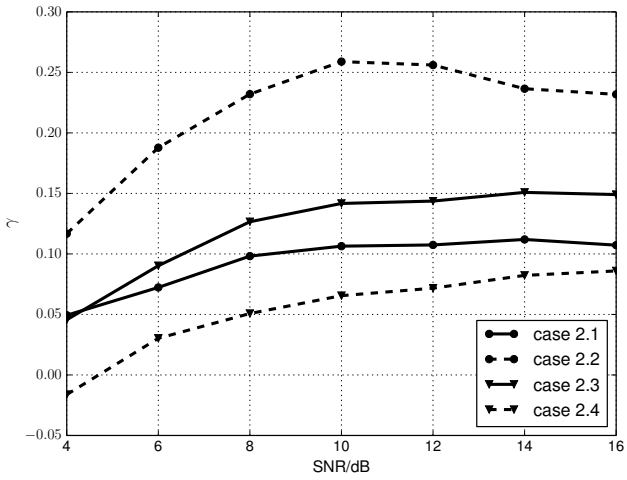
For comparison purpose, we examine the impact of hemodynamics on GC analysis as well. In most cases, GC cannot distinguish which one is the casual side. Due to space limits, only the results for case 2.4 are shown here. Please refer to Figure 8. As can be seen, based on the p-values, both $x(t)$ and $y(t)$ are identified as the causal side by GC. Our analysis indicates that GC is more sensitive to hemodynamic effects.

3.4.4 Summary of Results

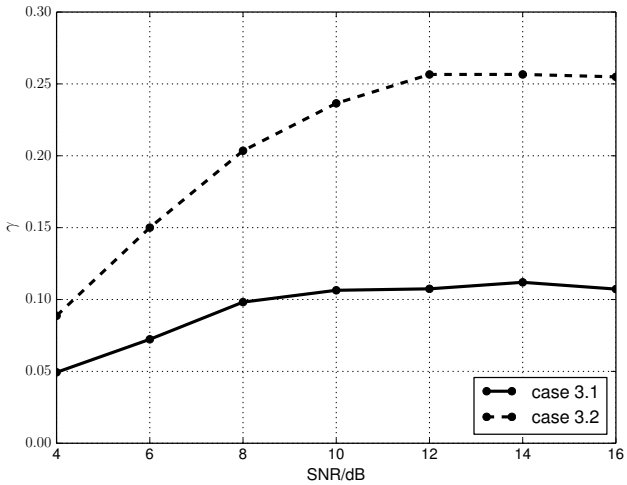
In the simulations, we first performed DI based causality analysis between the fMRI data and the simulated data, and compared the results with that of the GC analysis. Two sets of simulated data were generated from the fMRI data. Set I is obtained by convolving the



(a) Group 1



(b) Group 2



(c) Group 3

Figure 3.7: DI under different hemodyamic response functions.

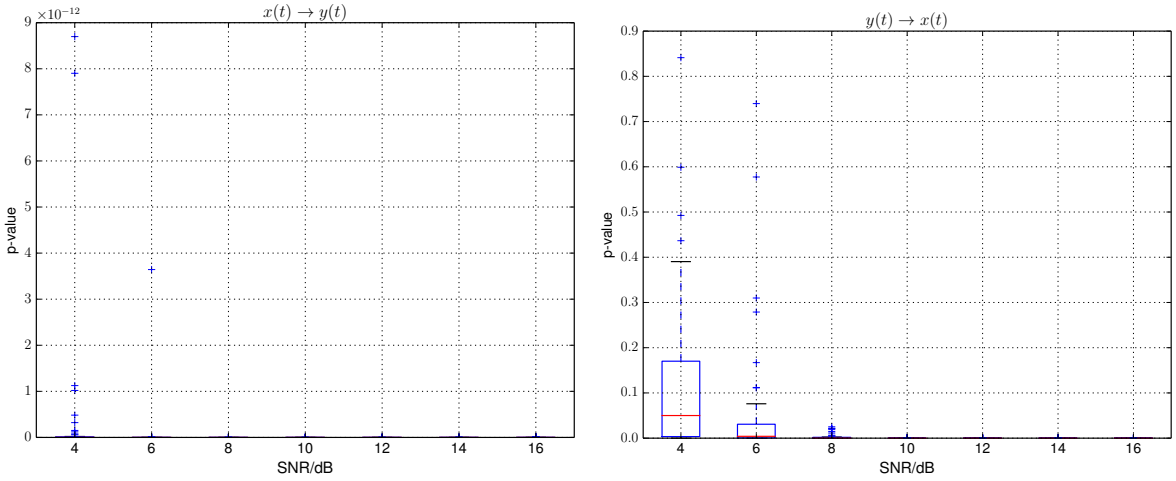


Figure 3.8: Granger Causality test results of case 2.4.

fMRI data with a simple causal model; Set II is obtained by mapping the positive points in the fMRI data to 1, and the negative points to zeros. The simulation results showed that the Granger Causality analysis could identify the cause clearly between the fMRI data and data set I, but failed for the test corresponding to data set II, due to the severe nonlinearity of the data. The directed information based approach, on the other hand, can identify the cause accurately in both cases as long as the SNR is above 5dB.

As can be seen, GC analysis is effective in detecting linear or nearly linear causal relationship, but has difficulty in capturing nonlinear causal relationships. The underlying argument is that GC analysis relies heavily on the linear prediction theory, or linear modeling of the involved parameters. The directed information based causality analysis, on the other hand, does not have any modeling constraints on the sequences to be evaluated, hence can be used to capture both linear and non-linear causal relationships.

We then applied both the DI based analysis and the GC analysis to examine the causal relationship in V1–PPA, V1–SMC, as well as the right and left homologous brain regions, including V1, PPA and SMC. From the DI based analysis, we observed that: (i) In the

resting state, there is no dominant cause for both the V1–PPA and V1–SMC pairs. (ii) In the stimulation based state, V1 turns out to be the cause in the V1-SMC pair. For the V1–PPA pair, although not as strong as in the V1–SMC pair, V1 is more likely to be the cause part. (iii) For both the resting state and the stimulus-based state, there is no dominant cause observed in the right and left homologous brain regions. For the GC analysis, it can be seen that: (i) The results for the V1–PPA are consistent with that of the DI analysis. (ii) For the V1–SMC pair, the results corresponding to resting state are also consistent with the DI analysis; however, in the stimulation based state, it shows that SMC is more likely to be the cause, which is contradicting with the expected sequential brain activation (V1 to SMC). In this paradigm, the activity in SMC is weak, which could be a reason that the GC analysis could not detect the sequential activity. (iii) The information flows between each pair of homologous regions were not balanced in most cases, which is contradicting to the expected non-sequential activation between them.

Finally, we evaluate the impact of hemodynamic effects on DI based causality analysis method using simulated data. we observed that: (i) even if the hemodynamic response function of the driving side changes slower than that of the other side, the proposed DI method can still identify the causal side accurately. (ii) However, when the power level of the driving side is much lower than that of the other side, then the causal relationship may be reversed or become ambiguous. This is because that: in the digitization process, higher power level maps to higher entropy; as in BOLD, higher fMRI amplitude implies more significant activity levels. We will investigate more on this in the future.

Our results indicate that DI is an effective technique to quantify the overall causal relationship. It is also observed that brain connectivity between two different regions generally involves dynamic two-way information transmission between them, rather than a fixed one-

way source to destination relationship.

3.5 Summary

In this chapter, we presented the directed information framework and showed how to apply it for fMRI causality analysis. We provided the detailed procedure on how to calculate the DI for two finite time series. The two major steps involved here are optimal bin size selection for data digitization, and probability estimation. We applied the DI based causality analysis to both the simulated data and experimental fMRI data, and compared the results with that of the Granger Causality analysis. Our results indicated that GC analysis is effective in detecting linear or nearly linear causal relationship, but has difficulty in capturing nonlinear causal relationships. On the other hand, DI based causality analysis is effective in capturing both linear and non-linear causal relationships. Moreover, it was observed that brain connectivity among different regions generally involves dynamic two-way information transmissions between them. Our results showed that when bidirectional information flow is present, DI is more effective than GC to quantify the overall causal relationship.

We would also like to point out that with DI based approach, the performance improves as the data size increases. This is because the probability estimation gets more accurate as we have more samples. For future work, we would continue our research on functional and effective brain connectivity by combining the conventional information theory, the directed information framework as well as the network-level information theory.

Chapter 4

Discrete DCM and Its Relationship with Directed Information and Granger Causality

In this chapter, we explores the discrete Dynamic Causal Modeling (DDCM) and its relationship with Directed Information (DI) and Granger Causality (GC). First, we demonstrate the relationship between DDCM and the continuous DCM. Second, we prove the conditional equivalence between DDCM and DI in characterizing the causal relationship between two brain regions. This equivalence between DDCM and DI also provides an effective method for DI estimation. Moreover, it is shown that when the hemodynamic system is invertible, the DI-based causal relationship between the neurostates of two brain regions is the same with that between the observed BOLD signals. Finally, we illustrate the similarities and differences between DDCM and GC. Although they share a similar mathematic form, the causality measures they utilize are completely different. The theoretical techniques are demonstrated using fMRI data obtained under both resting state and stimulus based state. Our numerical analysis is consistent with that reported in previous study.

4.1 Introduction

Causality analysis aims to find the relationship between causes and effects. It provides insightful information on how brain regions interact with each other during a cognitive task [6]. In general, causality analysis tries to determine whether the values of one time series are useful in predicting the future values of another time series. Since 1990s, a number of frameworks have been applied to fMRI based causality analysis. Among them, *Granger Causality* (GC), *Directed Information* (DI), and *Dynamic Causal Modeling* (DCM) are three representative approaches. In this chapter, we will revisit these three causality analysis frameworks, discuss the relationships between them, especially the relationship between DCM and DI.

Granger Causality The first practical causality analysis framework was proposed by Granger in 1969 [7]. The main idea is, if two signals X_1 and X_2 form a causal relationship, then, instead of using the past values of X_2 alone, the information contained in the past values of X_1 will help predict X_2. More specifically, the calculation of Granger Causality is based on the linear prediction models. Suppose $\mathbf{X}_1^n = [X_1(1), X_1(2), ..., X_1(n)]$ and $\mathbf{X}_2^n = [X_2(1), X_2(2), ..., X_2(n)]$ are two time series observed from two brain regions, respectively. Granger Causality compares the prediction errors e_r and $\tilde{e}_r$ in the following equations:

$$X_2(k+1) = \sum_{l=0}^{L-1} a_l X_2(k-l) + e_r, \tag{4.1}$$

$$X_2(k+1) = \sum_{l=0}^{L-1} [b_l X_1(k-l) + c_l X_2(k-l)] + \tilde{e}_r, \tag{4.2}$$

for $k = 1, 2, ..., n$. Here, e_r is the error of predicting X_2 based only on the previous values

of X_2, and $\tilde{e}_r$ is the error of predicting X_2 based on the previous values of both X_2 and X_1. If $\tilde{e}_r$ is much smaller than e_r, that is, the introduction of the previous values of X_1 can improve the prediction accuracy, then we say there is a Granger causal relationship between X_1 and X_2.

In literature, there have been growing interests in applying GC to identify causal interactions in the brain [8–11]. For example, in [10], Hu et al. applied GC analysis on fMRI data to evaluate the causal relationship among specific brain regions, so as to understand the impact of amnesic mild cognitive impairment (aMCI) to brain connectivity. In [11], David et al. applied GC (together with Dynamic Causal Modeling) to a combination of fMRI and EEG data. Their experiments showed that as the hemodynamics (i.e., the blood flow or the circulation) vary from region to region, GC may not be applied directly on the fMRI signals. However, when the hemodynamic effects were explicitly removed, GC test can perform effective causality analysis in linear relationships.

As a widely accepted technique, the validity and computational simplicity of Granger Causality have been appreciated. However, it has also been noticed that GC relies heavily on the linear prediction method. When there exist instantaneous and/or strong nonlinear interactions between two regions, GC analysis may lead to invalid results [11,27]. To address this problem, several approaches on nonlinear Granger Causality have been proposed. For example, in [56], Bezruchko et al. proposed an auto-regression model constructed in the form of a polynomial. More recently, Marinazzo et al. proposed a method to generalize GC to include the nonlinear case using the kernel technique [57]. The copula approach has been applied for GC assessment in [9, 58]. A comprehensive discussion on nonlinear GC could be found in [59].

Directed Information Directed Information is an information theoretic metric, which was

first introduced by Massey when studying communication channels with feedback [24]. It measures the directed information flow from one time series X to another time series Y, denoted as $I(X \to Y)$. If $I(X \to Y) > I(Y \to X)$, then we say X has more influence on Y, or X is the causal side in the connectivity.

DI is a universal method. Unlike GC, which mainly relies on the linear prediction theory, or linear modeling for the involved parameters, the DI-based causality analysis does not have any modeling constraint on the sequences to be evaluated, hence, can be used to characterize more general relationships. This advantage of DI has been reported in recent advances in causality analysis [22, 25, 26]. In [27], it was pointed out that GC analysis is effective in detecting linear or nearly linear causal relationship, but may have difficulty in capturing nonlinear causal relationships. On the other hand, DI-based causality analysis is more effective in capturing both linear and nonlinear causal relationships. In [28], Liu et al. applied DI to the EEG data and compared the result with that of GC. Their conclusion was that DI based approach could be superior to GC in capturing the instantaneous and nonlinear causal relationship in EEG data. Moreover, in [29], it was shown that the Granger Causality graphs of stochastic processes can be generated from the DI framework, and the authors indicated that the DI theory provides an adequate framework for the connectivity inference problems in neuroscience applications. A comprehensive investigation of DI can be found in [30].

Dynamic Causal Modeling In 2003, Friston proposed the framework of *Dynamic Causal Modeling* to describe the general interactions among a group of brain regions [15]. DCM assumes that the invisible neurostate X, the (external) input U, the observed BOLD signal Y, the parameter θ that characterizes the connectivities between different brain regions, and the independent noise Ω form a dynamic system that could be described by the following

equations:

$$\dot{X} = f(X, U, \theta) \text{ and } Y = \Lambda(X) + \Omega, \tag{4.3}$$

where Λ represents a cascade of differential equations which map the neurostate X to the observed BOLD signal Y. Relying on the EM algorithm, DCM has been implemented on both fMRI and EEG data [16]. In practical applications, due to the computational complexity, DCM is usually used as a confirmatory approach. That is, the users need to put forward different connectivity models and then compare them based on their likelihood evaluated under DCM [17].

Compared with GC, DCM provides a more comprehensive characterization of the dynamic interactions between multiple regions. In [73], Friston et al. pointed out that GC and DCM were complementary to each other: GC models the causal dependency among observed responses, while DCM models the causal interactions among the hidden neurostates. On the other hand, in [74], Friston provided an example to show that DCM and GC may generate different results given the same dataset. The underlining argument is that: DCM takes into account both the external input and the biological variations of the hemodynamic response, which are not involved in GC.

