Brain Connectivity Analysis using fMRI Data

by

Xuefei Cao

B.Sc., Fudan Unversity; Shanghai, China, 2015

M.Sc., Brown University; Providence, RI, 2018

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2020

This dissertation by Xuefei Cao is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Björn Sandstede, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Xi Luo, Ph.D., Reader

Date_____ _____

Matthew Harrison, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Andrew G. Campbell, Dean of the Graduate School

Vita

Education

- Brown University, Providence, RI, 2015 - 2020

  Ph.D. in Applied Mathematics, expected May 2020

  M.Sc. in Applied Mathematics, 2018

- Fudan University, Shanghai, China, 2011 - 2015

  B.S. in Mathematics and Applied Mathematics

Honors and Awards

- 2018, Honorable Mention of Student Paper Competition in Mental Health Statistics Section in JSM
- 2015, Outstanding Graduate (Fudan University)
- 2014, Honorable Prize of S.-T. Yau College Student Mathematics Contest of probability and statistics

Teaching Experience

- Teaching Assistant, Recent Applications of Probability and Statistics, Brown University, Spring 2017
- Teaching Assistant, Methods of Applied Mathematics I, II, Brown University, Fall 2016
- Teaching Assistant, Methods of Applied Mathematics I, II, Brown University, Fall 2017

Service and Professional Activities

- Organizer, Graduate Student Seminar, Brown University, Spring 2017 - Spring 2018
- Intern, Facebook Inc, Summer 2019
- Part-time Researcher, Facebook Inc, Fall 2019

Publications

- Cao X, Sandstede B, Luo X. *A Functional Data Method for Causal Dynamic Network Modeling of Task-Related fMRI*, Frontiers in neuroscience. 2019
- Cao X, Ke J, Sandstede B, Luo X. *Time-dependent Canonical Correlation Analysis for Multilevel Time Series*, bioxiv. 2019
- Cao X, Chen BC, Lim SN *Unsupervised Deep Metric Learning via Auxiliary Rotation Loss*, arxiv. 2019

Presentations

- Poster Presentation: *Sparse Causal Dynamic Network Modeling of fMRI* at JSM, Vancouver, July, 2018

Professional Associations

- American Mathematical Society (AMS)
- Association for Women in Mathematics (AWM)
- Society for Industrial and Applied Mathematics (SIAM)
- International Biometric Society (IBS)
- American Statistical Association (ASA)
- Eastern North American Region (ENAR)

Dedication

In memory of my grandfather Yongcheng Chen, whose love for me knew no bounds and, who encouraged me to pursue my own passion.

Acknowledgments

To my advisor Björn, thank you for the guidance, support, and mentorship throughout these years. His passion of learning new areas and great vision on scientific research in applied mathematics have inspired me to explore my research field. Aside from being an academic advisor in my research work, his endless support of my professional development has allowed me to grow as a researcher in the area I am passionate about. It is truly my honor to have Björn as my advisor and I am forever grateful for his continued guidance.

I would also like to thank Dr. Xi Luo and Dr. Matthew Harrison for taking the time to serve on my thesis committee and read my thesis. Also, to Dr. Xi Luo for collaborating and providing guidance and support for my research.

It has been my honor to know many people from the Division of Applied Math, Department of Mathematics and Department of Biostatistics at Brown. Thanks for their help and support in these past five years. To the many current and previous members of the Sandstede research group, who have been my colleagues and friends: thank you for sharing your endless knowledge and introducing me to learn different research areas.

Finally, I am eternally thankful for the love and support of my family and friends - especially that of my parents and my grandmother.

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter One

Introduction

1.1 Brain Connectivity Analysis

Neuroimaging has become a predominant technique in computational neuroscience [Friston, 2011]. Modern brain imaging techniques — such as diffusion MRI, functional MRI, Electroencephalography (EEG), and Magnetoencephalography (MEG) have shown increasingly larger datasets for studies on anatomical or functional brain connection patterns [Bandettini, 2009]. It has been a trend to obtain complementary information about neural connectivity and to promote a better understanding of the underlying brain mechanisms of different sensorimotor and cognitive tasks to be performed. Neurons or neural populations do not function alone. Rather, to achieve different tasks, they interact with other such populations. The human brain has about one hundred billion neurons, each establishing connections with several thousand other neurons in an intricate matrix [Rossini et al., 2019]. The brain is a complicated network that can be modeled by a set of nodes (anatomical/functional neuronal aggregates) and interconnecting edges (structural/functional connections). The brain's anatomical and functional organization is characterized by the segregation and integration of processed information. Functional interactions exist in both local regions and distant brain regions and thus brain networks consist of spatially distributed but functionally linked regions [Lang et al., 2012].

1.1.1 Functional Connectivity

Functional connectivity (FC) is defined as the functionally integrated relationship (i.e., the similar patterns of activation) between spatially separated brain regions. Functional connectivity can be quantified with measures of temporal dependency (e.g. correlations, coherence, or transfer entropy) of neuronal activation patterns

from anatomically separated brain regions [Van Den Heuvel and Pol, 2010]. Studying functional connectivity provides new insights on how the brain communicates and interacts. A common practice [Li et al., 2009] is to utilize resting-state fMRI to measure functional connectivity in BOLD (blood oxygenation level dependent) time-series of activations in distinct brain regions.

In general, some popular methods [Dennis and Thompson, 2014] to extract brain signals and assess functional connectivity (using the BOLD signal) include the seed-based approach [Joel et al., 2011], principal component analysis (PCA) [Zhong et al., 2009], independent components analysis (ICA) [Joel et al., 2011], and graph theory [Wang et al., 2010]. In the seed-based approach, researchers define a seed (brain region) by a priori knowledge or from a traditional task-dependent activation map acquired in a separate fMRI experiment [Van Den Heuvel and Pol, 2010]. The time series of the chosen seed is tested for correlations with the rest of the brain regions. The simplicity of this method gives it a strong advantage over other methods. However, the information of a seed-based method is limited to the functional connections of a predefined region and it is difficult to operate it on the whole-brain scale. Thus PCA or ICA methods are applied to search for underlying components of latent factors or sources that are orthogonal or statistically independent of each other. Graph-based brain network analyses of fMRI signals introduce a set of new metrics including path length, modularity, nodal centrality, which may be calculated to describe global and local brain network structure [Wang et al., 2010].

1.1.2 Dynamic Functional Connectivity

Functional connectivity has attracted a lot of attention in the last two decades. However, the dynamic behavior of functional connectivity was discovered recently, and

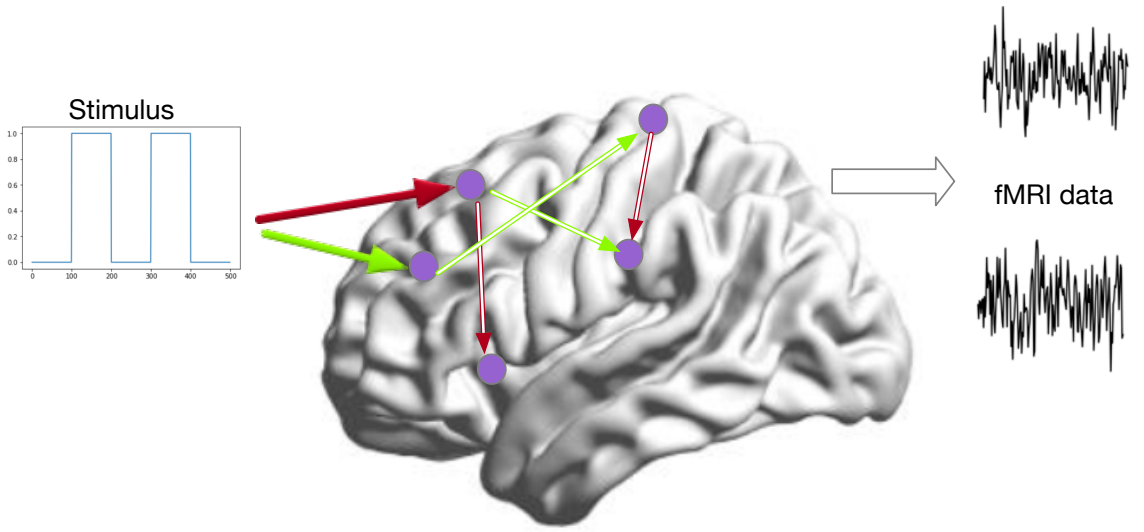

Figure 1.1: The BOLD response of multiple brain regions due to external stimuli. The green (red) arrows indicate positive (negative) causal connections.

it has been shown that connectivity between distinct brain regions involves meaningful variations in resting-state fMRI data [Chang and Glover, 2010, Preti et al., 2017]. Dynamic functional connectivity (dFC) was then proposed. Dynamic functional connectivity refers to the observed phenomenon that functional connectivity changes over a short time. One common strategy for dynamic functional connectivity study is to explore the variations of pairwise connectivity from different brain regions using a set of temporal sliding windows [Sakoğlu et al., 2010] (so-called sliding window method). This basic sliding window approach has been extensively studied and applied in neuroimaging community [Leonardi and Van De Ville, 2015, Zalesky and Breakspear, 2015]. Besides its simplicity, the sliding window technique has some major limitations. For example, it is not clear how the window length should be chosen. A window with too short length may introduce artificial fluctuations while too long windows would impede the detection of the temporal variations of interest [Preti et al., 2017].

1.1.3 Effective Connectivity

Effective connectivity is defined as the causal influences that neural units exert over another Friston [2011]. Studies of effective connectivity focus on explaining the observed dependencies (functional connectivity) and estimating parameters of proposed causal models [Friston, 2011]. For instance, one can study the causal relation between several different regions of subjects under a motor task. See more examples in our application section 5.6. In other words, effective connectivity investigates coupling or directed causal influence within different brain regions. Figure 1.1 is a simple illustration of the process. Some seminal works include [Friston et al., 2003, Marreiros et al., 2008, Stephan et al., 2010, 2008]. It worth noting that the analysis of effective connectivity can, in general, be reduced to model comparison, for example, the comparison of a model with different network structures.

1.1.4 Connectivity and Covariates

The variations in functional connectivity have been found to be associated with different factors including age and cognitive behaviors (so-called covariates) [Yang et al., 2015, Hull et al., 2017, Badhwar et al., 2017, Zonneveld et al., 2019] . It has been an important research area to study the relationship between functional connectivity (correlation matrices) and different subjects' auxiliary information. One common approach analyzing this relationship is to do the regression of each correlation coefficient on the external subjects' information. Although this method is easy to implement and interpret, it may introduce multiplicity issues and lose the global network information provided by the correlation matrices [Zhao et al., 2018]. Another direction is to model these matrices based on the external covariates [Hoff and

Niu, 2012]. However, estimating these matrices directly is computationally intensive and usually requires a large number of samples. Recently, Zhao et al. instead aims to find the projections for covariance matrices that are associated with the covariates, where the covariance matrix is not estimated.

1.2 Outlines of Thesis

This thesis is organized as follows.

To study and estimate effective brain connectivity more efficiently, section 2 introduces a causal dynamic network (CDN) method to estimate brain activations and connections simultaneously. Our method links the observed fMRI data with the latent neuronal states modeled by an ordinary differential equation (ODE) model. In section 3, we extend CDN and introduce a new optimization problem and an efficient iterative algorithm that uses a data-driven method to search from a huge number of ODE models. Different from previous approaches on effective brain connectivity 1.1.3, our method does not require hypothesizing the connections in advance.