To this end, it can be seen that the relationships between GC and DCM, and between GC and DI have been investigated in literature. While GC is efficient in detecting linear causal relationships, both DI and DCM can be used to characterize more general causal relationships. A missing link here is: what is the relationship between DCM and DI?

In this chapter, we aim to fill this missing link, and explore the connection between DCM and DI. First, we revisit the discrete DCM (DDCM), and demonstrate the relationship

between DDCM and the conventional continuous time DCM. Second, we show that under certain conditions, DDCM and DI are equivalent in characterizing the causal relationship between two brain regions. Recall that traditionally, the accuracy of DI estimation is based on the accuracy of probability or statistic estimation, and hence requires the data length be sufficiently long. This equivalence between DDCM and DI, in fact, also provides a simple but effective method for DI estimation under limited data length. More specifically, the major contributions of this chapter can be summarized as:

1. *We demonstrate the relationship between DDCM and the continuous time DCM, and confirm the validity of DDCM.* In DCM, the neural dynamics within the brain are described using a differential equation. Conventionally, DDCM is obtained from DCM in two steps: first, sampling the continuous DCM and approximating the differential equation with a difference equation; second, modeling the hemodynamic response as an LTI system, and characterizing it with a convolution. In this chapter, rather than using approximation, we prove that when the input to the neural dynamic system is a constant, then DDCM can be strictly derived from DCM under the noise free case. Our result further demonstrates the validity and accuracy of the DDCM model.

2. *We reveal the conditional equivalence between DDCM and DI in characterizing the causal relationship between two brain regions.* More specifically, assuming that the dynamical neural system is causal, the neurostate and the noise at each region are normally distributed, and the external input is a constant, we show that DDCM and DI are equivalent in characterizing the causal relationship between two brain regions. We also show that when the hemodynamic system is invertible, then the DI-based causal relationship between the neurostates of two brain regions is the same with that

between the observed BOLD signals. Finally, this equivalence between DDCM and DI provides a simple method for DI estimation.

3. *We illustrate the similarities and differences between DDCM and GC.* Although GC and DDCM share a similar mathematic form, DDCM determines the causality based on values of the connectivity coefficients, while GC determines the causality by comparing the prediction errors. In addition, both the external input and the biological variations of the hemodynamic response are taken into account in DDCM, but are not involved in GC.

4. *We validate the theoretical results with fMRI data obtained under both resting state and stimulus based state.* Numerical result shows that both DDCM and DI can capture the causal relationships between the primary visual cortex (V1), the parahippocampal place area (PPA), and between V1 and the sensorimotor cortex (SMC). As expected, in the stimulus based state, V1 has shown significant causal influence over PPA and SMC; in the resting state, no clear causal relationship can be observed. Our results are consistent with that reported in previous study [27].

The rest of this chapter is organized as follows. In Section II, we revisit DDCM and demonstrate its relationship with the continuous time DCM. In Section III, we prove the conditional equivalence between DDCM and DI in characterizing the causal relationship between two brain regions. The similarities and differences between DDCM and GC are illustrated in Section IV. We present the numerical results in Section V, and conclude in Section VI.

4.2 Discrete Dynamic Causal Modeling

In continuous time DCM, the invisible neurostates of d brain regions are denoted by a vector $\mathbf{X} = [X_1, X_2, ..., X_d]^t$, where each X_i, $i = 1, 2, ..., d$, represents the neurostate of the ith region. The basic idea of DCM is that, the neurostate $\mathbf{X}$, the external input U, the connectivity matrices $\tilde{A}$ and $\tilde{B}$ that describe the connections among brain regions, the observed BOLD signal $\mathbf{Y}$ and the independent noise can be formulated as a complex dynamic system, characterized as:

$$\dot{\mathbf{X}}(t) = \tilde{A}\mathbf{X}(t) + \tilde{B}U(t) + \mathbf{\Omega}_1(t), \tag{4.4}$$

$$\mathbf{Y}(t) = \tilde{\Lambda}(\mathbf{X}(t)) + \mathbf{\Omega}_2(t), \tag{4.5}$$

where $\mathbf{\Omega}_1(t)$ and $\mathbf{\Omega}_2(t)$ are the state noise and observation noise, and $\tilde{\Lambda}$ represents the mapping from the neurostate $\mathbf{X}(t)$ to the observed BOLD signal $\mathbf{Y}(t)$.

The functional connectivity in DCM is mainly characterized by matrix $\tilde{A}$ [8]. For instance, in a model with two brain regions, that is, $d = 2$ and $\mathbf{X} = [X_1, X_2]^t$, the connectivity matrix $\tilde{A}$ will be:

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}. \tag{4.6}$$

Here, for $i, j = 1, 2$, $\tilde{A}_{ii}$ measures the influence of the past values of X_i on its future values, and $\tilde{A}_{ji}$ measures the influence of past values of X_i on the future values of X_j. The absolute values of $\tilde{A}_{12}$ and $\tilde{A}_{21}$ describe the causal relationship between the two regions: when $|\tilde{A}_{12}| > |\tilde{A}_{21}|$, it means that X_2 has imposed more influence over X_1; and when $|\tilde{A}_{21}| > |\tilde{A}_{12}|$, it means that X_1 has imposed more influence over X_2 [17, 75].

It can be seen from equations (4.4) and (4.5) that the continuous time DCM characterizes the dynamic neural system using two continuous-time equations. However, parameter estimation in continuous time equations faces considerable challenges in practical applications. To overcome this difficulty, there have been efforts to simplify DCM to a more tractable form, such as the switching linear dynamic model (SLDS) [75] and multivariate dynamical model (MDS) [17]. In both approaches, the continuous time equations are discretized, and the mapping between the neurostate and the BOLD signal is approximated as an LTI system, characterized using a convolution. That is, the discrete DCM (DDCM) model can be obtained as:

$$\begin{aligned} \mathbf{X}(k+1) &= A\mathbf{X}(k) + BU(k) + \mathbf{\Omega}_1(k), & (4.7) \\ \mathbf{Y}(k) &= \sum_{m=0}^{M} \Lambda(m)\mathbf{X}(k-m) + \mathbf{\Omega}_2(k), & (4.8) \end{aligned}$$

where A is the connectivity matrix, $\{\Lambda(m), m = 0, 1, \cdots M\}$ denotes the convolution coefficients corresponding to the hemodynamic response, and $\mathbf{\Omega}_1(k)$ and $\mathbf{\Omega}_2(k)$ denote the noise terms independent of the brain state and the input.

Consider the case of two regions, region 1 and region 2, where equation (4.7) can be rewritten as:

$$\begin{bmatrix} X_1(k+1) \\ X_2(k+1) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} X_1(k) \\ X_2(k) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} U(k) + \begin{bmatrix} \Omega_{11}(k) \\ \Omega_{12}(k) \end{bmatrix}. \quad (4.9)$$

Similar to the continuous time DCM, coefficients A_{12} and A_{21} actually measure the causal relationship between region 1 and region 2. More specifically, if $|A_{21}| > |A_{12}|$, then X_1 is more likely to be the casual side, and vice versa. The same analysis holds when multiple

brain regions are under investigations [17, 75].

As can be seen, equation (4.7) is obtained by approximating the continuous time differential equation in DCM with a difference equation. Here, we will show that, in the noise-free case, when the external input is a constant, DDCM can be directly derived from DCM. For this purpose, we need to prove the following Lemma first.

Lemma 1 For a differential equation $\dot{\mathbf{X}}(t) = \tilde{A}\mathbf{X}(t) + \tilde{B}U(t)$, where $\tilde{A}$ and $\tilde{B}$ are deterministic matrices, the solution to the equation is:

$$\mathbf{X}(t) = e^{\tilde{A}t}\mathbf{X}(0) + \int_0^t e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau. \tag{4.10}$$

Proof Taking the derivatives on both side of equation (4.10), the left hand side of the equation would be $\dot{\mathbf{X}}(t)$. Given that $U(t)$ is causal, the right hand side would be:

$$\begin{aligned}
& \frac{d}{dt}[e^{\tilde{A}t}\mathbf{X}(0) + \int_0^\infty u(t-\tau)e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau] \\
= \; & \tilde{A}e^{\tilde{A}t}\mathbf{X}(0) + \int_0^\infty \frac{d}{dt}u(t-\tau)e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau + \int_0^\infty u(t-\tau)\frac{d}{dt}e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau \\
= \; & \tilde{A}e^{\tilde{A}t}\mathbf{X}(0) + \int_0^\infty \delta(t-\tau)e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau + \int_0^\infty u(t-\tau)\tilde{A}e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau \\
= \; & \tilde{A}e^{\tilde{A}t}\mathbf{X}(0) + e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)|_{\tau=t} + \tilde{A}\int_0^t e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau \\
= \; & \tilde{A}[e^{\tilde{A}t}\mathbf{X}(0) + \int_0^t e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau] + \tilde{B}U(t) \\
= \; & \tilde{A}\mathbf{X}(t) + \tilde{B}U(t),
\end{aligned} \tag{4.11}$$

in which $u(t)$ is the unit step function, and $\delta(t)$ the Dirac delta function. That means, $\mathbf{X}(t) = e^{\tilde{A}t}\mathbf{X}(0) + \int_0^t e^{\tilde{A}(t-\tau)}\tilde{B}U(\tau)d\tau$ is the solution to equation $\dot{\mathbf{X}}(t) = \tilde{A}\mathbf{X}(t) + \tilde{B}U(t)$. ■

To illustrate the relationship between DCM and DDCM, assume there is no background noise, and the external input is a constant, i.e., $U(t) = U$. In this case, sampling the

neurostate in equation (4.10) at $t = kT$, where $k = 0, 1, 2...$, and T the sampling period, we get:

$$\mathbf{X}(kT) = e^{\tilde{A}kT}\mathbf{X}(0) + \int_0^{kT} e^{\tilde{A}(kT-\tau)}\tilde{B}Ud\tau. \tag{4.12}$$

Accordingly, the neurostate at time $t = (k+1)T$ is

$$\mathbf{X}((k+1)T) = e^{\tilde{A}(k+1)T}\mathbf{X}(0) + \int_0^{(k+1)T} e^{\tilde{A}(\ (k+1)T-\tau\ )}\tilde{B}Ud\tau. \tag{4.13}$$

Subtracting $e^{\tilde{A}T}\mathbf{X}(kT)$ from $\mathbf{X}((k+1)T)$, we have:

$$\begin{aligned}
\mathbf{X}((k+1)T) - e^{\tilde{A}T}\mathbf{X}(kT) &= \int_0^{(k+1)T} e^{\tilde{A}(\ (k+1)T-\tau\ )}\tilde{B}Ud\tau - \int_0^{kT} e^{\tilde{A}(\ (k+1)T-\tau\ )}\tilde{B}Ud\tau \\
&= \int_{kT}^{(k+1)T} e^{\tilde{A}(\ (k+1)T-\tau)}\tilde{B}Ud\tau \\
&= \tilde{A}^{-1}(e^{\tilde{A}T} - I)\tilde{B}U.
\end{aligned} \tag{4.14}$$

Let $A = e^{\tilde{A}T}$ and $B = \tilde{A}^{-1}(e^{\tilde{A}T} - I)\tilde{B}$, we can rewrite equation (4.14) as:

$$\mathbf{X}(\ (k+1)T\ ) = A\mathbf{X}(kT) + BU. \tag{4.15}$$

When the sampling period T is fixed, we can simplify equation (4.15) as:

$$\mathbf{X}(k+1) = A\mathbf{X}(k) + BU. \tag{4.16}$$

This confirms the validity of the DDCM model, and the relationship between DCM and DDCM.

In the following sections, we will investigate the relationship of DDCM with the DI-based causality analysis framework and the Granger Causality analysis.

4.3 The Relationship between DDCM and Directed Information

In this section, we show that, under certain assumptions, DI and DDCM are equivalent in characterizing the causal relationship between different brain regions.

4.3.1 Directed Information based Causality Analysis

Directed Information is a causality analysis framework based on information theory. It was first introduced by Massey to study the capacity of a communication channel with feedbacks [24]. The amplitude of DI has a clear physical meaning: it reflects the information flow from one time series $\mathbf{X}_1^n$ to another, $\mathbf{X}_2^n$. In [29], it was pointed out that the directed information theory provides an effective and adequate framework for the connectivity inference problems in neuroscience applications, as the Granger Causality graphs could be derived using DI.

The directed information from one time series $\mathbf{X}_1^n$ to another $\mathbf{X}_2^n$ is calculated as [24]:

$$I(\mathbf{X}_1^n \to \mathbf{X}_2^n) = \sum_{k=1}^{n} [h(X_2(k)|\mathbf{X}_2^{k-1}) - h(X_2(k)|\mathbf{X}_2^{k-1}, \mathbf{X}_1^k)], \tag{4.17}$$

where $\mathbf{X}_i^k = [X_i(1), X_i(2), ..., X_i(k)]$, $i = 1, 2$, and h denotes the differential entropy operator. If $I(\mathbf{X}_1^n \to \mathbf{X}_2^n)$ is greater than $I(\mathbf{X}_2^n \to \mathbf{X}_1^n)$, we say $\mathbf{X}_1^n$ has more causal influence over $\mathbf{X}_2^n$; otherwise $\mathbf{X}_2^n$ has more causal influence over $\mathbf{X}_1^n$.

Practical evaluation of directed information measures has been a challenging problem. Over the last two decades, a limited number of directed information estimators have been purposed [3, 22, 30]. In the following subsections, we will derive the theoretical form of DI under certain assumptions, and investigate its relationship with DDCM.

4.3.2 DDCM and Directed Information

When deriving the relationship between DDCM and DI, we impose the following assumptions to make the problem more tractable: (i) The dynamic neural system under investigation is a causal system, which means for each brain region, the current value of the neurostate depends only on previous values of neurostates of the region and its related regions. (ii) For each region, both the neurostate and the background noise are normally distributed, and the variances are the same in related brain regions. More specifically, for each $k = 1, 2, \cdots, n$, the variances corresponding to the neurostate and the background noise are σ_x^2 and σ_0^2, respectively. (iii) The external input U is a constant. This assumption is reasonable when the changing rate of the external input is much slower than that of neurostates.

In the following analysis, let the uppercase letters (X,Y,...) denote random variables, and the lowercase letters (x,y,...) the possible values they can acquire. In particular, $x_1(k)$ and $x_2(k)$ denote the possible values $X_1(k)$ and $X_2(k)$ can acquire, and $\omega_{11}(k)$ and $\omega_{12}(k)$ denote the possible values $\Omega_{11}(k)$ and $\Omega_{12}(k)$ can acquire. Given a time series $\mathbf{X}^n = [X(1), X(2), ..., X(n)], n \in N$, for any $x(k), k = 1, 2, \ldots n$, $P(x(k))$ denotes the probability for $X(k)$ to take the value $x(k)$, and $P(x(k)|\mathbf{x}^{k-1})$ the conditional probability that the current sample $X(k)$ is $x(k)$, given that the previously observed sequence is $\mathbf{x}^{k-1} = [x(1), x(2), ..., x(k-1)]$.

Following equation (4.9), the conditional probability $P(x_2(k) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k)$ can be written

as:

$$\begin{aligned} P(x_2(k) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k) &= P(A_{21}x_1(k-1) + A_{22}x_2(k-1) + B_2U + \omega_{12}(k-1) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k) \\ &= P(A_{21}x_1(k-1) + A_{22}x_2(k-1) + B_2U + \omega_{12}(k-1) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^{k-1}) \\ &= P(\omega_{12}(k-1)). \end{aligned} \tag{4.18}$$

This implies that, for each $k = 1, 2, \cdots, n$, the conditional probability density function of the neurostate $X_2(k)$ given $\mathbf{X}_2^{k-1}$ and $\mathbf{X}_1^k$ is Gaussian with variance σ_0^2.

It is well known that given a Gaussian random variable Ξ with variance σ_ξ^2, the corresponding differential entropy $h(\Xi)$ can be calculated as:

$$h(\Xi) = \frac{1}{2}\log 2\pi e\sigma_\xi^2. \tag{4.19}$$

Therefore, based on equations (4.18) and (4.19), the differential entropy corresponding to the neurostate $X_2(k)$ given $\mathbf{X}_2^{k-1}$ and $\mathbf{X}_1^k$ can then be calculated as:

$$h(X_2(k)|\mathbf{X}_2^{k-1}, \mathbf{X}_1^k) = \frac{1}{2}\log 2\pi e\sigma_0^2. \tag{4.20}$$

Similarly, the conditional probability $P(x_2(k) \mid \mathbf{x}_2^{k-1})$ can be simplified as:

$$\begin{aligned} P(x_2(k) \mid \mathbf{x}_2^{k-1}) &= P(A_{21}x_1(k-1) + A_{22}x_2(k-1) + B_2U + \omega_{12}(k-1) \mid \mathbf{x}_2^{k-1}) \\ &= P(A_{21}x_1(k-1) + \omega_{12}(k-1)). \end{aligned} \tag{4.21}$$

As a result, the corresponding differential entropy will be:

$$h(X_2(k)|\mathbf{X}_2^{k-1}) = \frac{1}{2}\log 2\pi e(A_{21}^2\sigma_x^2 + \sigma_0^2), \tag{4.22}$$

where σ_x^2 is the variance of the neurostate, which is assumed to have no significant changes among related regions and within the observation frame.

Based on equations (4.18) to (4.22), the directed information can then be obtained as:

$$\begin{aligned} I(\mathbf{X}_1^n \to \mathbf{X}_2^n) &= \sum_{k=1}^{n}[\frac{1}{2}\log 2\pi e(A_{21}^2\sigma_x^2 + \sigma_0^2) - \frac{1}{2}\log 2\pi e\sigma_0^2] \\ &= \frac{n}{2}\log(1 + A_{21}^2\frac{\sigma_x^2}{\sigma_0^2}) \\ &= \frac{n}{2}\log(1 + cA_{21}^2); \end{aligned} \tag{4.23}$$

where $c = \sigma_x^2/\sigma_0^2$ is the ratio of the power of neural activities and the noise power. Similarly, we can prove that $I(\mathbf{X}_2^n \to \mathbf{X}_1^n) = (n/2)log(1 + cA_{12}^2)$.

It can be seen from equation (4.23) that after the parameters in DDCM have been obtained, the directed information from one region to another can be calculated accordingly. That is, equation (4.23) provides an effective method for the estimation of DI.

Note that when $c > 0$, $log(1 + cx^2)$ is a monotonically increasing function. Based on the discussions above, we can obtain the following proposition:

Proposition 4.1 *If $|A_{21}| > |A_{12}|$, then $I(\boldsymbol{X}_1^n \to \boldsymbol{X}_2^n) > I(\boldsymbol{X}_2^n \to \boldsymbol{X}_1^n)$, that is, region 1 is more likely to be the causal side; otherwise, we will have $I(\boldsymbol{X}_2^n \to \boldsymbol{X}_1^n) > I(\boldsymbol{X}_1^n \to \boldsymbol{X}_2^n)$, and region 2 is more likely to be the causal side.*

Proposition 4.1 is in accordance with the previous analysis for DDCM: the causality analysis can be carried out based on the absolute values of the coefficients of the connectivity

matrix. This means, for a causal dynamic neural system with a constant external input, when the neurostate and the background noise are normally distributed, DI and DDCM are equivalent in characterizing the causal relationship between two brain regions.

On the other hand, in practice, we can only observe the BOLD signal $\mathbf{Y}^n$ rather than the neurostate $\mathbf{X}^n$. That is, given two brain regions, region 1 and region 2, the calculation of DI can only be carried out on the observations $\mathbf{Y}_1^n$ and $\mathbf{Y}_2^n$ rather than the neurostates $\mathbf{X}_1^n$ and $\mathbf{X}_2^n$. The natural questions are: will the causal relationship obtained from the neurostates of two brain regions be maintained in the observed BOLD signals? Whether DDCM is still equivalent with DI concerning the observed BOLD signals in charactering the causal relationship between two brain regions?

In the following, we will prove that: in the observation noise-free case, as long as the hemodynamic system is invertible, DI calculated using the estimated neurostates is equal to the DI calculated using the observed signals.

Following equation (4.8), for each $k = 1, 2, \cdots n$, the values of the observed BOLD signals of region $l, l = 1, 2$, can be written as

$$y_l(k) = \sum_{m=0}^{M} \Lambda_l(m) x_l(k-m) + \Omega_{2l}(k). \tag{4.24}$$

When there is no observation noise, equation (4.24) can be rewritten in the matrix form:

$$\begin{bmatrix} y_l(1) \\ y_l(2) \\ y_l(3) \\ \vdots \\ y_l(k) \end{bmatrix} = \begin{bmatrix} \Lambda_l(0) & 0 & 0 & \cdots & 0 \\ \Lambda_l(1) & \Lambda_l(0) & 0 & \cdots & 0 \\ \Lambda_l(2) & \Lambda_l(1) & \Lambda_l(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Lambda_l(0) \end{bmatrix} \begin{bmatrix} x_l(1) \\ x_l(2) \\ x_l(3) \\ \vdots \\ x_l(k) \end{bmatrix}. \tag{4.25}$$

That is, we have

$$\mathbf{y}_l^k = \mathbf{\Lambda}_{l,k}\mathbf{x}_l^k, \quad k = 1, 2, \cdots n, \ l = 1, 2, \tag{4.26}$$

where $\mathbf{y}_l^k = [y_l(1), y_l(2), \cdots, y_l(k)]^T$, $\mathbf{x}_l^k = [x_l(1), x_l(2), \cdots, x_l(k)]^T$, and $\mathbf{\Lambda}_{l,k}$ is a $k \times k$ matrix shown in equation (4.25). It can be seen from equation (4.26) that the determinant of matrix $\mathbf{\Lambda}_{l,k}$ equals to $|\Lambda_l(0)|^k$. Therefore, if $\Lambda_l(0) \neq 0$, then $\mathbf{\Lambda}_{l,k}$ is invertible. Hence for $k = 1, 2, \cdots n,$ and $l = 1, 2$, the values of the neurostate $\mathbf{x}_l^k$ can be recovered from the observed BOLD signal $\mathbf{y}_l^k$ through $\mathbf{x}_l^k = \mathbf{\Lambda}_{l,k}^{-1}\mathbf{y}_l^k$.

As a result, the conditional probability $P(y_2(k) \mid \mathbf{y}_2^{k-1}, \mathbf{y}_1^k)$ can be written as:

$$\begin{aligned} P(\ y_2(k) \mid \mathbf{y}_2^{k-1}, \mathbf{y}_1^k\ ) &= P(\ y_2(k) \mid \mathbf{\Lambda}_{2,k-1}\mathbf{x}_2^{k-1}, \mathbf{\Lambda}_{1,k}\mathbf{x}_1^k\ ) \\ &= P(\ y_2(k) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k\ ) \\ &= P(\ \sum_{m=0}^{M} \Lambda_2(m)x_2(k-m) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k\ ) \\ &= P(\ \Lambda_2(0)x_2(k) \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k\ ) \\ &= P(\ \Lambda_2(0)[A_{21}x_1(k-1) + A_{22}x_2(k-1) + B_2U + \omega_{12}(k-1)] \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^k\ ) \\ &= P(\ \Lambda_2(0)[A_{21}x_1(k-1) + A_{22}x_2(k-1) + B_2U + \omega_{12}(k-1)] \mid \mathbf{x}_2^{k-1}, \mathbf{x}_1^{k-1}\ ) \\ &= P(\ \Lambda_2(0)\omega_{12}(k-1)\ ) \end{aligned} \tag{4.27}$$

This implies that, for each $k = 1, 2, \cdots, n$, the conditional probability density function of $Y_2(k)$ given $\mathbf{Y}_2^{k-1}$ and $\mathbf{Y}_1^k$ is Gaussian with variance $\Lambda_2^2(0)\sigma_0^2$. Therefore, Following equation (4.19), the differential entropy corresponding to $Y_2(k)$ given $\mathbf{Y}_2^{k-1}$ and $\mathbf{Y}_1^k$ can then be calculated as:

$$h(Y_2(k)|\mathbf{Y}_2^{k-1}, \mathbf{Y}_1^k) = \frac{1}{2}\log 2\pi e\Lambda_2^2(0)\sigma_0^2. \tag{4.28}$$

Similarly, the conditional probability $P(y_2(k)|\mathbf{y}_2^{k-1})$ can be simplified as:

$$\begin{aligned} P(y_2(k) \mid \mathbf{y}_2^{k-1}) &= P(\Lambda_2(0)[A_{21}x_1(k-1) + A_{22}x_2(k-1) + B_2U + \omega_{12}(k-1)] \mid \mathbf{\Lambda}_{2,k-1}\mathbf{x}_2^{k-1}) \\ &= P(\Lambda_2(0)[A_{21}x_1(k-1) + \omega_{12}(k-1)]). \end{aligned} \tag{4.29}$$

As a result, the corresponding differential entropy will be

$$h(Y_2(k)|\mathbf{Y}_2^{k-1}) = \frac{1}{2}\log 2\pi e\Lambda_2^2(0)(A_{21}^2\sigma_x^2 + \sigma_0^2), \tag{4.30}$$

where σ_x^2 is the variance of neurostate, which is assumed to have no significant differences among related regions.

Based on equations (4.28) and ((4.30), we have:

$$\begin{aligned} I(\mathbf{Y}_1^n \to \mathbf{Y}_2^n) &= \sum_{k=1}^{n}[\frac{1}{2}\log 2\pi e\Lambda_2^2(0)(A_{21}^2\sigma_x^2 + \sigma_0^2) - \frac{1}{2}\log 2\pi e\Lambda_2^2(0)\sigma_0^2] \\ &= \frac{n}{2}\log(1 + A_{21}^2\frac{\sigma_x^2}{\sigma_0^2}) \\ &= \frac{n}{2}\log(1 + cA_{21}^2); \end{aligned} \tag{4.31}$$

where $c = \sigma_x^2/\sigma_0^2$ is the ratio of the power of neural activities and the noise power. Similarly, we can prove that $I(\mathbf{Y}_2^n \to \mathbf{Y}_1^n) = (n/2)\log(1 + cA_{12}^2)$.

Comparing equation (4.31) with equation (4.23), we can see that in the noise-free case, as long as the hemodynamic system is invertible, DI obtained from the observed BOLD signal equals to the DI obtained from the neurostate, that is, $I(\mathbf{Y}_1^n \to \mathbf{Y}_2^n) = I(\mathbf{X}_1^n \to \mathbf{X}_2^n)$ and $I(\mathbf{Y}_2^n \to \mathbf{Y}_1^n) = I(\mathbf{X}_2^n \to \mathbf{X}_1^n)$.

Proposition 4.2 *If* $|A_{21}| > |A_{12}|$, *then* $I(\boldsymbol{Y}_1^n \to \boldsymbol{Y}_2^n) > I(\boldsymbol{Y}_2^n \to \boldsymbol{Y}_1^n)$, *that is, region* 1 *is more likely to be the causal side; otherwise, we will have* $I(\boldsymbol{Y}_2^n \to \boldsymbol{Y}_1^n) > I(\boldsymbol{Y}_1^n \to \boldsymbol{Y}_2^n)$, *and region* 2 *is more likely to be the causal side.*

Proposition 4.2 implies that in the noise-free case, as long as the hemodynamic system is invertible, the DI-based causal relationship between the neurostates of two brain regions is the same with that between the observed BOLD signals, and DCM and DI are still equivalent in characterizing the causal relationship between brain regions.

4.3.3 Discussions

In the previous subsection, we prove that under certain assumptions, DI and DDCM are equivalent to each other in characterizing the causal relationship between two brain regions. However, such an equivalence may not be generalized to the multi-region DDCM. The reason is that: when more than two regions are involved in DDCM, the causal relationship between two regions should take their interactions with other regions into account.

When multiple brain regions are under investigation, equation (4.7) would be

$$\begin{bmatrix} X_1(k+1) \\ \vdots \\ X_N(k+1) \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{N1} & \cdots & A_{NN} \end{bmatrix} \begin{bmatrix} X_1(k) \\ \vdots \\ X_N(k) \end{bmatrix} + \begin{bmatrix} B_1 \\ \vdots \\ B_N \end{bmatrix} U + \begin{bmatrix} \Omega_{11}(k) \\ \vdots \\ \Omega_{1N}(k) \end{bmatrix}, \quad (4.32)$$

where N is the number of regions. For any two different regions, region i and region j, $i, j = 1, 2, \cdots N$, under the same assumptions in Section III.B, the conditional probability $P(x_j(k)|\mathbf{x}_j^{k-1}, \mathbf{x}_i^k)$ can be written as:

$$\begin{aligned} P(x_j(k) \mid \mathbf{x}_j^{k-1}, \mathbf{x}_i^k) &= P(\sum_{l=1}^{N} A_{jl}x_l(k-1) + B_jU + \omega_{1j}(k-1) \mid \mathbf{x}_j^{k-1}, \mathbf{x}_i^k) \\ &= P(\sum_{l\neq i,j} A_{jl}x_l(k-1) + \omega_{1j}(k-1)). \end{aligned} \quad (4.33)$$

Similarly, the conditional probability $P(x_j(k) \mid \mathbf{x}_j^{k-1})$ can be simplified as:

$$\begin{aligned} P(x_j(k) \mid \mathbf{x}_j^{k-1}) &= P(\sum_{l=1}^{N} A_{jl}x_l(k-1) + B_jU + \omega_{1j}(k-1) \mid \mathbf{x}_j^{k-1}) \\ &= P(\sum_{l\neq j} A_{jl}x_l(k-1) + \omega_{1j}(k-1)). \end{aligned} \quad (4.34)$$

It can be seen from equations (4.33) to (4.34) that for each $k = 1, 2, \cdots n$, the items within the sum $\sum_{l\neq i,j} A_{jl}x_l(k-1)$ may be correlated with each other. As a result, DI cannot be derived with the same approach shown in equations (4.18) to (4.22). Therefore, unlike the two-region case, when multiple regions are involved, whether the causal relationships characterized by DDCM and DI are equivalent is still an open problem.