To better analyze the coupling temporal dynamics between the two time-varying data sources, we extend canonical correlation analysis and explore the time-dependent canonical correlation analysis (TDCCA) in section 4. It is a method for inferring time-dependent canonical vectors from multilevel time series data and we apply this method to fMRI data to study the dynamic functional brain connectivity. We are interested in the application of canonical correlation on dynamic functional brain connectivity study, while previous works focus on analysis of Pearson correlation.

In section 5, we focus on analyzing the relationship between functional connectiv-

ity and subject-level covariates and investigate a Common Vector Projection Regression (CVPR) model for multiple covariance matrix outcomes. Instead of focusing on estimating covariances directly in most of the existing works, we are interested in linking covariances and subject-level covariates or explainable variables.

Part II

Causal Dynamic Network modeling of fMRI

Chapter Two

A Functional Data Method for Causal Dynamic Network Modeling

2.1 Introduction

In recent years, functional magnetic resonance imaging (fMRI) has become a major tool to investigate dynamic brain networks. Earlier analysis methods focus on inferring brain regions activated by external experimental stimuli, for example using a general linear model approach [Friston et al., 1994]. More recently, it becomes important to model the inter-connections between brain regions, sometimes called connectivity analysis of fMRI, see for example a review [Smith, 2012].

Generally speaking, there are two types of connectivity modeling: functional connectivity and effective connectivity. Functional connectivity usually models the correlations or dependencies between multiple BOLD (blood-oxygen-level dependent) time series from multiple brain regions. It can be estimated using generic statistical methods, such as correlations, partial correlations [Marrelec et al., 2006], regularized inverse covariance [Banerjee et al., 2006]. Effective connectivity, on the other hand, aims to model the neuronal or causal connections between brain regions. Some effective methods can also model the effects of external experimental stimuli. Three major effective modeling approaches are dynamic causal modeling [Friston et al., 2003], structural equation modeling [McIntosh and Gonzalez-Lima, 1994], and Granger causality analysis (GCA) [Goebel et al., 2003, Roebroeck et al., 2005, Deshpande et al., 2008, Seth, 2010]. In particular, Granger causality utilizes a data-driven, "time-lagged" prediction criterion to determine the directionality of connections. Though GCA is usually employed to analyze BOLD signals, recent extensions [David et al., 2008, Wheelock et al., 2014, Grant et al., 2015] enable applications of GCA in the latent neuronal space. A closely related extension of GCA in the latent space is state-space multivariate dynamical systems (MDS) [Ryali et al., 2011, 2016a,b]. Dynamic causal modeling (DCM), being the more sophisticated mod-

eling approach, utilizes ordinary differential equation (ODE) models for the neuronal dynamics and hemodynamic response, and it is thus an important tool for understanding the biophysical (or neuronal) connections in the brain. However, it often requires prior knowledge or hypotheses of network connections and stimulus projections to narrow down the model space, and then it resorts to a computationally expensive approach to select the "best" fit model after computing all the candidate models. As the number of brain regions under investigation increases, the number of candidate models which need to be computed increases dramatically. Thus the computational burden of fitting many candidate DCM limits the number of regions (in the order of 10 regions) that can be considered. Therefore, DCM is considered to be a hypothesis-validation approach rather than a data-driven approach [Stephan et al., 2010]. Despite this limitation, DCM [Friston et al., 2003] has become a standard approach to investigate the brain mechanisms using fMRI, and has been validated in numerous studies [Penny et al., 2004, Lee et al., 2006, David et al., 2008, Dima et al., 2010, Reyt et al., 2010, Brodersen et al., 2011, Frässle et al., 2015] ,

Most recently, some progress has been made to refine and extend the standard DCM [Friston et al., 2003] described before. Several methods are developed to improve the biophysical interpretation of the DCM neuronal model, including: (1) two-state DCM [Marreiros et al., 2008] that comprises of excitatory and inhibitory neuronal populations in each brain region/node; (2) nonlinear DCM [Stephan et al., 2008] that accounts for nonlinear interactions of neuronal states between nodes; (3) stochastic DCM [Friston et al., 2011, Li et al., 2011] that allows neuronal fluctuations; (4) canonical microcircuit DCM [Friston et al., 2017] that models four neuronal layers per node. These DCMs bring the neuronal mass modeling closer to the underlying biophysical processes. Extensive validation of these DCMs is less developed, compared with the standard DCM. These extended DCMs also introduce more model

parameters which can be a computational burden for model inversion, especially for medium or large networks. These issues are mitigated for analyzing resting-state fMRI. Spectral DCM [Friston et al., 2014, Razi et al., 2015] reduces the computation burden using a resting-state formulation of a linear DCM (without the bilinear term in the standard DCM) in the frequency domain, which essentially avoids computing the time-variant parameters. This approach allows inverting larger DCM networks with dozens of nodes [Razi et al., 2017]. For task-related fMRI, a linear DCM in the frequency domain is also inverted efficiently using regression DCM [Frässle et al., 2017], and thus it becomes feasible to compute for hundreds of regions [Frässle et al., 2018].

In the statistical literature, inferring ODE parameters from data has been studied separately for general settings, sometimes called dynamic data analysis (DDA) or functional data analysis of ODE models, mostly for the following observation model

$$\boldsymbol{y}(t_i) = \boldsymbol{x}(t_i) + \boldsymbol{\epsilon}(t_i) \tag{2.1}$$

where the observed data $\boldsymbol{y}(t_i)$ equals the noise $\boldsymbol{\epsilon}(t_i)$ plus the underlying latent signal $\boldsymbol{x}(t_i)$ generated from a set of ODEs. $\boldsymbol{y}(t_i)$ is usually sampled at equally spaced time points t_i, $i = 1, \ldots, n$. There exist several methods to infer the ODE parameters of interest, including the nonlinear least-square (NLS) method [Bard, 1974, Domselaar and Hemker, 1975, Xue et al., 2010], the two-stage smoothing-based estimation method [Varah, 1982, Brunel et al., 2008, Gugushvili et al., 2012, Brunel et al., 2014], the principal differential analysis and iterated principal differential analysis [Ramsay, 1996, Poyton et al., 2006, Zhang et al., 2015], the generalized profiling method [Cao and Ramsay, 2009, Qi and Zhao, 2010], the Bayesian approaches [Girolami, 2008, Bhaumik and Ghosal, 2015, Chkrebtii et al., 2016]. This observational model used in these methods, however, can only hold for specific neuroimaging techniques,

such as electrocorticographic [Zhang et al., 2015]. Unfortunately, all these existing DDA methods under this observational model have limited applications in fMRI, because the observed BOLD data depend on the current and past neuronal states, here $\boldsymbol{x}(t_i)$. Another complication for applying these DDA methods is that fMRI data are measured on a different time scale than the neuronal states with moderate or small signal-to-noise ratios.

One widely used yet simple model for connecting the neuronal states and BOLD responses is the linear convolution model,

$$\boldsymbol{y}(t) = \int h(s)\boldsymbol{x}(t-s)ds + \boldsymbol{\epsilon}(t) \tag{2.2}$$

where $\boldsymbol{y}(t_i)$ is a noise contaminated convolution of $\boldsymbol{x}(t_i)$ and a hemodynamic response function (HRF). Similar convolution formulations were employed before to model the neuronal states [Karahanoğlu et al., 2013, Ryali et al., 2011]. Different forms of convolution functions have been proposed in the literature, see for example [Friston et al., 1998] and [Lindquist et al., 2009]. Nonetheless, this general HRF convolution approach, adopted by many experimental studies, clearly suggests that the DDA observation model (2.1) is not valid for fMRI. Therefore, these existing statistical methods for ODE modeling cannot be applied directly.

In this chapter, we introduce a new statistical modeling framework for estimating effective connectivity and activations simultaneously from fMRI data. We propose a Causal Dynamic Network (CDN) method using a functional/dynamic data analysis approach, which jointly models the neuronal states modeled by a DCM-type ODE model and the observed BOLD responses. Unlike DCM or its extensions (e.g, Marreiros et al. [2008], Friston et al. [2014], Razi et al. [2017]), our optimization-based method and algorithm compute the ODE parameters efficiently in a data-driven

fashion, instead of comparing potentially a huge number of candidate ODE models. This approach to consider a more general convolution-based observation model also adds to the vast DDA literature. To the best of our knowledge, this type of data-driven ODE modeling for fMRI has not been considered before, and no existing DDA methods can deal with a convolution observation model. We aim to address both challenges in this chapter.

The remainder of the chapter is organized as follows. We start with a brief review of DCM and discuss its computational issues. We then introduce our CDN model and estimation method in section 2.2.2. Section 3.3 illustrates the numerical performance of our model using extensive simulations, including simulations from both CDN and DCM. A real fMRI data analysis is presented in section 2.3.2.

2.2 Methods

2.2.1 DCM Revisited

Dynamic causal modeling (DCM) was introduced by Friston et al. [2003]. DCM provides a general framework to infer the causal activations and connections in brain networks under experimental stimuli. There are two components in DCM. The first component is a set of ODEs characterizing the variations of neuronal responses based on stimuli as

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{u}(t), \boldsymbol{x}(t), \boldsymbol{\theta}) \tag{2.3}$$

where $\boldsymbol{x}$ represents the neuronal states of d regions, $\boldsymbol{u}$ is the input from J stimuli, and $\boldsymbol{\theta}$ is the ODE parameter of interest. It is usually difficult to estimate $\boldsymbol{f}$ in a

general form, and instead, DCM employs a bilinear approximation for this dynamic system as follows

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + \sum_{j=1}^{J} u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t). \tag{2.4}$$

In this approximation, the entry A_{mn} in $\boldsymbol{A} = (A_{mn})_{dd}$ denotes the strength of intrinsic causal connection from the n-th region to m-th region. $\boldsymbol{B}_j = (B_{mnj})_{dd}$ denotes the influence of the j-th input stimulus $u_j(t)$ on the directional connection between these two regions. $\boldsymbol{C} = (C_{mj})_{dJ}$ denotes the effects of u_j on the m-th region. The parameters $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ have important biophysical interpretations. $\boldsymbol{A}$ models the intrinsic effective connectivity between brain regions, $\boldsymbol{B}$ models how these connections are influenced by external stimuli, and $\boldsymbol{C}$ models how brain regions are activated (or suppressed) by stimuli. The second part of DCM involves a set of hemodynamic state equations, motivated by the Balloon-Windkessel model [Buxton et al., 1998]. These differential equations describe how neuronal activity induces vasodilatory signals that lead to changes in blood volume and deoxyhemoglobin content. These biophysical changes result, in a non-linear form, the fMRI BOLD measures. The parameters in this set of DCM equations are not usually used or interpreted.

To estimate these parameters, DCM in the first stage employs the expectation maximization (EM) procedure to fit a candidate model, usually with specific zero and nonzero patterns in $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ provided by users. In the second stage, from a list of user-specified models, DCM uses the Bayes factor to select the "best" model for a given dataset or experiment. Because of the expensive Bayesian computations in the first stage, users usually need to restrict the number of candidate models and the complexity of such models (e.g. by restricting the number of brain regions considered in the model) [Penny et al., 2010]. Clearly, this DCM approach cannot be made data-driven without incurring a huge computational cost, because the number

of all possible models grows at least with the order $O(2^{d^2})$, where d is the number of brain regions.

2.2.2 Causal Dynamic Networks

2.2.2.1 Model

We propose the following CDN model

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + \sum_j u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t) \tag{2.5}$$

$$\boldsymbol{y}(t) = \int h(s)\boldsymbol{x}(t-s)ds + \boldsymbol{\epsilon}(t) \tag{2.6}$$

where $\boldsymbol{y}(t)$ is a $d \times 1$ vector of BOLD signals at time t of d regions, $\boldsymbol{\epsilon}(t)$ is the error process, $h(s)$ is a hemodynamic response function. The integration for vectors above should be understood as component-wise. Like DCM, each component of $\boldsymbol{y}(t)$ for a brain region depends only on its corresponding neuronal component in $\boldsymbol{x}(t)$ for the same region. Here, equations (2.5) and (2.6) are fitted separately to each scan session of each participant, and thus the model parameters should also be interpreted as for one session of one participant. Section 2.2.2.5 describes the approaches to combine these parameter estimates across sessions and participants into group-level (or population-level) estimates.