4.4 The Relationship between DDCM and Granger Causality

Granger Causality (GC) is the first practical causality analysis framework for time series. It can be applied directly to any given time series to detect the coupling among brain regions. The main idea is, if two signals X_1 and X_2 form a causal relationship, then, instead of using the past values of X_2 alone, the information contained in the past values of X_1 will help to predict X_2. More specifically, the calculation of Granger Causality is based on the linear prediction models. Suppose $\mathbf{X}_1^n = [X_1(1), X_1(2), ..., X_1(n)]$ and $\mathbf{X}_2^n = [X_2(1), X_2(2), ..., X_2(n)]$ are two time series observed from two brain regions. Granger Causality compares the prediction errors e_r and $\tilde{e}_r$ in the following equations:

$$X_2(k+1) = \sum_{l=0}^{L-1} a_l X_2(k-l) + e_r, \tag{4.35}$$

$$X_2(k+1) = \sum_{l=0}^{L-1} [b_l X_1(k-l) + c_l X_2(k-l)] + \tilde{e}_r, \tag{4.36}$$

for $k = 1, 2, ..., n$. Here, e_r is the error of predicting X_2 based only on the previous values of X_2, and $\tilde{e}_r$ is the error of predicting X_2 based on both the previous values of X_2 and the previous values of X_1. If $\tilde{e}_r$ is much smaller than e_r, that is, the introduction of the previous values of X_1 can improve the prediction accuracy, then we say there is a Granger causal relationship between X_1 and X_2. In practical analysis, GC can be tested using a nested model comparison based on the F statistics [49].

Since 1990s, there have been growing interests in applying Granger Causality analysis to identify causal interactions in neuroscience [8]. GC has been successfully applied to fMRI

data, EEG measurements, as well as neural level signals [9]. In these pioneering work, the validity and computational simplicity of Granger Causality have been widely recognized.

To show the differences between GC and DDCM, rewrite equations (4.35) and (4.36) as follows:

$$\begin{bmatrix} X_2(k+1) \\ X_2(k+1) \end{bmatrix} = \sum_{l=0}^{L-1} \begin{bmatrix} a_l & 0 \\ b_l & c_l \end{bmatrix} \begin{bmatrix} X_1(k-l) \\ X_2(k-l) \end{bmatrix} + \begin{bmatrix} e_r \\ \tilde{e}_r \end{bmatrix}. \tag{4.37}$$

Comparing equation (4.37) with equations (4.7) and (4.9), we can see that, as in the state equation in DDCM, GC applies an auto-regression model to the observed sequences. However, there are several major differences between GC and DDCM:

First, GC uses the *regression error* terms rather than the *coefficient matrices* to determine the causal relationship between the two sequences. This means, the causality measures used in GC and DDCM are completely different. *Second*, in [74], Friston provided an example to show that DCM and GC could generate different results given the same dataset. The underlining argument is that: both the external input U and the biological variance of the hemodynamic response are taken into account in DDCM but are not involved in GC. Finally, it has also been noticed in [11] that because GC relies on the linear prediction method, when there exist instantaneous and/or strong nonlinear interactions between two regions, GC analysis may lead to invalid results.

4.5 Numerical Analysis

In this section, we briefly describe how to validate the equivalence of DDCM and DI between two regions using experimental fMRI data obtained under both resting state and stimulus

based state.

4.5.1 Data Acquisition

Fourteen right-handed healthy college students (7 males, 23.4 ± 4.2 years of age) from Michigan State University volunteered to participate in this study and signed consent forms approved by the Michigan State University Institutional Review Board. The experiment was conducted on a 3T GE Signa HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil.

For each subject, fMRI datasets were collected on a visual stimulation condition with a scene-object fMRI paradigm and then on a resting-state condition. The parameters for the fMRI scan were: gradient-echo EPI, 36 contiguous 3-mm axial slices in an interleaved order, time of echo (TE) = 27.7 ms, time of repetition (TR) = 2500 ms, flip angle = 80°, field of view (FOV) = 22 cm, matrix size = 64×64, ramp sampling, and with the first four data points discarded.

On the visual stimulation fMRI condition, each volume of images were acquired 192 times (8 min) while each subject was presented with 12 blocks of visual stimulation after an initial 10 s "resting" period. In a predefined randomized order, the scenery pictures were presented in 6 blocks and the object pictures were presented in other 6 blocks. All pictures were unique. In each block, 10 pictures were presented continuously for 25 s (2.5 s for each picture), followed with a 15 s baseline condition (a white screen with a black fixation cross at the center). The subject needed to press his/her right index finger once when the screen was switched from the baseline to picture condition. Stimuli were displayed in color in full screen on a 1024×768 32-inch LCD monitor (Salvagione Design, Sausalito, CA) placed at the back of the magnet room. The LCD subtended $10.2^\circ \times 13.1^\circ$ of visual angle. On the

resting-state fMRI (rs-fMRI) condition, each volume of images were acquired 164 times (6 *min* and 50 *s*) after a subject was informed to relax, keep his/her eyes closed and stay awake throughout the scan. After the above functional data acquisition, high-resolution volumetric T1-weighted spoiled gradient-recalled (SPGR) images with cerebrospinal fluid suppressed were obtained to cover the whole brain with 120 1.5-*mm* sagittal slices, 8° flip angle and 24 *cm* FOV. These images were used to identify anatomical locations.

4.5.2 fMRI Data Pre-processing and Analysis

All stimulus fMRI data pre-processing and analysis for each subject were conducted with AFNI software (Cox, 1996) as described in Henderson et al. [61]. Essentially, slice-timing correction and rigid-body motion correction were conducted. Spatial blurring with a full width half maximum of 4 mm was applied to reduce random noise. Multiple linear regressions (using the "3dDeconvolve" routine in AFNI) were applied on a voxel-wise basis to find the magnitude change when each picture condition was presented, followed with general linear tests to find the statistical significances between stimulus conditions.

The regions of interest (ROI) in this study were defined in the Talairach coordinate space [67]. Regions showing preferential activation to scenes over objects (voxel-based p-value $< 10^{-4}$) in the right and left parahippocampal gyri were defined as the right and left PPA [61]. The right and left V1 ROIs were defined as the regions activated by pictures (voxel-based p-value $< 10^{-10}$) within Brodmann area 17. Because there was a high level of activation at and around V1, a highly conservative p value threshold was chosen to define relatively focal ROIs. The right and left SMC spherical ROIs with 6 mm radius were defined with the centers at (R36, P22, S54) and (L38, P26, S50) correspondingly in the Talairach coordinate space (R = Right, L = Left, P = Posterior, S = Superior). The SMC coordinate

locations were defined by Witt et al. [68] and the ROIs were created as in Zhu et al. [69]. The time courses from the stimulation fMRI dataset that were already pre-processed as above were detrended and had their baselines removed also. The spatially averaged time course at each of the above ROIs was generated for the causality analyses discussed later.

The rs-fMRI pre-processing was also processed in AFNI [70] as commonly applied in the field and as described in details in Zhu et al. [69]. Essentially, slice-timing correction and rigid-body motion correction were carried. Spatial blurring with a full width half maximum of 4 mm was applied to reduce random noise. The time courses were detrended and the baselines were removed. Brain global, cerebrospinal fluid and white-matter mean signals were modeled as nuisance variables and removed from the time courses. Finally, the time courses were band-pass filtered to the range of 0.009 Hz – 0.08 Hz. The spatially averaged time course at each of the above ROIs was generated for the causality analyses discussed later.

4.5.3 Result

In this subsection, we validate the equivalence of DDCM and DI between two regions using experimental fMRI data obtained under both resting state and stimulus based state.

Recall that in the scene-object paradigm, subjects viewed blocks of scenery and object pictures. They were asked to press a button with the right index finger when they saw a block of pictures. We test the robustness of our causality analysis techniques against some expected outcomes: under the stimulation fMRI paradigm, the primary visual cortex (V1) and nearby regions are activated first, followed with activation in the parahippocampal place area (PPA) for higher level scene processing. Some but relatively small activations in the left sensorimotor cortex (SMC) is also expected following V1 activations. Under the resting-state

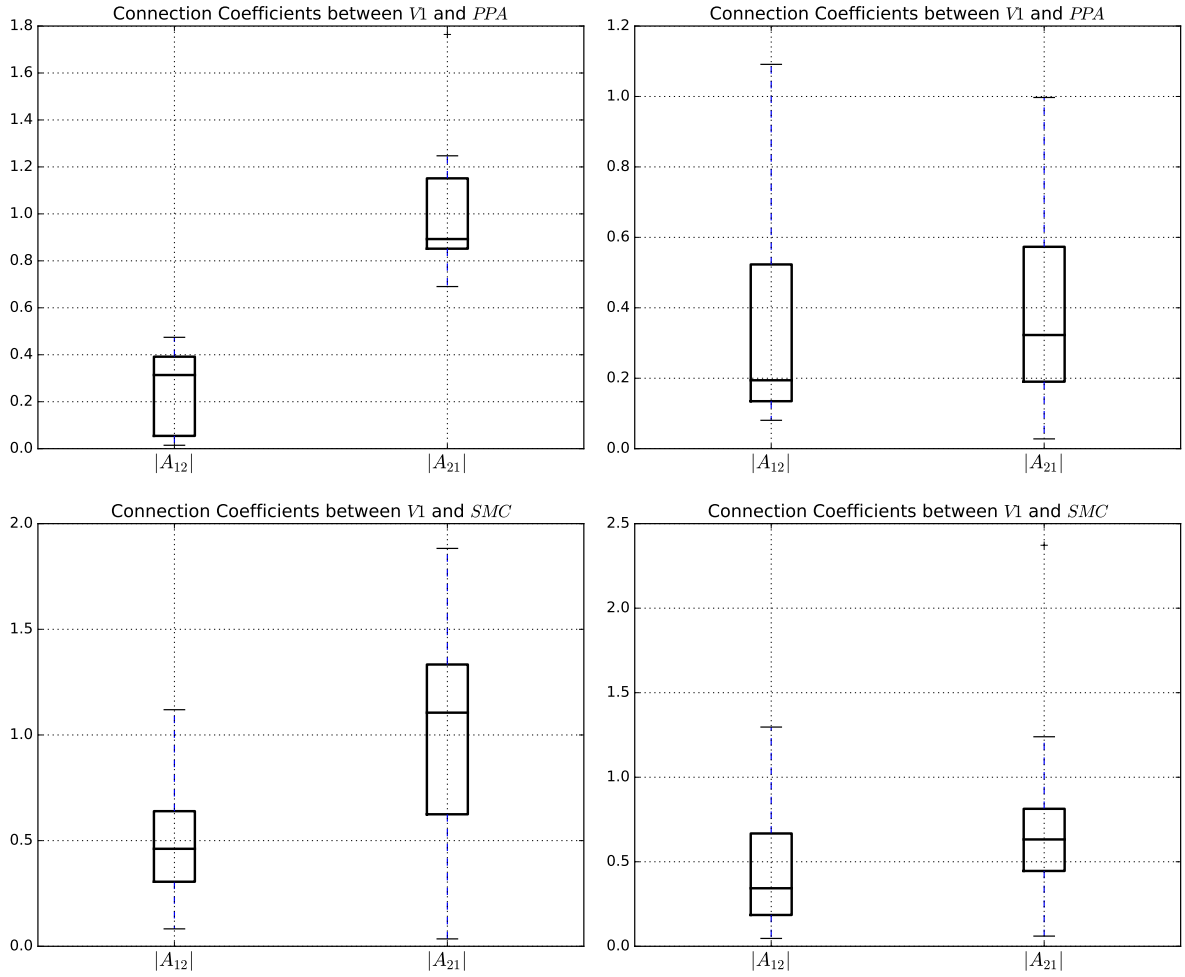


Figure 4.1: Estimations results of DDCM with the experimental fMRI data.

condition, neuronal activity is not expected to occur in a sequential manner among above regions.

The simulation result of V1 and PPA under both resting and stimulus based states are shown in Figure (4.1(a)) and (4.1(b)). It can be seen that under the resting state, V1 does not exhibit a dominating influence over PPA. However, under the stimulus based state, $|A_{21}|$ is increased considerably compared to $|A_{12}|$. In other words, V1 shows stronger influences over PPA as expected. Figure (4.1(c)) and (4.1(d)) have shown a similar pattern for the regions V1 and SMC.

In our previous study [27], we have carried out DI based causality analysis on the same

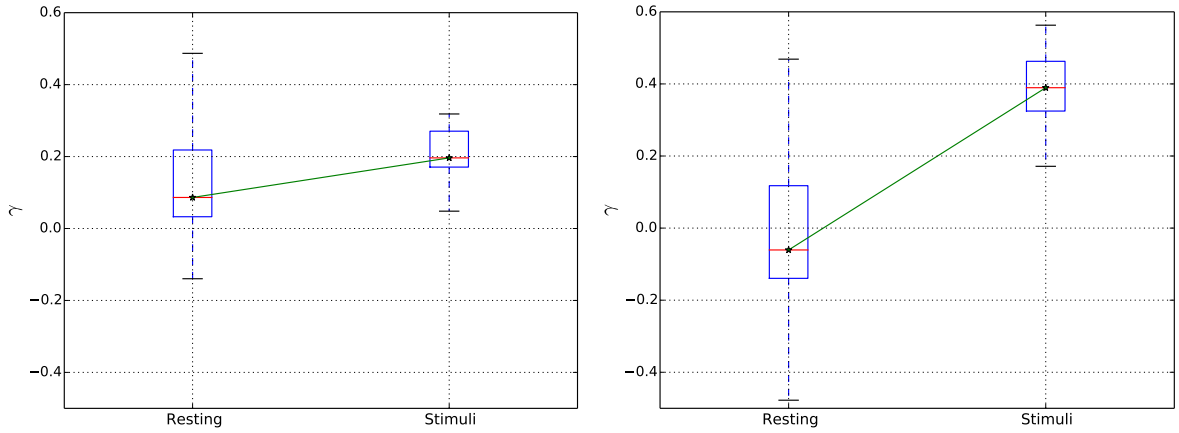


Figure 4.2: Results of directed information based causality analysis.

dataset. The corresponding result is shown in Figure (4.2), where the causal relationship between two brain regions is measured with the metric γ. Given two time series $\mathbf{X}_1^n$ and $\mathbf{X}_2^n$, γ is calculated as

$$\gamma = \frac{I(\mathbf{X}_1^n \to \mathbf{X}_2^n) - I(\mathbf{X}_2^n \to \mathbf{X}_1^n)}{I(\mathbf{X}_1^n; \mathbf{X}_2^n)}, \tag{4.38}$$

where $I(\mathbf{X}_1^n; \mathbf{X}_2^n)$ is the mutual information between two time series. Clearly, $\gamma \in [-1, 1]$. When $|\gamma|$ approaches 1, it can be said with high confidence that there does exist a causal influence between two brain regions; while if $|\gamma|$ is adjacent to 0, it is more likely that no clear causal relationship exists, or the samples in random sequences are subject to strong noises.

It can be seen that in the resting state, the medians of the γ values are within the range $[-0.1, 0.1]$. V1 does not exhibit a dominating causal influence over PPA and SMC. However, under the stimulus based state, the γ values for V1 $\to$ PPA and V1 $\to$ SMC increase significantly. In other words, under the stimulus based state, V1 shows stronger influences over PPA, as well as SMC.

From Figure (4.1) and Figure (4.2), we can see that the causality analysis result obtained with DDCM is in consistent with our previous analysis using the directed information in [27].

4.5.4 Summary of Result

In this section, we applied DDCM to the experimental fMRI data to examine the causal relationship in V1–PPA and V1–SMC region pairs. We observed that under the resting state there was no dominant causal influence for both pairs; while under the stimulus based states, V1 turned out to exhibit more influence over PPA and SMC. The result is consistent with the expectations and our previous result using DI.

4.6 Summary

This chapter investigated the discrete time DCM (DDCM) and its relationship with Directed Information (DI) and Granger Causality (GC). First, we demonstrated the relationship between DDCM and the continuous time DCM. Rather than using approximation, we proved that when the input to the neural dynamic system is a constant, then DDCM can be strictly derived from DCM under the noise free case. This result further validates the DDCM model. Second, based on information theory, we revealed the conditional equivalence between DDCM and DI in characterizing the causal relationship between two brain regions. More specifically, assuming that the dynamic neural system is causal, the neurostate and the noise at each region are normally distributed, and the external input is a constant, we showed that DDCM and DI are equivalent in characterizing the causal relationships between two brain regions. This equivalence between DDCM and DI provides a simple method for DI estimation. We also showed that when the hemodynamic system is invertible, the DI-based causal

relationship between the neurostates of two brain regions is the same with that between the observed BOLD signals. However, it should be pointed out that conditional equivalence between DDCM and DI needs further investigation when multiple regions are involved. Finally, we illustrated the similarities and differences between DDCM and GC. Although they share a similar mathematic form, the causality measures they utilize are completely different. Note that GC detects the causal relationship between the observed signals, and DCM detects the causal connections of the hidden neurostates. The conditional equivalence between GC and DCM remains an interesting problem. The theoretical techniques are demonstrated using fMRI data obtained under both resting state and stimulus based state. Our numerical analysis is consistent with that reported in previous study.

Chapter 5

Classification of Alzheimer's Disease, Mild Cognitive Impairment and Normal Control Subjects Using Resting-State fMRI based Network Connectivity Analysis

This chapter proposes a robust method for the Alzheimer's Disease (AD), mild cognitive impairment (MCI) and normal control (NC) subject classification under size limited fMRI data samples, by exploiting brain network connectivity pattern analysis. First, we select regions of interest (ROIs) within the default mode network (DMN) to formulate a sub-network. We calculate the Pearson correlation coefficients between all possible ROI pairs in the sub-network and use them to form a feature vector for each subject. Second, we propose a regularized linear discriminant analysis (LDA) approach, where we take shrinkage based regularization procedures to reduce the noise effect (including both biological variability and measurement errors) due to limited sample size. The feature vectors are then projected onto a one-dimensional axis using the proposed regularized LDA, where the differences between

AD, MCI and NC subjects are maximized. Based on the Central Limit Theorem, we show that when used for fMRI based brain analysis, LDA is equivalent to the optimal maximum likelihood based classification method. Finally, a decision tree based multi-class AdaBoost classifier, which is robust to noise effect, is applied to the projected one-dimensional vectors to carry out the classification task. Numerical analysis demonstrates that the combination of regularized LDA and the AdaBoost classifier can increase the classification accuracy significantly. Our analysis confirms the previous findings that the hippocampus and the isthmus of the cingulate cortex are closely involved in the development of AD and MCI.

5.1 Introduction

Alzheimer's Disease (AD) is the most common form of dementia, and causes problems with memory, thinking and behavior. It is a degenerative brain disorder, characterized by progressive deterioration of nerve cells, eventually leading to cell death. Mild Cognitive Impairment (MCI) is a condition in which people show a slight, but noticeable and measurable decline in cognitive capabilities, beyond what is considered normal for their age. Older people with MCI may or may not progress to AD, though they have a higher risk of doing so. Accurate distinction of AD and MCI from normal control (NC) subjects is critical for early diagnosis and treatment of brain disorders.

Traditional AD and MCI diagnosis methods are generally based on positron emission tomography (PET) and cerebrospinal fluid (CSF) [76]. In recent years, there has also been an increasing interest in noninvasive diagnosis methods based on electroencephalography (EEG) [77], structural magnetic resonance imaging (MRI) [78], and functional magnetic resonance imaging (fMRI) [36, 79].

In literature, the majority of existing noninvasive classification approaches rely on MRI and EEG. In [80], Pritchard et al. analyzed spectral-band measures of EEG data acquired from AD patients and NC subjects. They found that classifiers based on multivariate discriminant analysis [81] and the nearest neighbor approach could typically achieve a two-category (AD and NC) classification accuracy of 80% when applied to EEG data. At the same time, they showed that the accuracy could be improved if nonlinear EEG measures were added. In [82], Magnin et al. applied the support vector machine (SVM) classifier to the whole-brain anatomical MRI data acquired from AD patients and NC subjects. They formulated feature vectors for classification using gray matter information extracted from T1-weighted MR images of AD, MCI and NC subjects, and achieved a two-category classification accuracy of 94.5%. In [83], Korolev et al. combined data from clinical biomarkers, MRI signals, and plasma biomarkers and developed a classifier to predict whether an MCI patient would develop Alzheimer's disease over a three-year period. Their prediction accuracy was 80%.

More recently, fMRI, which maps brain activities to metabolic changes (such as the blood-oxygen-level dependent (BOLD) contrast) in cerebral blood flow, has also been used to classify AD, MCI and NC subjects [36, 79]. The underlying argument is that cerebral blood flow and neuronal activation are coupled. That is, when a particular brain region becomes active, blood flow to that region also increases.

Compared with EEG, fMRI data can display active brain areas more directly, has much better spatial resolution throughout the brain. Unlike structural MRI which mainly reflects the anatomical information of brain tissues and structure, fMRI focuses more on functional brain activities, and can provide more direct measurement on how different brain regions are involved in particular brain activities, hence provide more insight on the changes of functional brain connectivity during the evolution of MCI and AD.

In [36], Wang et al. extracted two intrinsically anti-correlated networks using resting state fMRI data from 14 AD patients and 14 NC subjects, and applied a Pseudo-Fisher Linear Discriminative Analysis (pFLDA) on the high dimensional feature vectors. Their *two-category* classification accuracy was 83%. In [79], Chen et al. applied the same technique to larger datasets. Similarly, the accuracy of *two-category* classification of AD patients and NC subjects was 82%.

While structural MRI has been widely applied to clinical diagnosis of brain disorders, fMRI has mainly been used for research purposes. As a result, the size of fMRI data samples is generally quite limited, which has become a major bottleneck in fMRI based AD, MCI and NC classification. The underlying reason is that, when the sample size is small, most existing classifiers suffer from severe noise effects, due to both biological variability and measurement noise.

Motivated by this observation, in this chapter, we develop a reliable method for AD, MCI and NC classification that is robust with respect to *size limited* fMRI data samples, by exploiting *brain network connectivity pattern analysis*. More specifically, we take multiple regions of interest (ROIs) in the brain, formulate a sub-network, and then analyze the network connectivity pattern by evaluating the correlation between all ROI pairs within the sub-network. The underlying argument is that: due to variability in the brain connectivity of each individual, the connectivity between two brain regions alone may not be sufficient to distinguish NC subjects from patients with cognitive impairments; brain network connectivity pattern analysis, which looks for subtle changes in the pattern of connectivity among multiple or all regions in the sub-network, may reveal more in-depth information.

The proposed classification scheme can be described as follows. *First*, we select an ROI sub-network and formulate the feature vectors by calculating the Pearson correlation coef-

ficients between all pairs of ROIs. In this chapter, we formulate the ROI sub-network by selecting regions within the default mode network (DMN), which denotes the network of brain regions that are active when the brain is at the resting state [84]. Prior resting-state fMRI studies have demonstrated that the DMN is affected by AD [42, 85–88]. More specifically, in this chapter, we select the right and left hippocampi and isthmus of the cingulate cortices (ICCs) (4 regions) as our ROI sub-network. This is because that both hippocampus and ICC are part of the DMN, and can be well defined anatomically through the FreeSurfer software [88], even in brains with abnormal anatomy [42]. It has also been demonstrated [42] that the functional connection between hippocampus and ICC was reduced in AD. *Second*, we propose a regularized linear discriminant analysis (LDA) approach, where we take shrinkage based regularization procedures to reduce the noise effect (including both biological variability and measurement errors) due to limited sample size. The feature vectors are then projected onto a one-dimensional axis using the proposed regularized LDA, where the differences between AD, MCI and NC subjects are maximized. Based on the Central Limit Theorem, we show that when used for fMRI based brain functioning classification, LDA is equivalent to the optimal maximum likelihood based classification method. *Finally,* a decision tree based *multi-class* AdaBoost classifier, which is robust to noise effect, is applied to the projected one-dimensional vectors to carry out the classification task.

The major results of this chapter can be summarized as:

1. *We propose a regularized LDA approach, which aims to reduce the noise effect by using two shrinkage methods.* The first shrinkage method moves the estimated mean of each class towards the overall mean, and the second one shifts the estimated covariance matrix for each class towards the identity matrix. Numerical analysis shows that: in comparison with the original LDA approach [41], the regularized LDA can reduce the

noise effect and increase the classification accuracy significantly.

2. *We investigate the relationship between LDA-based and Maximum Likelihood (ML) based classification or decision making methods.* Recall that LDA aims to separate two or more classes by projecting them into a subspace or direction where different classes show most significant differences [41]. In this chapter, we prove that when the original data are normally distributed, LDA is equivalent to maximizing the log-likelihood function of the projected data. Note that there are millions of neurons within one fMRI voxel, according to the Central Limit Theorem, the overall fMRI signal corresponding to each voxel follows the normal distribution approximately. This implies that when used for fMRI based brain functioning classification, LDA is equivalent to the optimal ML based classification method.

3. *We conduct the connectivity pattern classification of AD, MCI and NC subjects by applying the regularized LDA and AdaBoost classifier based approach.* First, we calculate the Pearson correlation coefficients between all possible pairs of the ROIs within the group to formulate the feature vectors. Second, the feature vectors are then projected onto a one-dimensional axis using the proposed regularized LDA, where the differences between AD, MCI and NC subjects are maximized. Finally, we construct the decision tree based on the projected feature vectors and carry out the classification using the multi-class AdaBoost classifier.

   In this chapter, we choose to utilize the AdaBoost classifier instead of the naive Bayesian classifier, since it has been observed consistently in literature that: the AdaBoost classifier could achieve significantly higher classification accuracy than the naive Bayesian classifier when the sample size is very limited [89]. Our numerical results

demonstrate that: (i) LDA-Bayesian classifier can achieve a three-category (AD, MCI and NC) classification accuracy of 44%; (ii) LDA-AdaBoost classifier can increase the accuracy to 69%; (iii) when AdaBoost is combined with the regularized LDA, the accuracy can be further increased to 75%.

As expected, it is also observed that compared with AD and NC subjects, it is more difficult for the classifier to identify MCI subjects. The classification accuracy for AD and NC are as high as 80% and 83%, respectively, while the accuracy for MCI is only 63%. Our analysis also confirms the previous findings that the hippocampus and the isthmus of the cingulate cortex are closely involved in the development of AD and MCI.

The rest of this chapter is organized as follows. In Section II, we present the proposed regularized LDA approach, and explore the relationship between LDA based and the Maximum Likelihood based classification methods. In Section III, we describe the ROI sub-network formulation, and elaborate how to perform AD, MCI and NC classification through connectivity pattern analysis. In Section IV we present the numerical results, and we conclude in Section V.

5.2 Regularized Linear Discriminant Analysis

In this section, first, we revisit the Linear Discriminant Analysis method. Second, we integrate two shrinkage methods with the original LDA to formulate the regularized LDA. Finally, we investigate the relationship between LDA and the ML estimation method.

5.2.1 Linear Discriminant Analysis

Linear Discriminant Analysis aims to separate two or more classes by projecting them into a subspace or direction where different classes show most significant differences [41]. Here, we illustrate the basic idea of LDA using the three-class case. Suppose we have a set of $d-$dimensional vector samples $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where n_1 of them are from the first class, denoted as C_1, and n_2 of them are from the second class, denoted as C_2, and the remaining $n_3 = n - n_1 - n_2$ of them are from the third class, denoted as C_3. For $i = 1, 2, 3$, the mean and scatter matrix (i.e., the scaled covariance matrix) of each of the three classes are defined as:

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}, \tag{5.1}$$

$$S_i = \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^t. \tag{5.2}$$

Consider the projection of vectors in X to a new $d-$dimensional space:

$$\mathbf{y} = W\mathbf{x}, \ \mathbf{x} \in X, \tag{5.3}$$

where W is a $d \times d$ matrix to be determined by the LDA algorithm. In this chapter, we only utilize the first dimension y of projected vector $\mathbf{y}$ where the differences among three classes are maximized. As a result, Equation (5.3) can be rewritten as:

$$y = \mathbf{w}^t \mathbf{x}, \tag{5.4}$$

where $\mathbf{w}^t$ is the first row of the matrix W. For $i = 1, 2, 3$, let

$$\tilde{C}_i = \{y = \mathbf{w}^t\mathbf{x} \mid \mathbf{x} \in C_i\}. \tag{5.5}$$

Define $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ as the overall mean, $S_W = \sum_{i=1}^{3} S_i$ as the within-class scatter matrix, and the between-class scatter matrix S_B as:

$$S_B = \sum_{i=1}^{3} n_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^t. \tag{5.6}$$

LDA seeks a transform vector $\mathbf{w}$ that maximizes the following objective function:

$$J(\mathbf{w}) \;=\; \frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_W \mathbf{w}}. \tag{5.7}$$

It can be proved [41, 81] that to maximize Equation (5.7), $\mathbf{w}$ should satisfy

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}, \tag{5.8}$$

for some constant λ. Performing eigenvalue decomposition to matrix $S_W^{-1}S_B$, LDA then chooses the eigenvector corresponding to the largest eigenvalues of the matrix $S_W^{-1}S_B$ as $\mathbf{w}$. As will be shown in Section III, various classifiers, such as the Bayesian classifier and the AdaBoost classifier can then be applied to the projected vectors $\{y_i = \mathbf{w}^t\mathbf{x}_i\}_{i=1}^{n}$ for further classification.