Equation (2.5) is the same as the neuronal state model in the standard DCM [Friston et al., 2003]. It uses the same bilinear approximation, and thus the interpretations of the key parameters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ are comparable to the standard DCM. Our goal is to infer the entries in $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ and their significance levels in this chapter.

It is important to note that our estimation approach does not need to prespecify the zero/nonzero patterns in these parameters. In this chapter, we will focus on this standard DCM. Several extensions of this DCM exist to account for more complex biophysical processes, for example [Marreiros et al., 2008, Stephan et al., 2008, Li et al., 2011, Friston et al., 2011, 2017]. Without parameter reduction, these extensions can lead to increased computational costs. To reduce the computation burden, [Frässle et al., 2017, 2018] employ a linear DCM (by setting the bilinear coefficient $\boldsymbol{B} = \mathbf{0}$), and compute only the frequency spectra in order to avoid the estimation of $\boldsymbol{x}(t)$ in the time domain.

The neuronal state model (2.5), originally developed for task-related fMRI, is sometimes referred to as *deterministic* DCM. An extension of this model, called *stochastic* DCM, includes an additive term representing endogenous neuronal fluctuations. Stochastic DCMs have been mostly used for modeling resting-state fMRI [Friston et al., 2011, Li et al., 2011] , under the following simplified form

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{\omega}(t) \tag{2.7}$$

where $\boldsymbol{\omega}(t)$ represents neuronal fluctuations (and possibly external stimulus input). Without any assumptions for $\boldsymbol{\omega}(t)$, there exists a trivial and less meaningful solution in which $\boldsymbol{\omega}(t)$ is simply the temporal derivative of $\boldsymbol{x}(t)$, shown by setting $\boldsymbol{A} = \mathbf{0}$. Under the smoothness assumption, as pointed out by [Friston et al., 2011, Daunizeau et al., 2012], there exists a mathematical connection between deterministic and stochastic DCMs. One can replace $\boldsymbol{\omega}(t) = \boldsymbol{F}\boldsymbol{\phi}(t)$ by a linear combination of a reasonable number of (cosine) basis functions $\boldsymbol{\phi}(t)$, and thus the random fluctuations in stochastic DCM are absorbed by the modeling coefficient $\boldsymbol{F}$ (similar to $\boldsymbol{C}$) in a deterministic DCM. In practice, however, this deterministic DCM can yield poor estimation [Daunizeau et al., 2012], and inverting (2.7) directly is computationally

more expensive. Because $\boldsymbol{x}(t)$ in resting-state fMRI does not have immediate interpretation, more recent methods usually transform this equation to the spectral domain [Friston et al., 2014, Razi et al., 2017]. This transformation essentially avoids estimating $\boldsymbol{x}(t)$ (and its associated parameters) to reduce computational costs. Our method introduced later will estimate $\boldsymbol{x}(t)$ because it can be interpreted along with the stimulus sequences in task-related fMRI. It is possible to apply our model to resting-state fMRI by setting $\boldsymbol{B}$ and $\boldsymbol{C}$ to zero. This simplification does not account for resting-state noise and requires estimating $\boldsymbol{x}(t)$. Despite these concerns, this initial work will provide preliminary validations of our method in resting-state fMRI, without much modification of our algorithm and incurring excessive computational costs. Comprehensive modeling of resting-state fMRI is beyond the scope of this work, and will be discussed as future work in Section 3.4.

Equation (2.6) is the relationship between neuronal states and fMRI BOLD signals where h is a hemodynamic response function. A similar formulation is used in [Karahanoğlu et al., 2013]. This is an approximation for the Balloon ODE model in DCM [Henson and Friston, 2007], and it is used here to avoid the heavy computation in fitting ODEs of less importance. Our method allows any hemodynamic response functions to be used. For simplicity, we employ the widely used canonical hemodynamic response function (HRF) [Friston et al., 1998] in this work. The canonical HRF is mathematically expressed as $\gamma(16t; 6, 1/6) - \gamma(16t; 16, 1/16)/6$, where the gamma density function $\gamma(t; \eta, \upsilon) = \upsilon^{\eta} t^{\eta-1} \exp(-\upsilon t)/\Gamma(\eta)$, time t is in seconds, and $\Gamma(\eta)$ is the gamma function.

Together, equation (2.5) and (2.6) can be interpreted as a latent space model in statistics. Our use of a two-equation model is similar to MDS [Ryali et al., 2011, 2016a,b]. However, there is an important difference. The first equation in MDS is for discrete latent signals because it is considered as an extension to GCA, whereas

ours is an ODE model for continuous neuronal states. Thus, these two models are fundamentally different in modeling the neuronal states, and thus the model parameters are not equivalent.

FMRI machines usually sample $\boldsymbol{y}(t)$ at equally spaced time intervals (usually 1–2 seconds): t_1, t_2, ..., t_i, ..., t_T. When fitting our CDN model to data, we replace Equation (2.6) in our CDN model with the following observation equation, for $i = 1, \ldots, T$,

$$\begin{aligned} \boldsymbol{y}(t_i) &= \int h(s)\boldsymbol{x}(t_i - s)ds + \boldsymbol{\epsilon}(t_i) \\ &= h \star \boldsymbol{x}(t_i) + \boldsymbol{\epsilon}(t_i) \end{aligned} \tag{2.8}$$

where $\star$ denotes the convolution operation by the integral above. This observation model is different from the popular model (2.1) in the statistical literature. This difference warrants a new estimation approach. For example, under the classic observation model (2.1), a general strategy in many existing methods is to approximate $\boldsymbol{x}(t)$ from smoothing $\boldsymbol{y}(t_i)$ before estimating the ODE parameters. However, this is clearly not applicable to our observation model or fMRI data. As a separate note, although functional data analysis models (see review [Wang et al., 2016]), especially functional convolution model [Asencio et al., 2014], may appear in a similar form as Equation (2.8), their estimation goal is fundamentally different from ours. They assume $\boldsymbol{y}$ and $\boldsymbol{x}$ are observed in a regression-type setting, mostly with no intention to estimate the ODE parameters associated with $\boldsymbol{x}$. Here we aim to estimate latent $\boldsymbol{x}(t)$ and more importantly the ODE parameters that generate it. Due to these many differences from existing models, we develop a new estimation approach in the next section.

2.2.2.2 Estimation

To estimate $\boldsymbol{x}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, we propose to minimize the following loss

$$l(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{t_i} \|\boldsymbol{y}(t_i) - h \star \boldsymbol{x}(t_i))\|^2 + \lambda \int \left\| \frac{d\boldsymbol{x}(t)}{dt} - (A\boldsymbol{x}(t) + \sum_j u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t)) \right\|^2 dt$$

where $\boldsymbol{\theta} = [\boldsymbol{A} \quad \boldsymbol{B} \quad \boldsymbol{C}]$, $\lambda > 0$ is a tuning parameter, and $\|\cdot\|$ is the ℓ_2 norm. The first part of the loss function is the data fitting error and the second part is regarded as the fidelity to our dynamic system. The ℓ_2 norm loss was also employed before in [Karahanoğlu et al., 2013], with the same goal to deconvolute the HRF.

Following a functional data analysis technique, we represent $\boldsymbol{x}(t)$ using truncated basis function expansions. Let $\boldsymbol{x}(t) = \boldsymbol{\Gamma}\boldsymbol{\Phi}(t)$ where $\boldsymbol{\Phi}(t)$ is a $p \times 1$ vector with entries denoting the basis function value at t and p is the number of basis functions. We choose p to be reasonably large to avoid modeling bias. We will use the classical cubic spline basis to model the neuronal state responses to smooth stimuli and the piece-wise linear basis for box-car stimuli. Certainly many other choices are possible here. The derivative $d\boldsymbol{x}(t)/dt$, integrations, solutions for ODE are approximated by numerical methods (e.g. fourth order Runge-Kutta), following the standard idea of solving ODEs numerically as also suggested in the textbook [Ramsay, 2006].

With the basis representation, our loss function becomes

$$l(\boldsymbol{\Gamma}, \boldsymbol{\theta}) = \sum_{t_i} \|\boldsymbol{y}(t_i) - \boldsymbol{\Gamma}[h \star \boldsymbol{\Phi}(t_i)]\|^2 + \lambda \int \left\| \boldsymbol{\Gamma}\boldsymbol{\Phi}'(t) - (\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\Phi}(t) + \sum_j u_j(t)\boldsymbol{B_j}\boldsymbol{\Gamma}\boldsymbol{\Phi}(t) + \boldsymbol{C}\boldsymbol{u}(t)) \right\|^2 dt.$$

The number of parameters in this loss function is $dp + d^2 + Jd^2 + dJ$, and the summands are due to estimating $\boldsymbol{x}(t)$, $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ respectively . Here, we will sacrifice a little mathematical rigor for simplified analysis of the estimation equation. The parameters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ play similar roles as regression coefficients, and thus they are less identifiable if the predictors in the design matrix $(\boldsymbol{x}(t), u_1\boldsymbol{x}(t), \ldots, u_J\boldsymbol{x}(t), \boldsymbol{u}(t))$ exhibit higher colinearity. The colinearity can increase as the network size increases, because the number of predictors included in the model grows quadratically with the predominant order $(J+1)d^2$, far exceeding the fMRI data size growing linearly with d. This is especially problematic for estimating $\boldsymbol{B}$, which could be understood as an interaction term in regression. The predictors associated with $\boldsymbol{B}$ are products of those ones with $\boldsymbol{A}$ and $\boldsymbol{C}$, and thus they can be highly correlated with them. Consider a simple box-car task desgin as an example, where $u_j(t) = 1$ during the jth task and zero otherwise. The colinearity is affected by the task duration. The reason is that the $\boldsymbol{B}$ predictors $(u_j(t)\boldsymbol{x}(t))$ equal to the corresponding $\boldsymbol{A}$ predictors $(\boldsymbol{x}(t))$ within the task window and zero otherwise. Thus, extremely long task duration will make these predictors highly correlated overall because they are identical most of the time. Similarly, $u_j(t)\boldsymbol{x}(t)$ can also be highly correlated with $u_j(t)$ if the duration is extremely short and $\boldsymbol{x}(t)$ has small variations within the task window.

Though the loss function $l(\boldsymbol{\Gamma}, \boldsymbol{\theta})$ is not jointly convex for $(\boldsymbol{\Gamma}, \boldsymbol{\theta})$, it is easy to see

that it is convex in $\boldsymbol{\Gamma}$ given $\boldsymbol{\theta}$ and vice versa. We thus propose to minimize the loss function using iterative block-wise updates for $\boldsymbol{\Gamma}$ and $\boldsymbol{\theta}$ respectively, summarized in Algorithm 1. Importantly, the iterative updates are given by the explicit form. This strategy is essentially an application of the alternate convex search algorithm for biconvex optimization, see the review [Gorski et al., 2007].

Algorithm 1 Estimation for CDN

Input: BOLD signal $\boldsymbol{y}$, stimulus $\boldsymbol{u}$, basis function $\boldsymbol{\Phi}$, tuning parameter λ
Initialize $\boldsymbol{\Gamma}$, $\boldsymbol{\theta}$
repeat
  Update $\boldsymbol{\Gamma}$ of $l(\boldsymbol{\Gamma}, \boldsymbol{\theta})$ given $\boldsymbol{\theta}$ by gradient descent with backtracking line search
  Solve for the minimizer $\boldsymbol{\theta}$ of $l(\boldsymbol{\Gamma}, \boldsymbol{\theta})$ given $\boldsymbol{\Gamma}$, see Section 2.2.2.3
until convergence
Return $\boldsymbol{\Gamma}$, $\boldsymbol{\theta}$

2.2.2.3 Parameter Updates

We derive the parameter updates in our algorithm. We introduce the following notations in order to illustrate our updating procedure. $\Phi = (\phi_1, ..., \phi_p)$ is our selected basis for estimating neuronal activity. Let

$$P_1[n,m] = \int_{t_0}^{t_T} \phi_n(t)\phi'_m(t)dt, \qquad P_2^j[n,m] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)\phi'_m(t)dt,$$

$$P_3[j,n] = \int_{t_0}^{t_T} u_j(t)\phi'_m(t)dt, \qquad P_4[n,m] = \int_{t_0}^{t_T} \phi_n(t)\phi_m(t)dt,$$

$$P_5^j[n,m] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)\phi_m(t)dt, \qquad P_6[j,n] = \int_{t_0}^{t_T} u_j(t)\phi_n(t)dt,$$

$$P_7^{jk}[n,m] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)\phi_m(t)u_k(t)dt, \qquad P_8^j[l,n] = \int_{t_0}^{t_T} u_l(t)\phi_n(t)u_j(t)dt,$$

$$P_9[n,m] = \int_{t_0}^{t_T} u_n(t)u_m(t)dt.$$

Combing these into matrix form, we denote

$$\begin{aligned}
\theta &= [A, B_1, B_2, .., B_J, C], \\
\Psi_A &= [\Gamma P_4 \Gamma', \Gamma P_5^1 \Gamma', ..., \Gamma P_5^J \Gamma', \Gamma P_6']', \\
\Xi_j &= [\Gamma P_5^j \Gamma', \Gamma P_7^{1j} \Gamma', ..., \Gamma P_7^{Jj} \Gamma', \Gamma (P_8^j)']', \\
\Psi_C &= [P_6 \Gamma', P_8^1 \Gamma', ..., P_8^J \Gamma', P_9]', \\
\mathrm{Y} &= [\Gamma P_1' \Gamma', \Gamma P_2^1 \Gamma', ..., \Gamma P_2^J \Gamma', \Gamma P_3'].
\end{aligned}$$

Setting the gradient with respect to θ to zero yields

$$\theta([\Psi_A, \Xi_1, ..., \Xi_J, \Psi_C]) = Y \tag{2.9}$$

and thus the update for θ is given by

$$\theta = Y([\Psi_A, \Xi_1, ..., \Xi_J, \Psi_C])^{-1} \tag{2.10}$$

One can derive the update for Γ by taking gradient.

2.2.2.4 Tuning Parameter Selection

To select the tuning parameter λ for our model, we use a cross-validation (CV) procedure as follows. Given two time series $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ from the same participant, generated by the same dynamic system (with potentially different inputs), we compute our estimator $\hat{\boldsymbol{\theta}}_\lambda$ by applying Algorithm 1 to $\boldsymbol{y}_1$, using λ on a grid. Based on $\hat{\boldsymbol{\theta}}_\lambda$, we generate $\hat{\boldsymbol{x}}_2$ using the ODE with varying initial conditions that yield the cross validation (CV) loss, defined as the smallest $\|\boldsymbol{y}_2 - h \star \hat{\boldsymbol{x}}_2(t)\|^2$. We then select the λ value that minimizes the CV loss. One can also consider varying p in this

procedure to minimize the CV loss, but we find that our approach is not sensitive for a reasonably large p. For example, we use $p \geq 50$ in our numerical studies. In principle, this CV procedure can be applied to each participant's data to select a subject-specific tuning parameter λ. In this work, for simplicity, we select one λ that minimizes the average CV loss across all participants.

2.2.2.5 Group-level Estimation

Based on the fitted CDN parameters $(\boldsymbol{A}_{kq}, \boldsymbol{B}_{kq}, \boldsymbol{C}_{kq})$ from session k of participant q, we in this chapter average the parameters across sessions and participant to yield a group-level estimate for $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. We use this simple averaging approach here because our simulated and real datasets contain a small fixed number (sometimes, one) of sessions from a relatively large number of participants. Based on our estimates, various alternative approaches can also be applied to obtain group-level inference. For example, when the within-participant dependence/cross-participant variability is a concern or the number of sessions for each participant varies a lot, one may also apply a mixed effects model to $(\boldsymbol{A}_{kq}, \boldsymbol{B}_{kq}, \boldsymbol{C}_{kq})$ and obtain the population estimate. This mixed effects approach was used before for the group level (or second stage) analysis in task activation studies [Worsley et al., 2002].

2.2.2.6 Inference and P-values

Because our model can be computed very efficiently (usually in seconds or a few minutes), we propose to use bootstrap (across subjects) method to assess the significance of each entry in $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. The resulting p-values can be used to pick the final model with only statistically significant entries depending on users' choice

of significance levels, without resorting to expensive model selection in DCM. For addressing the multiple comparison issue, popular p-value correction methods, such as Bonferroni or false discovery rate (FDR) [Benjamini and Hochberg, 1995] correction, should be used to select an appropriate threshold. In this chapter, we will use FDR to correct separately the p-values from each matrix $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and choose a conservative threshold of 0.01 to further control the overall false discovery rates.

2.3 Results

2.3.1 Simulations

In this section, we evaluate the numerical performance of our CDN method using two types of simulation models: DCM and our statistical model CDN. The former provides a more realistic fMRI data generating model while it is computationally more expensive. The latter is computationally inexpensive and will serve to validate the statistical properties of our algorithm. Whenever possible, we compare with three other methods: GCA, MDS, and DCM.

We use the Python Neuroimaging toolbox (Nitime) for the GCA method, and the generative DCM model from Smith et al. [2011]. We fit the DCM model using SPM12. The MDS implementation is included in the work [Ryali et al., 2016b], and the software code of Patel's tau is obtained through personal communications from Professor Stephen Smith, the first author of Smith et al. [2011]. Our CDN method is implemented in Python package cdn-fmri, publicly available from `https://pypi.org/project/cdn-fmri/`. The grid $\lambda = [0.01, 0.1, 1, 10, 100]$ is used for our CV tuning parameter selection.

2.3.1.1 Simulated Data From CDN

We consider three simulation scenarios regarding the ODE parameters in our CDN model:

S1: $\boldsymbol{B} = 0$ and $\boldsymbol{C} = 0$

S2: $\boldsymbol{B} = 0$

S3: $\boldsymbol{A} \neq 0$,$\boldsymbol{B} \neq 0$ and $\boldsymbol{C} \neq 0$

The first one does not consider either the stimulus activations or the stimulus modified connections, and the second one only excludes the stimulus modified connections. Because GCA models neither of these two effects, we include these two scenarios to provide biased advantages to GCA. The last scenario includes both the stimulus activations and the stimulus modified connections, and is a more realistic model for many task-related fMRI experiments.

For these scenarios, we consider a medium-sized network with 10 nodes with boxcar stimuli and different signal-to-noise (SNR) levels (3, 1, 0.5). We use the piece-wise linear basis. The simulations are repeated 50 times for each setting to represent fMRI data from 50 subjects, assuming for simplicity that they have the same underlying structure of brain connectivity, i.e. $\boldsymbol{\theta}$. We set that repetition time (TR) to be 0.72 s to demonstrate an ideal setting for GCA, because longer TR usually leads to low identification accuracy for directionality [Smith et al., 2011]. We also set the $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ based on the simulation code provided by Smith et al. [2011]. To validate the explanation of the colinearity issue in Section 2.2.2.2, we also reduce the task duration by 50% in one examplary S3 scenario.

AUC (area under the ROC curve) is used to evaluate the performance of recovering nonzero entries parameters. The AUC is calculated based on comparing the estimates against the true zero/non-zero statuses for all the entries of $\boldsymbol{A}$. We compare the performance of CDN with GCA and MDS in Figure 2.1. We are unable to compute the results for DCM here because the computation for this medium-sized network takes more than several days for each replication. It should be noted that the GCA connectivity estimates are based on the BOLD signals while ours are based on the estimated neuronal states. As shown in Figure 2.1, one can see that CDN performs much better than GCA in identifying the true connections for all SNRs and both stimulus types, partly because CDN models the neuronal states and stimulus activations instead of modeling the BOLD signals only as in GCA. Furthermore, CDN shows robust estimation results across different simulation scenarios. The AUCs of MDS increase as the SNRs increase. It has similar performance as CDN under high SNRs, but has suboptimal performance similar to GCA under low SNRs. Under low SNRs, it also tends to have high variability in AUCs.

To assess the statistical estimation accuracy of CDN, we use AUC and a scaled Frobenius norm for the connectivity-related parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ across different scenarios. Because $\boldsymbol{B}$ is a tensor, we define $||\boldsymbol{B}||_F = \sqrt{\sum_{m,n,j} B_{mnj}^2}$. The scaled Frobenious norm is defined as

$$\text{Fro}(\tilde{\boldsymbol{A}}, \boldsymbol{A}) = \frac{||(\boldsymbol{A} - \tilde{\boldsymbol{A}})||_F}{||\boldsymbol{A}||_F}$$

where $\tilde{\boldsymbol{A}}$ is an estimator for the true $\boldsymbol{A}$. This metric can be interpreted as the relative deviations from the truth.

For the sake of space. we only report the average performance metrics for box-car stimuli with varying SNR (3,1) in Table 2.1. The results are similar for other settings.

This table shows that our CDN algorithm estimates all the model parameters well, and the accuracy improves with increasing SNRs. As explained in Section 2.2.2.2, the accuracy for estimation $\boldsymbol{B}$ decreases for a shortened task duration.

2.3.1.2 Simulated Data From DCM

To test the robustness of our proposed model, we also compared the estimation performance using simulated data from DCM. We used the code and a simulation setting from Smith et al. [2011]. Due to the heavy computation cost of DCM, we considered a five node brain network with a shared stimulus input. To make this network complex enough to penalize the performance of our method, as shown by the previous simulation study, we let the stimulus input to all nodes, see Figure 2.2. Due to the high computation cost of DCM, we do not allow the connections to vary with the stimulus, and set $\boldsymbol{B} = 0$. We fixed TR=3 s to penalize our method because our CDN model is based on convolutions. To assess the robustness of using a canonical HRF in our model, our simulation setup also includes hemodynamic variability by introducing random perturbations to the ODE model parameters. The same approach is also used by [Smith et al., 2011]. We simulated 200 subjects with box-car shape stimuli which include five non-overlapped stimuli. We included one session for each stimulus which lasted 70 seconds.

Figure 2.3 assesses the performance of estimating $\boldsymbol{x}$ from one representative subject's simulated BOLD data. Our CDN estimates capture the main temporal variations in the neuronal states, even though the data are simulated using a relatively large TR. This shows that CDN is robust in recovering the neuronal states, even if the simulation model is different from our CDN model.

Applying DCM usually requires specifying a candidate model of stimulus activation and latent/induced connections. Based on the model, only selected entries in $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ are estimated while all the others are set to zero. To compare the accuracy of estimating latent connections $\boldsymbol{A}$, we specify $\boldsymbol{B}$ and $\boldsymbol{C}$ to have the same pattern as the truth, and let DCM estimate all the entries in ($\boldsymbol{A}$ using a fully connected network model. As before, we use AUC as before to evaluate the estimation accuracy. Table 2.2 compares the average AUC values for estimating $\boldsymbol{A}$ by CDN and DCM. CDN yields higher AUC values for both $\boldsymbol{A}$ using only a fraction of the computation time of DCM.