5.2.2 Regularized LDA

The original LDA algorithm has been widely applied in supervised learning problems [81]. However, as mentioned earlier, when the total number of subjects is small, the estimated statics suffer considerably from the noise effect caused by both biological variability and measurement error, leading to low classification accuracy. For our fMRI based AD, MCI and NC classification, due to the very limited sample size, LDA together with the Bayesian classifier could only achieve an accuracy that is under 50%. To reduce the noise effect, we propose to regulate the original LDA using two shrinkage methods.

5.2.2.0.3 Shrinkage of the Mean The first shrinkage method, originally proposed by Tibshirani et al. [90] for gene expression profiling, adjusts the estimated mean vectors. In our case, let $C = \bigcup_{i=1}^{3} C_i$ be the whole sample set. Recall that for any $\mathbf{x} \in C$, $\mathbf{x} = [\mathbf{x}(1), \cdots, \mathbf{x}(d)]^t$. For $i = 1, 2, 3$, $k = 1, \cdots, d$, define $\mu_{i,k} \triangleq \frac{1}{n_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}(k)$, and $\mu_k \triangleq \frac{1}{n} \sum_{\mathbf{x} \in C} \mathbf{x}(k)$, where $n_i = |C_i|$ and $n = n_1 + n_2 + n_3$. Let $\boldsymbol{\mu}_i = [\mu_{i,1}, \mu_{i,2}, ..., \mu_{i,d}]^t$, and $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_d]^t$. The algorithm first calculates the following variances:

$$\xi_k^2 = \frac{1}{n-3} \sum_{i=1}^{3} \sum_{\mathbf{x} \in C_i} [\mathbf{x}(k) - \mu_{i,k}]^2, \qquad k = 1, \cdots, d. \tag{5.9}$$

Then for $i = 1, 2, 3$, and $k = 1, \cdots, d$, the scaled distance of $\mu_{i,k}$ to the centroid μ_k is calculated as:

$$d_{i,k} = \frac{\mu_{i,k} - \mu_k}{m_i \xi_k}, \tag{5.10}$$

where $m_i = \sqrt{1/n_i + 1/n}$. After that, the distance is shrunken as follows:

$$d_{i,k} \leftarrow sign(d_{i,k})max(0, |d_{i,k}| - \delta), \tag{5.11}$$

where δ is a positive step size determined through cross-validation [81]. Now based on Equation (5.10), the shrinkage is achieved as follows:

$$\mu_{i,k} \leftarrow \mu_k + m_i \xi_k d_{i,k}. \tag{5.12}$$

As can be seen from (5.11) and (5.12), each dimension of $\boldsymbol{\mu}_i$ has been shrunken towards the overall mean. This shrinkage method is essentially based on the t test between the mean of each class and the overall mean at every dimension. Recall that the t score of two sets of random variables $\{z_1\}$ and $\{z_2\}$, which have the same standard deviation S_z, is defined as:

$$t = \frac{\mu_{z_1} - \mu_{z_2}}{s_z \sqrt{\frac{1}{n_{z_1}} + \frac{1}{n_{z_2}}}}, \tag{5.13}$$

where $\mu_{z_1} = E\{z_1\}$, $\mu_{z_2} = E\{z_2\}$, and n_{Z_1} and n_{Z_2} are the sample sizes [91]. In this shrinkage method, for $i = 1, 2, 3$, $k = 1, \cdots, d$, the t score between each $\mu_{i,k}$ and μ_k pair is defined as a distance in (5.10). If the distance $d_{i,k}$ is small, i.e., if $d_{i,k} < \delta$, then most likely it is caused by the noise effect. In this case, the shrinkage method forces $d_{i,k}$ to be zero, and therefore reduces the noise effect.

5.2.2.0.4 Shrinkage of the Covariance Matrix The second shrinkage method, proposed by Friedman et al. [92], regulates the estimation of covariance matrix S_i for each class

as:

$$S_i \leftarrow (1-\epsilon)S_i + \epsilon I, \qquad i = 1, 2, 3, \tag{5.14}$$

where I is the identity matrix, and ϵ a positive number determined through cross-validation. The basic idea of this shrinkage method is that: when the sample size is small, the estimated covariance matrix $S_i,\ i = 1, 2, 3$, generally becomes non-invertible. By adding a small perturbation to the slightly scaled covariance matrix, the adjusted or shrunken S_i will generally become invertible as expected.

After the regularized LDA transform, the feature vectors are projected into a set of real valued numbers. After that, a selected classifier can be applied to the transformed data for further classification.

5.2.3 LDA and the ML Estimation

In this subsection, we demonstrate that when the original data from all classes are normally distributed, then LDA is equivalent to the ML method. For $i = 1, 2, 3$, assuming each vector $\mathbf{x}_j$ in class C_i has the same probability density function (pdf):

$$f_X(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{2\pi^d |\Sigma_i|}} exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t {\Sigma_i}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\}. \tag{5.15}$$

Consider a general linear transform defined by:

$$\mathbf{y} = W\mathbf{x}. \tag{5.16}$$

where W is a $d \times d$ matrix. For the transformed data, the probability density function becomes:

$$f_Y(\mathbf{y}; \tilde{\boldsymbol{\mu}}_i, \tilde{\Sigma}_i) = \frac{1}{\sqrt{2\pi^d |\tilde{\Sigma}_i|}} exp\{-\frac{1}{2}(\mathbf{y} - \tilde{\boldsymbol{\mu}}_i)^t \tilde{\Sigma}_i^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}}_i)\}, \tag{5.17}$$

where $i = 1, 2, 3$, $\tilde{\boldsymbol{\mu}}_i = W\boldsymbol{\mu}_i$, and $\tilde{\Sigma}_i = W\Sigma_i W^t$.

Recall that in LDA, we try to find W such that the difference among different classes is maximized in the transformed space. Without loss of generality, we assume that the major difference lies in the first dimension of the transformed vector $\mathbf{y}$ only, and the remaining $d-1$ dimensions make little contributions. Under this assumption, $\tilde{\boldsymbol{\mu}}_i$ and $\tilde{\Sigma}_i$ can be decomposed as:

$$\tilde{\boldsymbol{\mu}}_i = \begin{bmatrix} \tilde{\mu}_i^1 \\ \tilde{\boldsymbol{\mu}}^{d-1} \end{bmatrix}, \ \tilde{\Sigma}_i = \begin{bmatrix} \tilde{\Sigma}_i^1 & \mathbf{0} \\ \mathbf{0} & \tilde{\Sigma}^{d-1} \end{bmatrix}, \tag{5.18}$$

since for each i, $\tilde{\boldsymbol{\mu}}_i^{d-1} \approx \tilde{\boldsymbol{\mu}}^{d-1}$, $\tilde{\Sigma}_i^{d-1} \approx \tilde{\Sigma}^{d-1}$. Accordingly, the matrix W can also be decomposed as

$$W = \begin{bmatrix} W^1 \\ W^{d-1} \end{bmatrix}. \tag{5.19}$$

In this case, we have $\tilde{\mu}_i^1 = W^1\mu_i$, $\tilde{\boldsymbol{\mu}}_i^{d-1} = W^{d-1}\tilde{\boldsymbol{\mu}}_i$, $\tilde{\Sigma}_i^1 = W^1\Sigma_i {W^1}^t$ and $\tilde{\Sigma}_i^{d-1} = W^{d-1}{\Sigma_i}{W^{d-1}}^t$.

For fairness, in LDA based classification, the sample size of the three classes is assumed to be the same, i.e., $n_1 = n_2 = n_3 = n/3$. With the probability density function given in

(5.17), the log-likelihood function of the transformed data can be written as [93]:

$$\begin{aligned} L(W) &= \sum_{i=1}^{3} \sum_{\mathbf{y} \in \tilde{C}_i} log\ f_Y(\mathbf{y}; \tilde{\boldsymbol{\mu}}_i, \tilde{\Sigma}_i) \\ &= nlog|W| - \frac{n}{2} log(2\pi)^d - \sum_{i=1}^{3} \frac{n_i}{2} log|\tilde{\Sigma}_i^1| \\ &- \frac{1}{2} \sum_{i=1}^{3} \sum_{\mathbf{x} \in C_i} (W^1\mathbf{x} - \tilde{\mu}_i^1)^t (\tilde{\Sigma}_i^1)^{-1} (W^1\mathbf{x} - \tilde{\mu}_i^1) \\ &- \frac{n}{2} log|\tilde{\Sigma}^{d-1}| \\ &- \frac{1}{2} \sum_{\mathbf{x} \in C} (W^{d-1}\mathbf{x} - \tilde{\boldsymbol{\mu}}^{d-1})^t (\tilde{\Sigma}^{d-1})^{-1} (W^{d-1}\mathbf{x} - \tilde{\boldsymbol{\mu}}^{d-1}) \end{aligned} \tag{5.20}$$

To find the optimal W that maximizes $L(W)$, set $\frac{\partial L(W)}{\partial \tilde{\Sigma}_i^1} = 0$ and $\frac{\partial L(W)}{\partial \tilde{\Sigma}^{d-1}} = 0$, we get:

$$\tilde{\Sigma}_i^1 = W^1 S_W W^{1^t}, \tag{5.21}$$

$$\tilde{\Sigma}^{d-1} = W^{d-1} S_B W^{d-1^t}. \tag{5.22}$$

Substitute (5.21) and (5.22) into (5.20) and remove the constant items, the optimization of $L(W)$ is equivalent to optimizing the following function:

$$L_{eq}(W) = nlog|W| - \frac{n}{2} log|W^1 S_W W^{1^t}| - \frac{n}{2} log|W^{d-1} S_B W^{d-1^t}|. \tag{5.23}$$

The optimal choice of W will satisfy the differential equation:

$$\frac{dL_{eq}(W)}{dW} = 0. \tag{5.24}$$

It was shown in [94–96] that the partial differential equations are satisfied when W is com-

posed of eigenvectors of the matrix $S_W^{-1}S_B$.

If we only keep the eigenvector corresponding to the largest eigenvalue of $S_W^{-1}S_B$, then we obtain the LDA algorithm presented in Section 5.2.1. As can be seen, LDA is equivalent to the ML method.

5.3 Classification of AD, MCI and NC Subjects based on Connectivity Pattern Analysis

In this section, we formulate the ROI sub-network, and perform AD, MCI and NC Subjects classification through connectivity pattern analysis, by exploiting the proposed Regularized LDA.

5.3.1 ROI Sub-Network Formulation and Connectivity Pattern Analysis

The default mode network (DMN) is one of the well studied networks at the resting state [84]. Prior resting-state fMRI studies have demonstrated that the DMN is affected by AD [42,85–88]. Both hippocampus and ICC are part of the DMN, and can be well defined anatomically through the FreeSurfer software [88], even in brains with abnormal anatomy [42]. The paper by Zhu et al. [42] specifically demonstrated that the functional connection between hippocampus and ICC was reduced in AD.

Motivated by the observations above, in this chapter, we select the right and left hippocampi and ICCs (4 regions) as our ROI sub-network. Our connectivity pattern analysis is carried out following the procedure below.

First, we calculate the Pearson correlation coefficients between all possible pairs of the ROIs within the group to formulate the feature vectors. As we now have 4 regions in the ROI sub-network, for each subject i, we can obtain a $d-$dimensional ($d = 6$) vector $\mathbf{x}_i$, consisting of the Pearson correlation coefficients for each pair of ROIs. When we have n subjects, we get the feature vector set $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$.

Second, using the proposed regularized LDA, we map $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ to a one-dimensional subspace or axis, where the differences between AD, MCI and NC subjects are maximized, and denote the projected vectors as $\{y_1, \cdots, y_n\}$.

Finally, we construct the decision tree based on $\{y_1, \cdots, y_n\}$ and carry out the classification using the multi-class AdaBoost classifier.

In the following subsections, we will provide more details on decision tree construction, and multi-class classification using AdaBoost.

5.3.2 Basic Decision Tree Construction

In the classification procedure, we will construct $T = 50$ basic decision trees. Each decision tree divides the LDA projected data set $y_i, i = 1, \cdots, n$, into K regions, and each region is called a leaf node. Here, the number of regions, K, and the boundaries for all the regions are chosen by the decision tree algorithm to minimize the Gini impurity coefficient I_G [97]. More specifically, assuming $y_i \in c_i$, where $c_i \in \{\tilde{C}_1, \tilde{C}_2, \tilde{C}_3\}$. Here $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3$ denote the projected data set corresponding to AD, MCI and NC subjects, respectively. For $k = 1, 2, ..., K$, without loss of generality, suppose k_m data samples $\{(y_{k_1}, c_{k_1}), (y_{k_2}, c_{k_2}), ..., (y_{k_m}, c_{k_m})\}$ are assigned to node k, where $y_{k_l} \in \{y_i\}, c_{k_l} \in \{\tilde{C}_1, \tilde{C}_2, \tilde{C}_3\}$, $k_l = 1, \cdots, k_m$. The Gini

impurity coefficient of node k is calculated as:

$$I_G(k) = \sum_{k_i=1}^{k_m} f_{k_i}(1 - f_{k_i}), \tag{5.25}$$

where $f_{k_i} = \frac{number\ of\ occurance\ of\ c_{k_i}\ within\ node\ k}{k_m}$.

For any given input y to be classified, if y falls within the boundaries of node k, then it will be assigned to node k, and paired with the majority class inside this node. Note that in our case, $y_i, i = 1, \cdots, n$, are all real-valued numbers. That is, $\{y_i\} \in R$. In this case, the boundary between two neighboring regions is reduced to a point, and hence each reigon corresponds to an interval on the R axis.

The decision tree is a weak classifier. In most applications, it needs to be incorporated with an ensemble classifier to achieve higher accuracy. Some representative ensemble algorithms include Bagging and Boosting [98]. Note that the naive Bayesian classifier is subject to severe overfitting problems when the sample size is limited, leading to low classification accuracy. In the following, we will apply the AdaBoost algorithm to construct the ensemble classifier, due to its robustness under noise effect [81, 89].