2.3.1.3 Simulated Data for Resting-state fMRI

As a simple extension, we use the modification described in Section 2.2.2.1 to fit our model to a simulated resting-state fMRI dataset (simulation scenario 4) from [Smith et al., 2011]. This dataset contains 50 brain nodes from 50 subjects, and the true network contains 10 subnetworks of size 5. More details on the simulation procedure and parameters are described in [Smith et al., 2011]. In the same paper, Patel's τ [Ramsey et al., 2014], an effective connectivity method, was the top performer for recovering the directionality of connections, and thus we also compare with it here.

Figure 2.4 compares the simulated and estimated BOLD time series from 4 representative nodes. Our model fits the simulated resting-state fMRI data well overall, though the fitted time series have a slightly smaller variation probably because our method ignores the neuronal fluctuation term. Based on the ROC evaluation metric described before, our method achieves an AUC value of 0.95 for recovering the intrinsic connectivity matrix on this simulated dataset. We are unable to fit the large-scale stochastic DCM method in [Razi et al., 2017] to this dataset because of

its long algorithmic convergence time. Patel's τ achieves an AUC value of 0.55 on this simulated dataset.

2.3.2 Application

2.3.2.1 Application to Task FMRI with Block Design

In this section, we apply our method to analyze a block-design fMRI dataset from the Human Connectome Project (HCP). The task is a language processing task developed by [Binder et al., 2011]. The scan session interleaves 4 blocks of a story task and 4 blocks of a math task. The length of each block varies with an average of approximately 30 seconds, and the lengths of the story task and math task are roughly matched. The story task asks the participants to classify the topic of the story as revenge or reciprocity for example, after they hear a brief (around 5-9 sentences) story adapted from Aesop's fables, The math task requires the participants to do serial addition and subtraction calculations, and then choose the correct answer. These questions were generated from the same text-to-speech method used in the story task. The participants were then asked to press a button under the right index finger to select the first choice, or a button under the right middle finger to select the second choice. Details about the task design were described in [Binder et al., 2011]. The HCP task fMRI study, including the imaging protocol, was described in [Barch et al., 2013]. We analyze the data from 100 examplary subjects, and we follow the suggested HCP preprocessing pipeline [Glasser et al., 2013] to preprocess the data.

We are interested in modeling a language network of regions that were implicated in a previous study [Turken and Dronkers, 2011] using both functional connectivity and diffusion tensor imaging. This network consists of four main regions, including

MTG (posterior middle temporal gyrus, MNI: -50,-38, 2), STG (superior temporal gyrus, MNI: -49,7,-12), STS (posterior superior temporal sulcus, MNI: -45, -62, 21) and IFG (inferior frontal gyrus, pars orbitalis, MNI: -44, 28, -7). After preprocessing the data, we extract the average BOLD time series using 8 mm radius balls centered around these four coordinates. We compute the p-values using 10000 bootstraps. Statistical significance was assessed using a p-value threshold of 0.01 after the FDR correction.

Figure 2.5 and 2.6 show the statistically significant activations and connections estimated by CDN. Our results show that these four ROIs are activated by the story task, but not the math task. This finding is consistent with the previous fMRI activation study [Binder et al., 2011]. The non-activation result of the math task fits prior evidence that math calculations do not engage the temporal lobe [Cappelletti et al., 2001, Crutch and Warrington, 2002]. Without considering directionality, the recovered intrinsic connections match the functional connectivity findings in [Turken and Dronkers, 2011]. This result showing a well-connected network is consistent with the notion that language comprehension engages a complex network of multiple regions interacting via multiple pathways [Mesulam, 1998, Dronkers et al., 2000].

The directionality recovered by CDN help better elucidate the roles of these ROIs. One key region highlighted by our results is MTG, because it is the converging point of various directional pathways in this network. MTG is regarded as a high-level region contributing to language comprehension. The key role of MTG is supported by various types of prior evidence. For example, MTG was a shared node of six different networks analyzed by a conjunction analysis [Koyama et al., 2010], and it was also found to be also among the most connected node in a resting-state fMRI network analysis of the cerebral cortex [Buckner et al., 2009]. A lesion analysis also showed that patients with MTG-lesions perform worse in various language tasks

[Dronkers et al., 2004], highlighting its critical role in the language comprehension process. In our CDN results, STG has directional connections pointing to all other three ROIs, suggesting that it may serve as one entry point of this network. Thus it is likely to be involved in the upstream of the language comprehension process. This implication fits the role of STG in basic morphosyntactic processing [Turken and Dronkers, 2011]. Another prior functional imaging evidence supporting this finding is that both speech and complex nonspeech sound activated STG [Wise et al., 1991, Binder et al., 1997].

Comparison of the task-dependent connections (Figure 2.6) shows that the story task modifies more network connections than the math task. In fact, almost all of the connections are impacted by the story task. This is again consistent with prior evidence that language comprehension engages multiple routes of the language comprehension network [Mesulam, 1998, Dronkers et al., 2000].

2.3.2.2 Application to Task FMRI with Event-related Design

We apply our CDN approach to a publicly available event-related fMRI dataset, downloaded from OpenfMRI.org under the access number ds000030. In the experiment, healthy subjects perform the stop-go response inhibition task inside the fMRI scanner. This task consists of two types of trials: go and stop. Specifically, on a go trial, subjects were instructed to press a button quickly when a go stimulus was presented on a computer screen; on a stop trial, subjects were to withhold from pressing when a go stimulus is followed shortly by a stop signal. We preprocess data using a suggested preprocessing pipeline based on FSL, a standard fMRI analysis software. See [Poldrack et al., 2016] for a detailed description of the experiment, dataset and preprocessing steps.

We are interested in studying the regions and their interconnections under either the go or stop stimuli. We select six brain regions implicated in prior publications, which are also validated by the meta-analysis tool from neurosynth.org. These six regions include M1 (primary motor cortex, MNI coordinate: -41, -20, 62), pos-preSMA (posterior presupplementary motor area, MNI: -4, -8, 60), ant-preSMA (anterior presupplementary motor area, MNI: -4, 36, 56) , SMA (supplementary motor area, MNI: -3, 6, 50), Thalamus (MNI: -12, -13, 7) , and STN (subthalamic nucleus, MNI: 6, -18, -2). The exact brain MNI coordinates for these regions are taken from Aron and Poldrack [2006], [Yeo et al., 2011], and Luo et al. [2013]. From the data of 100 healthy subjects, we extracted the average BOLD time series using 8 mm radius balls centered around these 6 coordinates. Each time series was standardized to mean zero and unit variance. We selected the tuning parameters using the approach described in Section 2.2.2.4, and computed the p-values using 10000 bootstraps as described in Section 2.2.2.6. Statistical significance was assessed using a threshold level of 0.01.

Figure 2.7 shows the statistically significant intrinsic connections and stimulus inputs estimated by CDN. These results identify different connectivity patterns between the anterior and posterior preSMAs. The posterior preSMA is closely connected to Thalamus, STN, and two motor regions (M1 and SMA), while the anterior preSMA is only connected to Thalamus and STN, but not the two motor regions. This finding corroborates the functional connectivity result based on resting-state fMRI in [Zhang et al., 2011]. Such finding supports the notion that the anterior preSMA is responsible for response inhibition [Haggard, 2008] while the posterior preSMA, along with the SMA, are involved in error detection [Li et al., 2006]. Unlike those results from functional connectivity analysis, CDN also infers connectivity directions and stimulus effects, and this leads to a more detailed understanding of

these connections and task inputs. For example, M1 receives input from the go stimulus only, while Thalamus and STN receive input from the stop stimulus only. This is well expected from the roles of these regions. M1 is responsible for button pressing under the go stimulus. Thalamus and STN are important regions implicated for response inhibition in several previous fMRI studies, see a review [Aron, 2007]. Based the directional connections, CDN recovers a pathway STN→Thalamus→M1. This pathway is consonant with the existing theory and previous fMRI results [Aron, 2007].

Figure 2.8 shows the connections induced by the go and stop stimuli respectively. Overall, the stop stimulus induces more connectivity strength changes across all the six regions, while go modifies a smaller number of connections. The results on task-induced connections also help better understand the network connection changes under different tasks. For example, CDN identifies a positive connection from the posterior preSMA to the anterior preSMA under stop, while the corresponding intrinsic and go-task dependent connections are not statistically significant. The latter negative finding on effective connectivity replicates the functional connectivity finding in a resting-state fMRI analysis [Zhang et al., 2011]. Importantly, the recovered connection from the posterior to anterior preSMA under stop provides an fMRI evidence for the theory that the posterior preSMA detects response conflicts and then projects to the anterior preSMA engaged in response inhibition, see a meta analysis of human neuroimaging studies on the preSMA [Ridderinkhof et al., 2004].

To check the model fit of CDN, we plot the representative BOLD signal and neuronal state time series from all the six regions in Figure 2.9. The BOLD time series constructed from CDN closely track the variation in the measured BOLD time series. This demonstrates that our CDN model fits the actual data well.

2.3.2.3 Application to Resting-state fMRI

In this section, we test the applicability of our method for recovering a medium-sized network using resting-state fMRI. The dataset is collected by the Human Connectome Project [Smith et al., 2013], and is preprocessed using a minimal pipeline [Glasser et al., 2013]. For the sake of manageable computation time, we analyze one scan session from each of 50 selected healthy subjects in this cohort. Each session contains 1200 scans with a temporal resolution of 0.72s. We adopt the same 36 ROI coordinates used in a recent large-scale resting-state DCM study [Razi et al., 2017]. For each ROI, the average BOLD time series from an 8 mm sphere are extracted for analysis. The MNI coordinates and network annotations of these ROIs are listed in Table 2.3.

To assess the validity of our method, Figure 2.10 compares our CDN estimate with the functional connectivity estimate based on correlations and a large-scale resting-state DCM (lrDCM) [Razi et al., 2017]. By visual comparison, our CDN estimate is close to rDCM, especially those connections with large magnitude. Both CDN and lrDCM estimates contain both positive and negative connections, while the correlation estimate shows mostly positive connections. Those strongest connections (in magnitude) are identified by all methods, for example between left and right V1. Our CDN and lrDCM also identify the asymmetry of brain connections while correlations cannot do so.

Because the values of correlation, CDN, lrDCM estimates are not directly comparable, as they are fitted from different models, we use the following binarization step to convert all the numerical estimates to network connections before the comparison. The rationale is that larger estimated values (in magnitude) in all the models are typically interpreted as stronger connections, and neuroscientists usually resort to

these connectivity methods to investigate if connections between certain brain nodes exist or not. Our binarization step converts the largest $100 \times q\%$ (in magnitude) of off-diagonal entries to 1, and sets the rest to 0, where q is a parameter. After converting to binary matrices, we then compute the percentages of network connections found by both methods or one method only. When $q = 20\%$, the resulting binary matrices of CDN and lrDCM overlap by 80.09%, while 9.95% of the CDN matrix entries are nonzero while the corresponding lrDCM entries are zero. This shows that CDN and lrDCM yield very similar estimates. It is worth mentioning that our CDN algorithm takes about 10 minutes in total while lrDCM takes about 19 hours with more memory use. The percentages for comparing CDN and correlations are 83.87% and 8.10% respectively. Figure 2.11 show the similar comparison results when q varies. Overall, there are large overlaps between the connections recovered by CDN and those by other methods.

2.4 Discussion

We develop a novel causal dynamic method to study brain activations and causal connections simultaneously from fMRI. Driven by the data nature of fMRI, CDN uses a functional data analysis approach to fit ODEs of neuronal states from BOLD time series. Unlike DCM, our model is estimated by an optimization algorithm, and this allows data-driven estimation of the ODE parameters, without DCM's requirement for hypothesizing the connections and stimulus inputs. The high computational efficiency of our algorithm also reduces the computation time.

Based on the task fMRI simulation studies, we show that our CDN approach is robust and accurate for recovering the ODE parameters for modeling dynamic brain

networks. Compared with other effective connectivity methods, including GCA, DCM and MDS, CDN performs the best in terms of parameter estimation and network identification across a wide range of scenarios. In the low SNR settings, CDN has the least variability followed by MDS, suggesting the robustness of our method. In the high SNR settings, both CDN and MDS achieve close to perfect network identification, while CDN yields a smaller parameter estimation loss in a realistic DCM simulation. GCA is relatively sensitive to SNRs, and its performance usually has substantial variability. Regarding the computational speed, CDN requires only a small fraction of the computation times of DCM and MDS. This makes it a very competitive approach for inferring effective connectivity. Patel's tau was a top method for recovering directional connectivity in a previous large-scale simulation study [Smith et al., 2011]. However, it only yields a mediocre recovery accuracy on our simulated resting-state dataset. A previous experimental validation [Wang et al., 2017] also found that its accuracy for estimating directionality was not better than chance. Thus, further research is needed to investigate the settings that impact the accuracy of Patel's tau. On the same simulated dataset, CDN achieves much higher network recovery accuracy than Patel's tau. All these comparisons suggest that CDN is an accurate and computationally tractable method for modeling effective connectivity and task activation.

For the analysis of the story/math task fMRI data, our activation and connectivity results are consistent with prior evidence on the language comprehension process and network. Our effective connectivity finding complements the structural and functional connectivity findings [Turken and Dronkers, 2011]. Notably, the directional connections in our results help identify the critical regions involved in the process. As we select only a few regions well supported by prior evidence in our analysis, the resulting network is relatively small and may omit other regions and pathways in-

volved. In order to obtain a more comprehensive view of the process, future analysis will need to target a larger number of candidate areas. For the analysis of the stop/go task fMRI, our effective connectivity results corroborate the functional connectivity findings using resting-state fMRI [Zhang et al., 2011], and these findings provide additional support for the existing theory and fMRI results on response inhibition [Ridderinkhof et al., 2004, Aron, 2007]. Our CDN results for both activation and connectivity also enhance the understanding of the distinct roles of the anterior and posterior preSMAs, and delineate the pathways involved in response inhibition under different tasks. The network size studied here may also be a limitation because we omit several other regions, including insula, caudate, and the inferior frontal cortex. These regions may mediate or contribute to the response inhibition process [Zhang et al., 2011, Aron, 2007].

CDN in this work is presented as a purely data-driven method. Other data-driven methods for effective connectivity were proposed in the literature, including latent-space GCA [David et al., 2008, Wheelock et al., 2014, Grant et al., 2015] and state-space multivariate dynamical systems [Ryali et al., 2011, 2016a,b]. In order to extend these data-driven methods, it is possible to constrain the network structures and input nodes based on prior knowledge or integrating with other sources of data. For example, one can restrict CDN to fit only those CDN connection parameters between nodes that correspond to significant resting-state functional connectivities [Razi et al., 2017] or anatomical connections [Sotero et al., 2010].

The number of parameters in CDN grows with the numbers of stimuli and nodes. This can be a potential issue to scale our model to large-scale networks with multiple stimuli, especially when the number of parameters exceeds the sample size. This relatively small sample size setting yields a so-called under-determined system, because there might exist multiple sets of parameters that yield the same fit of data. This

is further complicated by the colinearity issue introduced by the fMRI task designs. One possible direction is to introduce informative Bayesian priors to help invert such large systems with colinearity. We leave to future research on extending our model for large-scale networks.

The bilinear approximation used in this initial work is a simplified model for the biophysical mechanisms. It neglects several micro and macro neurophysicological processes, see a review [Daunizeau et al., 2011]. Therefore, the effective connectivity estimates may be distinct from anatomical connections or synaptic connections, because they fail to model several critical processes, such as gated connections [Stephan et al., 2008], extrinsic and intrinsic connections [Marreiros et al., 2008, Friston et al., 2017], and neuronal fluctuations [Li et al., 2011, Friston et al., 2011]. For example, the physiological noise could represent neuronal input from other regions not included in the model or resting-state fluctuations. Without modeling the neuronal noise, the resulting estimated connections between two regions may be indirect ones due to the shared input from other regions. It may also lead to higher estimation errors and over confidence [Daunizeau et al., 2012]. Without accounting for these biophysical processes, this bilinear approximation also limits the possibility of combining data from multiple modalities or understanding the nature of the BOLD response [Friston et al., 2017]. Though this approximation has been validated before for identifying the existence of effective connections, it may introduce estimation bias on the magnitude of these connections [Friston et al., 2011].

To extend our method to biophysically more realistic models, there remain several challenges for future research. More model parameters associated with these more sophisticated models will increase the computation burden. One should also be aware of the overfitting issue, especially for task-related fMRI where the recording sessions tend to be relatively short with limited repetitions of certain stimuli of interest. There

is clearly a trade-off between model complexity and data evidence, in addition to the concern of computation time. Bayesian priors, for example based on anatomical evidence, could be helpful for constraining the model complexity and computation time. However, these priors should be chosen carefully to ensure reliability and robustness [Frässle et al., 2015]. We should also note that another direction is to trade biophysical plausibility with large network sizes. For example, [Frässle et al., 2017, 2018] replace the bilinear approximation by a linear one in order to estimate large networks from task-related fMRI.

Another limitation of this first work is its canonical model for HRF. Though our model with the canonical HRF yields relatively robust results for one dataset simulated with HRF variations [Smith et al., 2011], this approach in general may lead to modeling bias. For example, the directionality of connections is determined from the temporal ordering of the neuronal states, as in the ODE model. Such temporal ordering may be reconstructed incorrectly from the observed BOLD signals, if the model does not account for the HRF variations across regions and subjects. Modeling HRF variations remains a challenging topic for various connectivity models as well, including DCM [Handwerker et al., 2012a] and functional connectivity [Rangaprakash et al., 2018]. In order to mitigate this issue, there are two possible extensions of our approach. One approach is to replace the canonical HRF used in our model by empirically estimated HRF for each region and each subject. A similar approach has been successful in resting-state functional connectivity estimation [Wu et al., 2013]. This step serves as data-driven deconvolution of the HRF, in the spirit of those latent-space GCA methods [David et al., 2008, Wheelock et al., 2014, Grant et al., 2015]. Because we treat the estimated HRF as given, one can easily adapt our algorithm to fit this extended model. However, to avoid double dipping of the same data, one may need to include a separate scan session to estimate HRF

[Aguirre et al., 1998]. A natural alternative without extra scanning sessions is to split out one part of the existing time series of each individual, and use that part to estimate the HRF. Either way can be challenging for task fMRI experiment designs, especially when the scanning time may be limited for task fMRI. Further research is also needed to study the optimal allocation of the splits in order to optimally balance the sample sizes available for two estimation steps. Another direction is to treat the HRF as unknown in our model. One will then need to add additional steps in our algorithm to estimate the HRF, for example using flexible functional bases. One may build on the joint estimation framework for detecting brain activation and estimating HRF simultaneously, see [Vincent et al., 2010, Chaari et al., 2013, Vincent et al., 2014]. Future research is needed to extend this framework to incorporate the ODE connectivity model. It is also important to note the possible challenges along this direction. First, though DCM also takes this direction and estimates the HRF empirically, one cannot assume the estimated HRF is accurate [Handwerker et al., 2012a]. Moreover, these additional parameters may introduce potential estimation issues related to collinearity and identifiability [Vincent et al., 2010]. More development, such as introducing regularization [Karahanoğlu et al., 2013] or Bayesian priors [Ryali et al., 2011], may be needed to ensure robust and stable estimation. With these two possible directions and potential issues, we will leave to future research to develop and compare these two possible extensions.

Our method here is mainly developed for task-related fMRI. It shows some promising results on analyzing resting-state fMRI with medium sized networks. However, our method, like others based on the deterministic DCM principal, can be computationally challenging for inverting large-scale networks, partly because our method needs to estimate the hidden neuronal states. For resting-state fMRI, the neuronal states recovered by CDN may serve as input data to other connectivity

methods, in order to minimize the confounding effect of HRF variability [Handwerker et al., 2004, Rangaprakash et al., 2018]. In the mean time, many large-scale resting-state DCM methods use the spectral domain characterization to avoid the estimation of the neuronal states and associated time-variant parameters [Friston et al., 2014, Razi et al., 2015, 2017, Frässle et al., 2017]. One future direction is to develop the spectral domain extension of our method, for analyzing large-scale resting-state fMRI data.

2.5 Data Availability

The stop/go task dataset was provided by the Consortium for Neuropsychiatric Phenomics LA Study from University of California, Los Angeles. The language/math task and resting-state fMRI datasets were provided by the Human Connectome Project (HCP; Principal Investigators: Bruce Rosen, M.D., Ph.D., Arthur W. Toga, Ph.D., Van J. Weeden, MD). The HCP was funded by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). HCP is the result of efforts of researchers from the University of Southern California, Martinos Center for Biomedical Imaging at Massachusetts General Hospital (MGH), Washington University, and the University of Minnesota. The HCP data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California. These studies that provided the data have been approved by their governing institutional review boards. The stop/go task-related fMRI dataset analyzed for this study can be found on openfmri.org under accession number ds000030 (https://openfmri.org/dataset/ds000030/). The language/math task and resting-state datasets can be found on the Human Connectome Project at

https://www.humanconnectome.org.

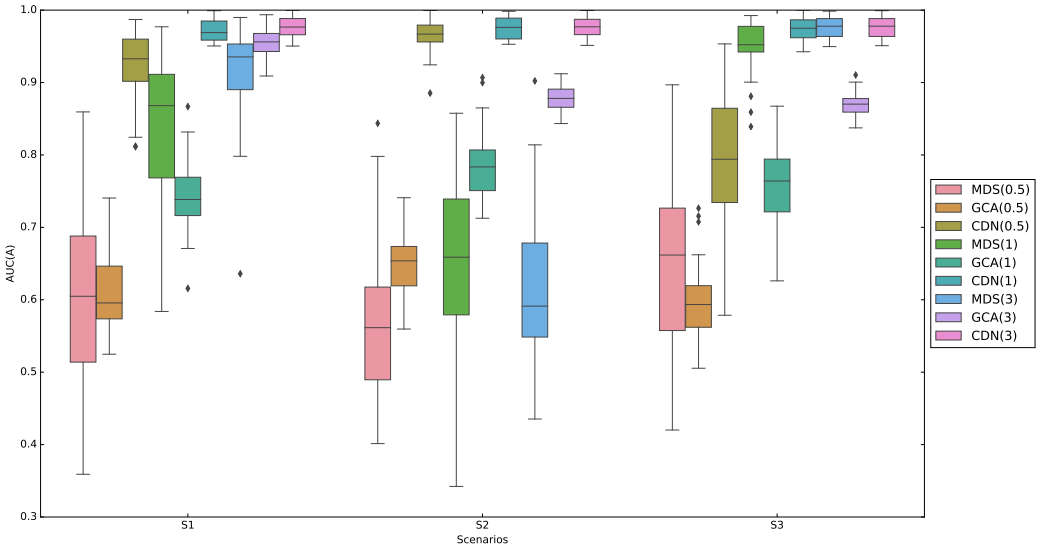

Figure 2.1: Comparison of AUCs by GCA and CDN for recovering nonzero intrinsic network connections under three simulation scenarios (S1–S3 described in Section 2.3.1.1) and varying signal-to-noise ratios (SNRs). For better visualization, we subtract random uniform jitters (between 0 to 0.05) for those cases when all points equal to 1.