5.3.3 The Multi-Class AdaBoost Classifier

The multi-class AdaBoost classifier is built upon an ensemble of weak decision tree classifiers [89]. Given a set of labeled data $\{(y_1, c_1), (y_2, c_2), ..., (y_n, c_n)\}$, where $c_i \in \{\tilde{C}_1, \tilde{C}_2, \tilde{C}_3\}$, the algorithm first starts with an empty ensemble and T decision trees, as constructed above. Each sample y_i in the date set is given an initial weight $w_i = 1/n$, where $i = 1, 2, \cdots, n$. Then for $t = 1, 2, \cdots, T$, the algorithm will iteratively implement the following procedures:

1. Weighted Classification Error Calculation Apply a weak decision tree classifier t to

the samples and calculate the weighted classification error. More specifically, let

$$\mathbf{I}(c_i, \tilde{c_i}) = \begin{cases} 1 & if \quad c_i \neq \tilde{c_i}, \\ 0 & if \quad c_i = \tilde{c_i}, \end{cases} \tag{5.26}$$

where $\tilde{c_i}$ is the assigned class for sample y_i, and c_i is the true class y_i belongs to. Then the weighted classification error e_t would be

$$e_t = \sum_{i=1}^{n} w_i \ \mathbf{I}(c_i, \tilde{c_i}). \tag{5.27}$$

2. Voting Weight Assignment Based on the weighted classification error e_t, the algorithm will assign a voting weight α_t for the weak decision tree classifier t as follows:

$$\alpha_t = ln\frac{1 - e_t}{e_t} + ln2, \tag{5.28}$$

and then add classifier t into the ensemble.

3. Weight Update Before the next iteration, the weight of each data sample y_i is updated as follows:

$$w_i \quad \leftarrow \quad w_i e^{\alpha_t \mathbf{I}(c_i, \tilde{c_i})}, \tag{5.29}$$

$$w_i \quad \leftarrow \quad \frac{w_i}{\sum_{i=1}^{n} w_i}. \tag{5.30}$$

The procedure in (5.30) ensures that the weights $\{w_i\}, i = 1, 2, \cdots, n$, form a probability distribution with $\sum_{i=1}^{n} w_i = 1$. As can be seen, after the update, those samples which have been incorrectly classified in current iteration will have higher weights in the next iteration.

4. Final Classification After T iterations, there will be T decision trees in the ensemble.

The final classification is a weighted majority votes of each of those classifiers.

As will be shown in the next Section, the combination of regularized LDA and AdaBoost can achieve much higher accuracy in AD, MCI and NC classification than the conventional approach based on the original LDA and the Bayesian classifier.

5.4 Numerical Analysis

5.4.1 fMRI Data Acquisition

In our data collection process, 10 patients with mild-to-moderate probable Alzheimer's Disease, 11 patients with MCI and 12 age- and education-matched healthy NC subjects were recruited to participate in this study. The fMRI experiment was conducted on a GE 3T $Signa^{®}$ HDx MR scanner (GE Healthcare, Waukesha, WI) with an 8-channel head coil. To study resting-state brain function, echo-planar images, starting from the most inferior regions of the brain, were acquired for 7 minutes with the following parameters: 38 contiguous $3mm$ axial slices in an interleaved order, time of echo $= 27.7ms$, time of repetition $= 2500ms$, flip angle $= 80^{\circ}$, field of view $= 22cm$, matrix size $= 64 \times 64$, ramp sampling, and with the first four data points discarded. Each volume of slices was acquired 164 times. Common pre-processing procedures on resting state fMRI data were carried as detailed in [42].

5.4.2 Performance Comparison of Different Classification Algorithms

In this subsection, we present the classification performance of the proposed method and compare it to existing methods. Since the size of data samples is small, the performance

of the classifier is evaluated by the Leave-One-Out (LOO) cross-validation. As described earlier, the ROIs used are the hippocampus and ICC from both hemispheres of the brain.

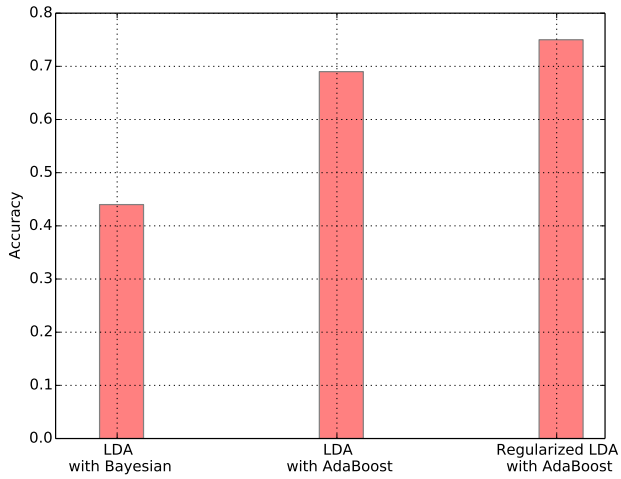


Figure 5.1: Comparison of 3-category (AD, MCI, NC) classification results.

Figure (5.1) shows the performance of three classifiers. In the first one, a naive Bayesian classifier is employed after the original LDA transform. As can be seen, its final accuracy is only 44%. As explained in Section II, the main reason of such an unsatisfying performance is that: when the number of data samples is small, the estimation of class means and covariance matrices in LDA suffers from severe noise effect, leading to overfitting. In the second one, the original LDA is combined with the AdaBoost classifier. As can be seen, accuracy is increased to 69% by AdaBoost. The third one is what we proposed, the regularized LDA is combined with the AdaBoost classifier. The shrinkage operations in the regularized LDA can reduce the noise effect in the estimation, and further improve the accuracy to 75%.

Figure 5.2 shows the classification results of all the three classes of subjects using regularized LDA and the AdaBoost classifier. As expected, compared with NC subjects and AD patients, it is more difficult for the classifier to identify MCI patients, and the classification accuracy for MCI patients is much lower than that for AD and NC subjects.

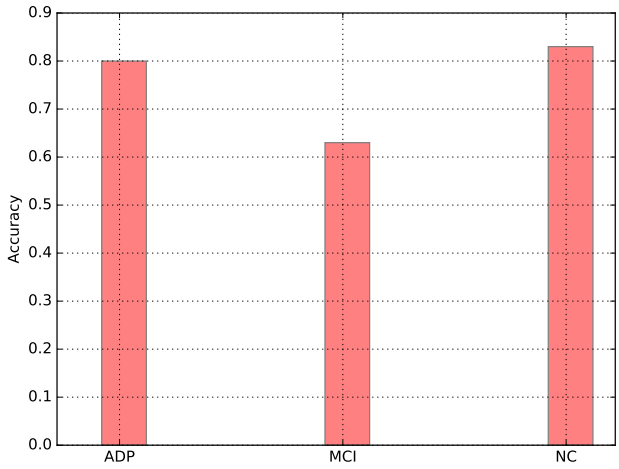


Figure 5.2: Regularized LDA with AdaBoost classifier: classification accuracy for different categories.

5.5 Summary

This chapter proposed a reliable method for AD, MCI and NC subject classification that is highly robust under *size limited* fMRI data samples, by exploiting *brain network connectivity pattern analysis*. To do it, *first*, we selected the right and left hippocampi regions and isthmus of the cingulate cortices (ICCs) as our ROI sub-network, and calculated the Pearson correlation coefficients between all possible ROI pairs and used them to form a feature vector for each subject. *Second*, the vectors were projected into a one-dimensional axis using the proposed regularized LDA approach, where the differences between AD, MCI and NC subjects were maximized. Shrinkage based regularization procedures were taken to reduce the noise effect due to the limited sample size. *Finally*, a decision tree based multi-class AdaBoost classifier, which is robust to noise effect, was applied to the projected one-dimensional vectors to perform the classification.

Both the theoretical and numerical analysis demonstrated that: (i) The regularization methods and the AdaBoost classifier can increase the classification accuracy significantly;

(ii) Brain network connectivity analysis, which evaluates the changes in the pattern of connectivity among multiple or all regions in the sub-network, can reveal in-depth information about brain connectivity and result in relatively accurate classification of AD, MCI and NC, especially when the sample size is very limited; (iii) Our analysis confirms the previous findings that the hippocampus and the isthmus of the cingulate cortex are closely involved in the development of AD and MCI.

Potentially, the proposed framework can be applied to other classification problems as well, especially under limited sample size.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Brain functional connectivity analysis has been shown as an essential method in revealing the organization of brain networks, identifying function hubs and discovering biological patterns. In this research, we developed innovative modeling and analysis methodologies by exploiting advanced techniques in network communication and information theory, and exploring how the proposed techniques can help us understand cognitive process and identify brain problems.

First, in Chapter 2, we considered the measurement of non-directional connectivities between brain regions using mutual information. We proposed a novel approach for the estimation of MI, which was composed of three major components: de-correlation, kernel based estimation of probability density function and Monte Carlo Integration. The proposed MI estimator was applied to experimental fMRI data obtained from Alzheimer's Disease patients and normal subjects. The numerical result was consistent with clinical observations.

Next, in Chapter 3, we presented the directed information framework and showed how to apply it for fMRI causality analysis. We provided the detailed procedure on how to calculate the DI for two finite time series. The two major steps involved here are optimal bin size selection for data digitization, and probability estimation. We applied the DI based causality analysis to both the simulated data and experimental fMRI data, and compared the

results with that of the Granger Causality analysis. Our results indicated that GC analysis is effective in detecting linear or nearly linear causal relationship, but has difficulty in capturing nonlinear causal relationships. On the other hand, DI based causality analysis is effective in capturing both linear and non-linear causal relationships. Moreover, it was observed that brain connectivity among different regions generally involves dynamic two-way information transmissions between them. Our results showed that when bidirectional information flow is present, DI is more effective than GC to quantify the overall causal relationship.

Then, in Chapter 4, we investigated the discrete time DCM (DDCM) and its relationships with Directed Information and Granger Causality. First, we demonstrated the relationship between DDCM and the continuous time DCM. Rather than using approximation, we proved that when the input to the neural dynamic system is a constant, then DDCM can be strictly derived from DCM under the noise free case. This result further validates the DDCM model. Second, based on information theory, we revealed the conditional equivalence between DDCM and DI in characterizing the causal relationships between two brain regions. More specifically, assuming that the dynamic neural system is causal, the neurostate and the noise at each region are normally distributed, and the external input is a constant, we showed that DDCM and DI are equivalent in characterizing the causal relationships between two brain regions. This equivalence between DDCM and DI provides a simple method for DI estimation. We also showed that when the hemodynamic system is invertible, the DI-based causal relationship between the neurostates of two brain regions is the same with that between the observed BOLD signals. However, it should be pointed out that conditional equivalence between DDCM and DI needs further investigation when multiple regions are involved. Finally, we illustrated the similarities and differences between DDCM and GC. Although they share a similar mathematic form, the causality measures they utilize are

completely different. Note that GC detects the causal relationship between the observed signals, and DCM detects the causal connections of the hidden neurostates. The conditional equivalence between GC and DCM remains an interesting problem. The theoretical techniques were demonstrated using fMRI data obtained under both resting state and stimulus based state. Our numerical analysis was consistent with that reported in previous study.

Finally, in Chapter 5, we proposed a reliable method for AD, MCI and NC subject classification that is highly robust under *size limited* fMRI data samples, by exploiting *brain network connectivity pattern analysis*. To do it, *first*, we selected the right and left hippocampi regions and isthmus of the cingulate cortices (ICCs) as our ROI sub-network, and calculated the Pearson correlation coefficients between all possible ROI pairs and used them to form a feature vector for each subject. *Second*, the vectors were projected into a one-dimensional axis using the proposed regularized LDA approach, where the differences between AD, MCI and NC subjects were maximized. Shrinkage based regularization procedures were taken to reduce the noise effect due to the limited sample size. *Finally*, a decision tree based multi-class AdaBoost classifier, which is robust to noise effect, was applied to the projected one-dimensional vectors to perform the classification.

Both the theoretical and numerical analysis demonstrated that: (i) The regularization methods and the AdaBoost classifier can increase the classification accuracy significantly; (ii) Brain network connectivity analysis, which evaluates the changes in the pattern of connectivity among multiple or all regions in the sub-network, can reveal in-depth information about brain connectivity and result in relatively accurate classification of AD, MCI and NC, especially when the sample size is very limited; (iii) Our analysis confirms the previous findings that the hippocampus and the isthmus of the cingulate cortex are closely involved in the development of AD and MCI.

Potentially, the proposed framework can be applied to other classification problems as well, especially under limited sample size.

6.2 Future Work

We propose the following directions for the future research.

Brain network connectivity pattern analysis – from Pearson correlation to mutual information

- In Chapter 2 and Chapter 5, interesting results had been drawn based on mutual information and network connectivity pattern analysis. Further research can be conducted on MI (including multi-voxel MI) based network connectivity pattern analysis. It is important to develop new classification algorithms based on the mutual information characteristics of part of the subjects, compare the classification results with those from the Pearson-based approach, and explore the possibility of developing MI as a new biomarker for AD diagnosis, especially at the early stage.

- In addition to the existing data set, future research can also be conducted on the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu/) data set, where more than 200 resting-state fMRI datasets have been collected from aged normal subjects, and MCI and AD patients. Some subjects have multiple time points, and the dataset size continues to grow. The ADNI dataset is freely available to researchers.

Understanding causality from a network perspective

- For causality analysis, traditionally, only two regions are considered at a time. Note that the dependence relationship between two time series may not be conserved when additional observations are taken into account. Further research is needed to study the causality between two brain regions while taking additional regions into consideration, and investigate the relationship between the extended discrete DCM model and multivariate DI theory.

Stability analysis of the brain network

- Brain network stability is essentially an uncharted area. While the stability on functional brain activity measures has been studied by examining the variations of various measures along the time and spatial domain [99], and the stability of brain network topology has been studied in [100], existing work on brain network stability has been very limited. It is important for the future research to explore how to characterize brain stability quantitatively, model the information processing at the region level, and evaluate network communication pattern changes in different groups of subjects.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] A. Tsai, I. Fisher, JohnW., C. Wible, I. Wells, WilliamM., J. Kim, and A. Willsky, "Analysis of functional MRI data using mutual information," in *Medical Image Computing and Computer-Assisted Intervention MICCAI99*, ser. Lecture Notes in Computer Science, C. Taylor and A. Colchester, Eds. Springer Berlin Heidelberg, 1999, vol. 1679, pp. 473–480.

[2] V. Michel, C. Damon, and B. Thirion, "Mutual information-based feature selection enhances fMRI brain activity classification," in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008.*, May 2008, pp. 592–595.

[3] B. Chai *et al.*, "Exploring functional connectivities of the human brain using multivariate information analysis," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 270–278.

[4] L. M. Parkes, J. V. Schwarzbach, A. A. Bouts, P. Pullens, C. M. Kerskens, D. G. Norris *et al.*, "Quantifying the spatial resolution of the gradient echo and spin echo bold response at 3 tesla," *Magnetic resonance in medicine*, vol. 54, no. 6, pp. 1465–1472, 2005.