Table 2.1: Estimation accuracy of CDN under the simulation scenarios described in Section 2.3.1.1. Box-car stimuli are used as a representative example. The accuracy metrics are averaged over 50 replications.

Scenario	J	SNR	AUC of A	AUC of B	AUC of C	Fro(A)	Fro(B)	Fro(C)
+1	10	0.5	1.00	-	-	0.52	-	-
+1	10	1	1.00	-	-	0.11	-	-
+2	10	0.5	1.00	-	0.98	0.38	-	0.56
+2	10	1	1.00	-	0.98	0.26	-	0.44
+3	1	0.5	0.95	0.71	-	0.55	0.92	0.56
+3	1	1	1.00	0.74	-	0.36	0.88	0.54
+3	1	1 (50% duration)	1.00	0.70	-	0.25	1.07	0.55

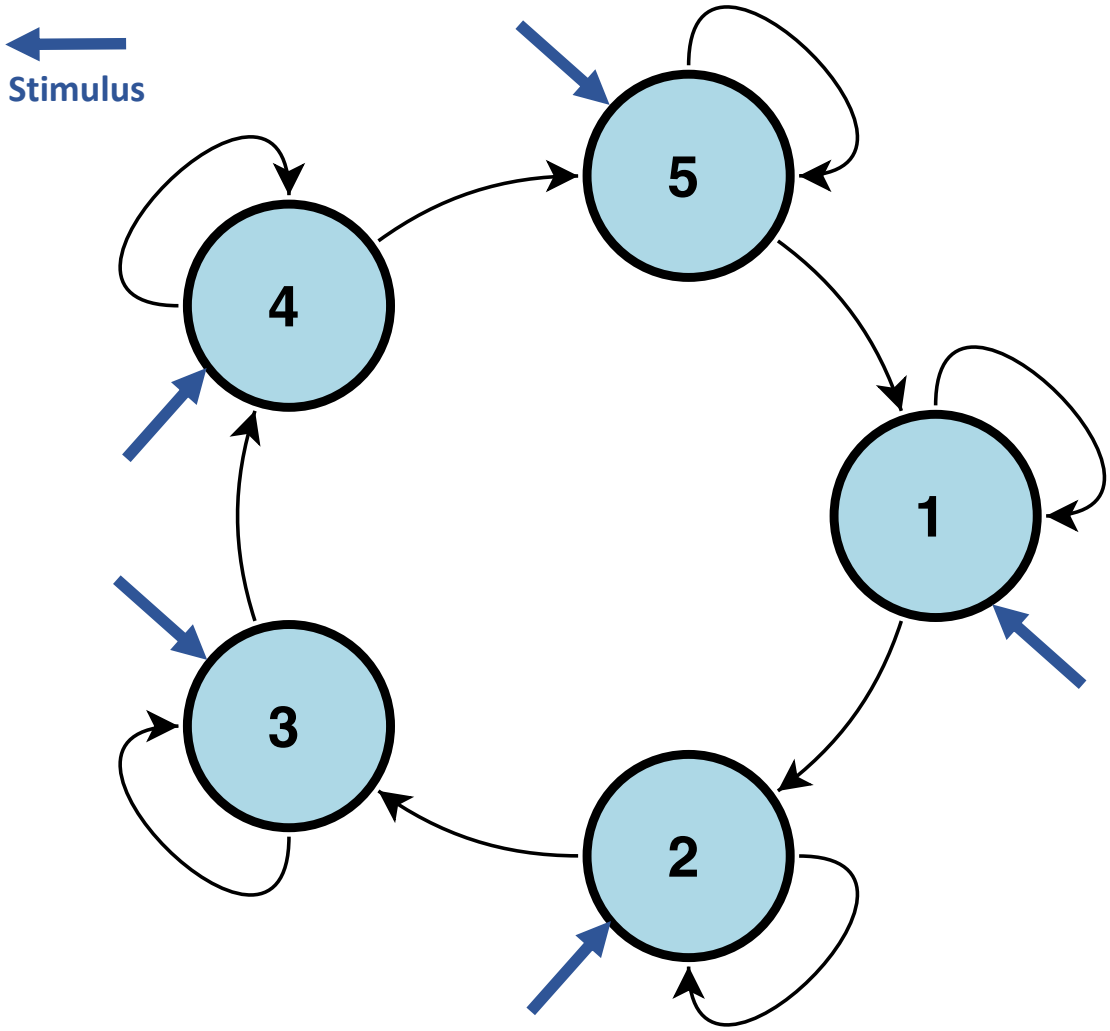

Figure 2.2: The network structure used to simulate DCM data in Section 2.3.1.2. All nodes are influenced by the same stimulus.

Table 2.2: Comparison of the AUCs, Frobenius losses, computation times of CDN and DCM, All computations are conducted in a computer with 8 Intel CPU cores (at 2.6 GHz) and sufficient memory for both algorithms. The best performance under each comparison criterion is highlighted in bold.

Method	AUC (A)	Fro loss	Computation cost
CDN	**0.77**	**1.29**	**18.95 seconds**
DCM	0.56	1.42	24 hours and 15 minutes
MDS	**0.77**	1.57	580 seconds

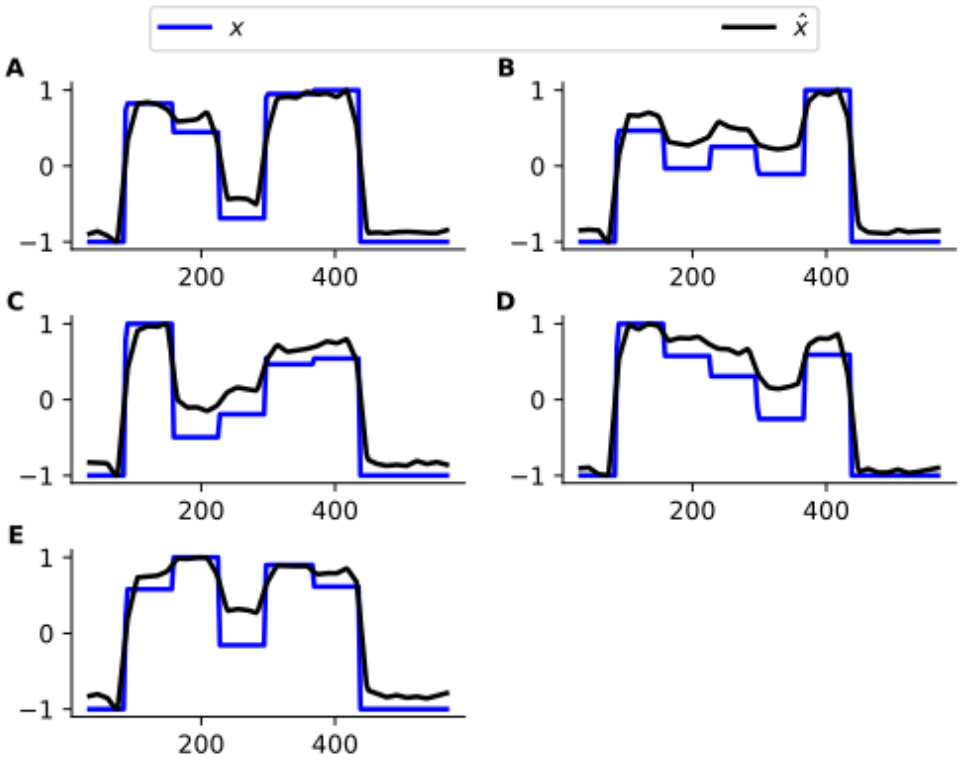

Figure 2.3: A representative example of the neuronal state time series of five nodes (A–E), simulated by DCM and their estimates by CDN . The DCM network model is plotted in Figure 2.2 . Because neuronal states have arbitrary units, all curves are scaled to have the same minimum and maximum values to illustrate the differences.

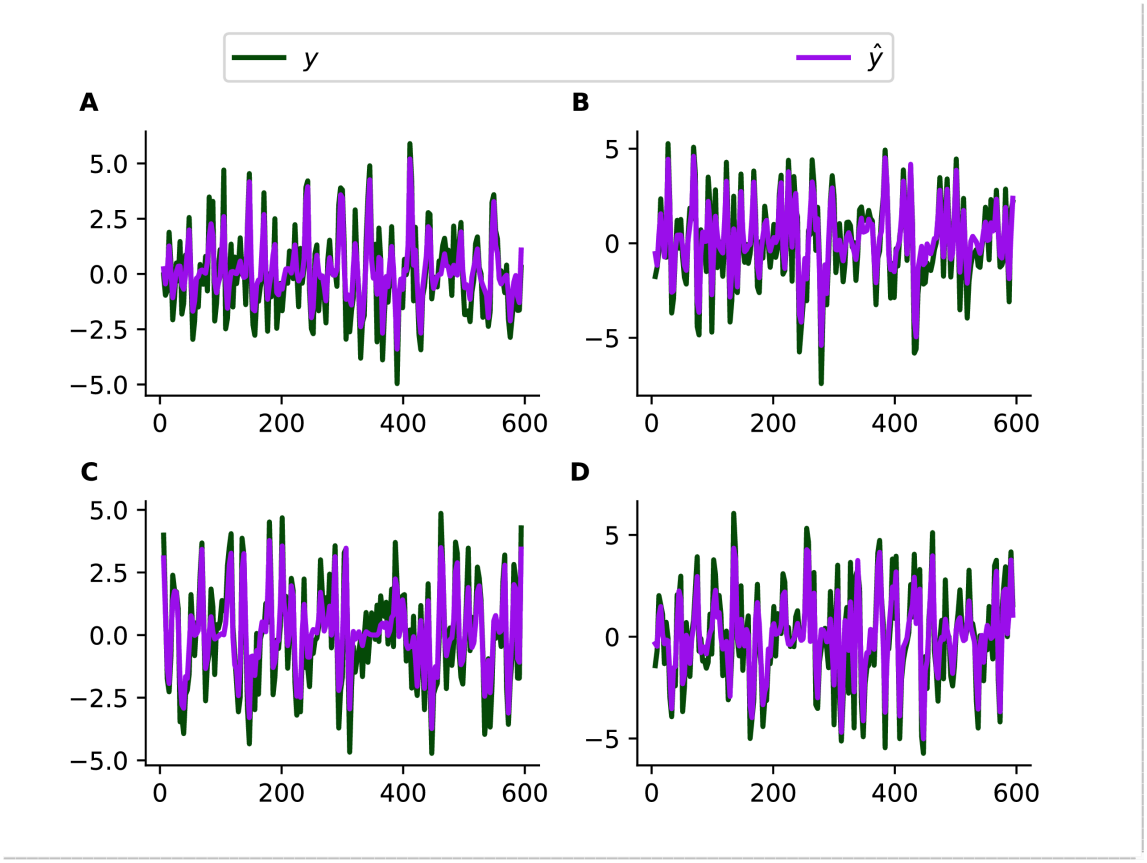

Figure 2.4: Simulated (y) and fitted ($\hat{y}$, by CDN) BOLD time series from four representative regions using a simulated resting-state fMRI dataset from [Smith et al., 2011].