[5] V. Gmez-Verdejo, M. Martnez-Ramn, J. Florensa-Vila, and A. Oliviero, "Analysis of fMRI time series with mutual information," *Medical Image Analysis*, vol. 16, no. 2, pp. 451 – 458, 2012.

[6] A. Roebroeck *et al.*, "Causal time series analysis of functional magnetic resonance imaging data," in *JMLR: Workshop and Conference Proceedings 12*, 2011, pp. 65–94.

[7] C.Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[8] K. J. Friston, "Functional and effective connectivity: a review," *Brain connectivity*, vol. 1, no. 1, pp. 13–36, 2011.

[9] M. Hu and H. Liang, "A copula approach to assessing granger causality," *NeuroImage*, vol. 100, pp. 125–134, 2014.

[10] P. Liang, Z. Li, G. Deshpande, Z. Wang, X. Hu, and K. Li, "Altered causal connectivity of resting state brain networks in amnesic mci," *PloS one*, vol. 9, no. 3, p. e88476, 2014.

[11] O. David *et al.*, "Identifying neural drivers with functional mri: An electrophysiological validation," *PLOS BIOLOGY*, vol. 6, pp. 2683–2697, Dec 2008.

[12] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. New York: Cambridge University Press, Sep 2009.

[13] E. E. Schadt *et al.*, "An integrative genomics approach to infer causal associations between gene expression and disease," *Nature Genetics*, vol. 37, pp. 710–717, July 2005.

[14] M. Luessi *et al.*, "Variational bayesian causal connectivity analysis for fMRI," *Frontiers in Neuroinformatics*, vol. 8, pp. 1–16, May 2014.

[15] K. Friston *et al.*, "Dynamic causal modeling," *NeuroImage*, vol. 19, pp. 1273–1302, 2003.

[16] K. Stephan *et al.*, "Dynamic causal models of neural system dynamics: current state and future extensions," *J. Biosci*, vol. 32, pp. 129–144, Jan 2007.

[17] S. Ryali *et al.*, "Multivariate dynamical systems models for estimating causal interactionsin fMRI," *NeuroImage*, vol. 54, pp. 807–823, 2011.

[18] R. Vicente *et al.*, "Transfer entropy – a model-free measure of effective connectivity for the neurosciences," *J Comput Neurosci*, vol. 30, pp. 45–67, 2011.

[19] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLOS Computational Biology*, vol. 2, pp. 1301–1312, Oct 2006.

[20] R. Vicente *et al.*, "Transfer entropya model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, pp. 45–67, Aug 2010.

[21] J. T. Lizier *et al.*, "Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity," *Journal of Computational Neuroscience*, vol. 30, pp. 85–107, Aug 2010.

[22] C. Quinn *et al.*, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *J Comput Neurosci*, vol. 30, pp. 17–44, Feb 2011.

[23] M. Wibral *et al.*, *Directed Information Measures in Neuroscience*. Springer, 2014.

[24] J. Massey, "Causality,feedback, and directed information," in *Int. Symp. Inf. Theory Appl.*, Honolulu, HI, Nov. 1990, pp. 303–305.

[25] H.H.Permuter *et al.*, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 3248–3259, Jun. 2009.

[26] N. Soltani and A. Goldsmith, "Inferring neural connectivity via measured delay in directed information estimates," in *IEEE Int Symp. on Inf. Theory*, July 2013, pp. 2503–2507.

[27] Z. Wang, A. Alahmadi, D. Zhu, and T. Li, "Causality analysis of fmri data based on the directed information theory framework," *Biomedical Engineering, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[28] Y. Liu and S. Aviyente, "Quantification of effective connectivity in the brain using a measure of directed information," *Computational and Mathematical Methods in Medicine*, vol. 2012, 2012.

[29] P.-O. Amblard and O. J. Michel, "On directed information theory and Granger Causality graphs," *J Comput Neurosci*, vol. 30, pp. 7–16, Feb 2010.

[30] J. Jiao *et al.*, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, 2014.

[31] A. Razi and K. J. Friston, "The connected brain: Causality, models, and intrinsic dynamics," *IEEE Signal Processing Magazine*, vol. 33, no. 3, pp. 14–35, 2016.

[32] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo *et al.*, "Dynamic functional connectivity: promise, issues, and interpretations," *Neuroimage*, vol. 80, pp. 360–378, 2013.

[33] G. K. Cooray, B. Sengupta, P. Douglas, M. Englund, R. Wickstrom, and K. Friston, "Characterising seizures in anti-nmda-receptor encephalitis with dynamic causal modelling," *NeuroImage*, vol. 118, pp. 508–519, 2015.

[34] G. K. Cooray, B. Sengupta, P. K. Douglas, and K. Friston, "Dynamic causal modelling of electrographic seizure activity using bayesian belief updating," *NeuroImage*, vol. 125, pp. 1142–1154, 2016.

[35] D. S. Bassett, M. Yang, N. F. Wymbs, and S. T. Grafton, "Learning-induced autonomy of sensorimotor systems," *Nature neuroscience*, vol. 18, no. 5, pp. 744–751, 2015.

[36] K. Wang *et al.*, "Discriminative analysis of early Alzheimers disease based on two intrinsically anti-correlated networks with resting-state fMRI," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pp. 340–347, 2006.

[37] N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar *et al.*, "Prediction of individual brain maturity using fmri," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.

[38] G. Bellucci, S. Chernyak, M. Hoffman, G. Deshpande, O. Dal Monte, K. M. Knutson, J. Grafman, and F. Krueger, "Effective connectivity of brain regions underlying third-party punishment: functional mri and granger causality evidence," *Social neuroscience*, vol. 12, no. 2, pp. 124–134, 2017.

[39] A. Duggento, G. Valenza, L. Passamonti, M. Guerrisi, R. Barbieri, and N. Toschi, "Reconstructing multivariate causal structure between functional brain networks through a laguerre-volterra based granger causality approach," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 5477–5480.

[40] X. Hu, S. Hu, J. Zhang, W. Kong, and Y. Cao, "A fatal drawback of the widely used granger causality in neuroscience," in *Information Science and Technology (ICIST), 2016 Sixth International Conference on*. IEEE, 2016, pp. 61–65.

[41] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[42] D. C. Zhu *et al.*, "Alzheimer's disease and amnestic mild cognitive impairment weaken connections within the default-mode network: a multi-modal imaging study," *Journal of Alzheimer's Disease*, vol. 34, no. 4, pp. 969–984, 2013.

[43] B. Afshin-Pour, H. Soltanian-Zadeh, G.-A. Hossein-Zadeh, C. L. Grady, and S. C. Strother, "A mutual information-based metric for evaluation of fMRI data-processing approaches," *Humman Brain Mapping*, vol. 32, pp. 699–715, 2011.

[44] Q. Wang *et al.*, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2392–2405, May 2009.

[45] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Miscellaneous Clustering Methods, in Cluster Analysis*, 5th ed. Chichester, UK: Wiley, 2011.

[46] J. S. Racine, "Nonparametric econometrics: A primer," *Foundations and Trends in Econometrics*, vol. 3, no. 1, pp. 1–88, 2008.

[47] T. Weise, *Global Optimization Algorithms – Theory and Application*. Germany: it-weise.de (self-published), 2009. [Online]. Available: http://www.it-weise.de/projects/book.pdf

[48] C. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer, 2005.

[49] 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.

[50] C. Bernasconi and P. Konig, "On the directionality of cortical interactions studied by structural analysis of electrophysiological recordings," *Biological Cybernetics*, vol. 81, pp. 199–210, 1999.

[51] R. Goebela *et al.*, "Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger Causality mapping," *Magnetic Resonance Imaging*, vol. 21, pp. 1251–1261, 2003.

[52] A. Roebroeck *et al.*, "Mapping directed influence over the brain using Granger Causality and fMRI," *NeuroImage*, vol. 258, pp. 230–2421, 2005.

[53] S. L. Bressler *et al.*, "Top-down control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention," *The Journal of Neuroscience*, vol. 28, pp. 10 056–10 061, 2008.

[54] X. Wen *et al.*, "Is Granger Causality a viable technique for analyzing fMRI data?" *PLOS ONE*, vol. 8, pp. 1–11, July 2013.

[55] ——, "Causal interactions in attention networks predict behavioral performance," *The Journal of Neuroscience*, vol. 32, pp. 1284–1292, Jan 2012.

[56] B. P. Bezruchko *et al.*, "Modeling nonlinear oscillatory systems and diagnostics of coupling between them using chaotic time series analysis: applications in neurophysiology," *Physics-Uspekhi*, vol. 51, no. 3, pp. 304–310, 2008.

[57] D. Marinazzo *et al.*, "Kernel method for nonlinear Granger Causality," *Physical Review Letters*, vol. 100, no. 14, p. 144103, 2008.

[58] S. Iyengar, J. Dauwels, P. Varshney, and A. Cichocki, "Quantifying eeg synchrony using copulas," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 505–508.

[59] D. Marinazzo, W. Liao, H. Chen, and S. Stramaglia, "Nonlinear connectivity by Granger Causality," *Neuroimage*, vol. 58, no. 2, pp. 330–338, 2011.

[60] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, 2003.

[61] J. Henderson *et al.*, "Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: An fMRI study," *Visual Cognition*, vol. 19, no. 7, pp. 910–927, 2011.

[62] R. Epstein and N. Kanwisher, "A cortical representation of the local visual environment," *Nature*, vol. 392, no. 6676, pp. 598–601, 1998.

[63] J. Jeong *et al.*, "Mutual information analysis of the eeg in patients with alzheimer's disease," *Clinical Neurophysiology*, vol. 112, pp. 827–835, 2001.

[64] L. Bettencourt *et al.*, "Functional structure of cortical neuronal networks grown in vitro," *Physical Review*, vol. 75, no. 2, pp. 021 915–1–021 915–10, 2007.

[65] D. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–10, 1979.

[66] F. Willemx *et al.*, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, May. 1995.

[67] J. Talairach and P. Tournoux, *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging.* New York: Georg Thieme Verlag., 1988.

[68] S. T. Witt *et al.*, "Functional neuroimaging correlates of finger-tapping task variations: an ale meta-analysis," *Neuroimage*, vol. 42, no. 1, pp. 343–356, 2008.

[69] D. C. Zhu and S. Majumdar, "Integration of resting-state fmri and diffusion-weighted mri connectivity analyses of the human brain: limitations and improvement," *J Neuroimaging*, vol. 24, no. 2, pp. 1763–186, 2014.

[70] R. W. Cox, "Afni: software for analysis and visualization of functional magnetic resonance neuroimages," *Comput Biomed Res*, vol. 29, no. 3, pp. 162–173, 1996.

[71] S. Smith, "Overview of fMRI analysis," *The British Journal of Radiology*, vol. 77, pp. S167–S175, 2004.

[72] K. Schreiber and B. Krekelberg, "The statistical analysis of multi-voxel patterns in functional imaging," *PLoS ONE*, vol. 8, no. 7, p. e69328, 07 2013.

[73] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with granger causality and dynamic causal modelling," *Current opinion in neurobiology*, vol. 23, no. 2, pp. 172–178, 2013.

[74] K. Friston, "Dynamic causal modeling and granger causality comments on: The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution," *Neuroimage*, vol. 58, no. 2, pp. 303–305, 2011.

[75] J. F. Smith *et al.*, "Identification and validation of effective connectivity networks in functional magnetic resonance imaging using switching linear dynamic systems," *Neuroimage*, vol. 52, pp. 1027–1040, 2010.

[76] E. Westman *et al.*, "Combining MRI and measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion," *Neuroimage*, vol. 62, no. 1, pp. 229–238, 2012.

[77] C. Lehmann *et al.*, "Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG)," *Journal of neuroscience methods*, vol. 161, no. 2, pp. 342–350, 2007.

[78] R. Cuingnet *et al.*, "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[79] G. Chen *et al.*, "Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging," *Radiology*, vol. 259, no. 1, pp. 213–221, 2011.

[80] W. S. Pritchard *et al.*, "EEG-based, neural-net predictive classification of Alzheimer's disease versus control subjects is augmented by non-linear EEG measures," *Electroencephalography and clinical Neurophysiology*, vol. 91, no. 2, pp. 118–130, 1994.

[81] R. O. Duda *et al.*, *Pattern classification*. John Wiley & Sons, 2012.

[82] B. Magnin *et al.*, "Support vector machine-based classification of Alzheimers disease from whole-brain anatomical MRI," *Neuroradiology*, vol. 51, no. 2, pp. 73–83, 2009.

[83] I. O. Korolev *et al.*, "Predicting progression from mild cognitive impairment to alzheimer's dementia using clinical, mri, and plasma biomarkers via probabilistic pattern classification," *PLoS ONE*, vol. 11, no. 2, pp. 1–25, 02 2016.

[84] B. Yeo *et al.*, "The organization of the human cerebran cortex estimated by intrinsic functional connection," *J Neurophysiol*, vol. 106, pp. 1125–1165, June 2011.

[85] M. A. Binnewijzend *et al.*, "Resting-state fmri changes in alzheimer's disease and mild cognitive impairment," *Neurobiology of aging*, vol. 33, no. 9, pp. 2018–2028, 2012.

[86] M. D. Greicius *et al.*, "Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 13, pp. 4637–4642, 2004.

[87] H.-Y. Zhang *et al.*, "Resting brain connectivity: Changes during the progress of alzheimer disease 1," *Radiology*, vol. 256, no. 2, pp. 598–606, 2010.

[88] B. Fischl and thers, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[89] J. Zhu *et al.*, "Multi-class AdaBoost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[90] R. Tibshirani *et al.*, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.

[91] J. Rice, *Mathematical statistics and data analysis*. Cengage Learning, 2006.

[92] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[93] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[94] H. Zhou *et al.*, "On projections of Gaussian distributions using maximum likelihood criteria," in *Information Theory and Applications Workshop, 2009.* IEEE, 2009, pp. 431–438.

[95] S. R. Searle, "Matrix algebra useful for statistics," *New York*, vol. 1982, 1982.

[96] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech communication*, vol. 26, no. 4, pp. 283–297, 1998.

[97] L. Breiman *et al.*, *Classification and regression trees.* CRC press, 1984.

[98] Z.-H. Zhou, *Ensemble methods: foundations and algorithms.* CRC Press, 2012.

[99] K. A. Garrison, D. Scheinost, E. S. Finn, X. Shen, and R. T. Constable, "The (in) stability of functional brain network measures across thresholds," *Neuroimage*, vol. 118, pp. 651–661, 2015.

[100] J. K. Rzucidlo, P. L. Roseman, P. J. Laurienti, and D. Dagenbach, "Stability of whole brain and regional network topology within and between resting and cognitive states," *PloS one*, vol. 8, no. 8, p. e70275, 2013.