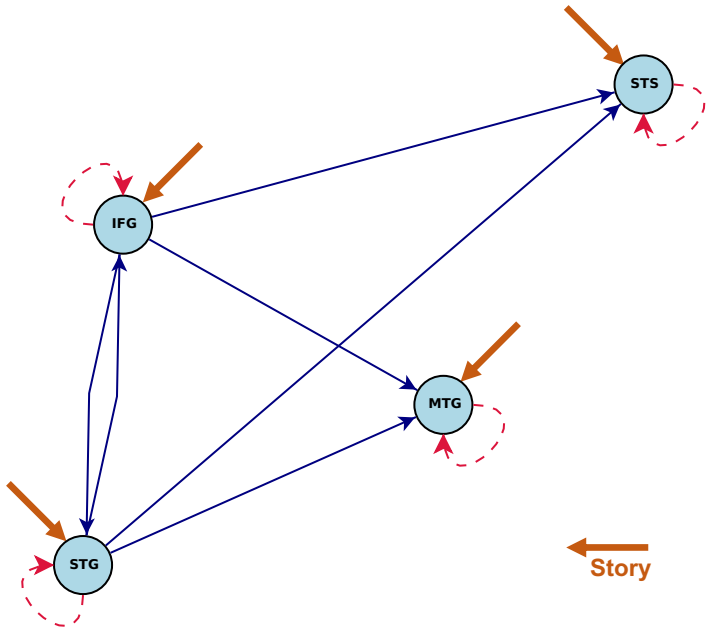

Figure 2.5: Estimated intrinsic connections in the math-story task experiment. Blue solid lines denote the positive causal connections, and red ashed lines denote negative connections. All connections and positive stimulus projections drawn are statistically significant at level 0.01, FDR corrected.

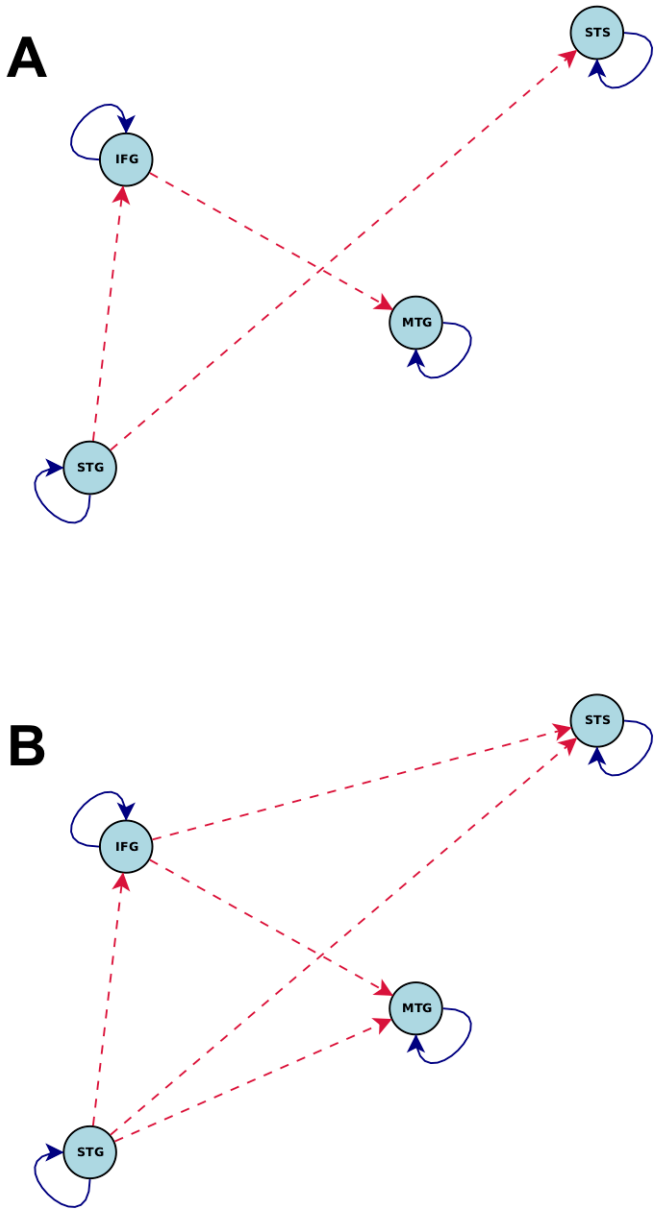

Figure 2.6: Connections induced by the (A) Math and (B) Story stimuli estimated by CDN in the math-story task experiment. Blue solid lines denote positive connections and red dashed arrows denote negative connections. All connections drawn are statistically significant at level 0.01, FDR corrected.

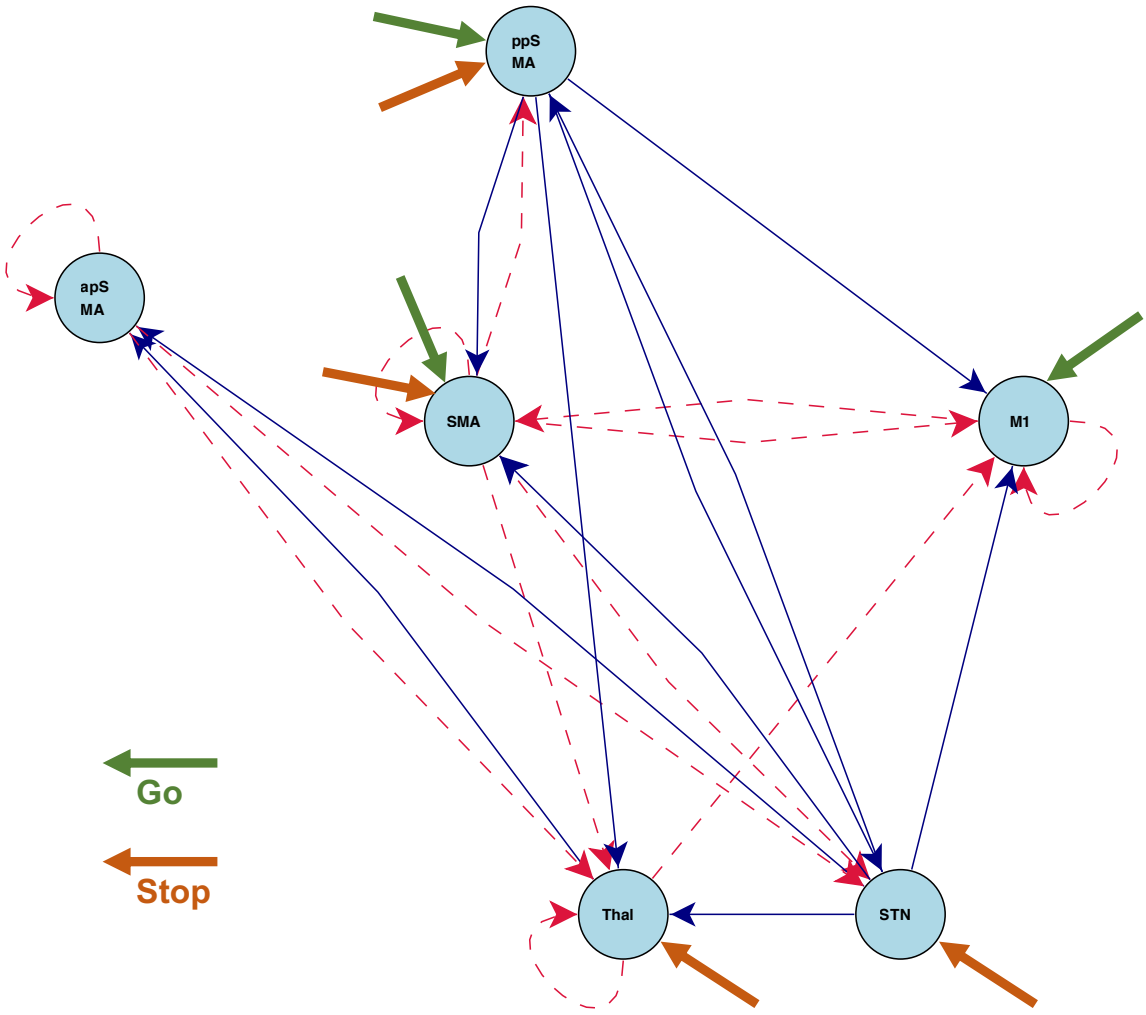

Figure 2.7: Estimated intrinsic connections in the stop-go task experiment. Blue solid lines denote the positive causal connections, and red ashed lines denote negative connections. All connections and positive stimulus projections drawn are statistically significant at level 0.01, FDR corrected.

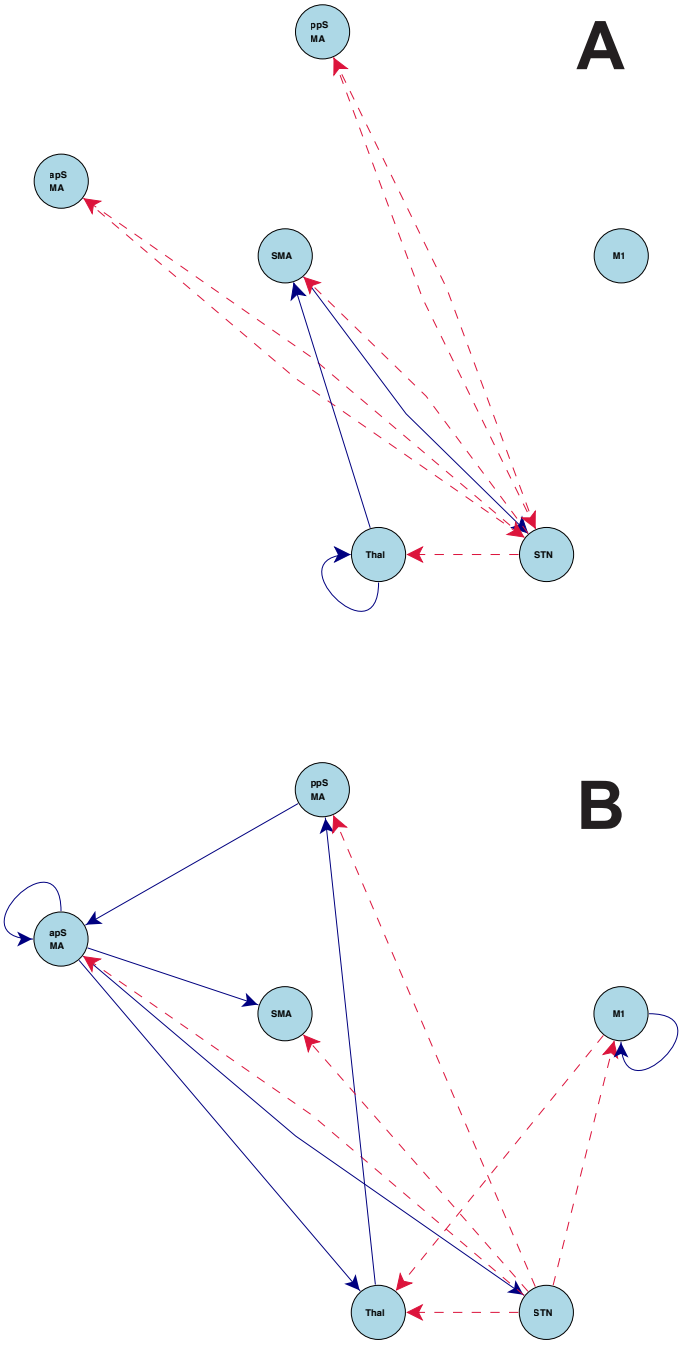

Figure 2.8: Connections induced by the (A) Go and (B) Stop stimuli estimated by CDN in the stop-go task experiment. Blue solid lines denote positive connections and red dashed arrows denote negative connections. All connections drawn are statistically significant at level 0.01, FDR corrected.

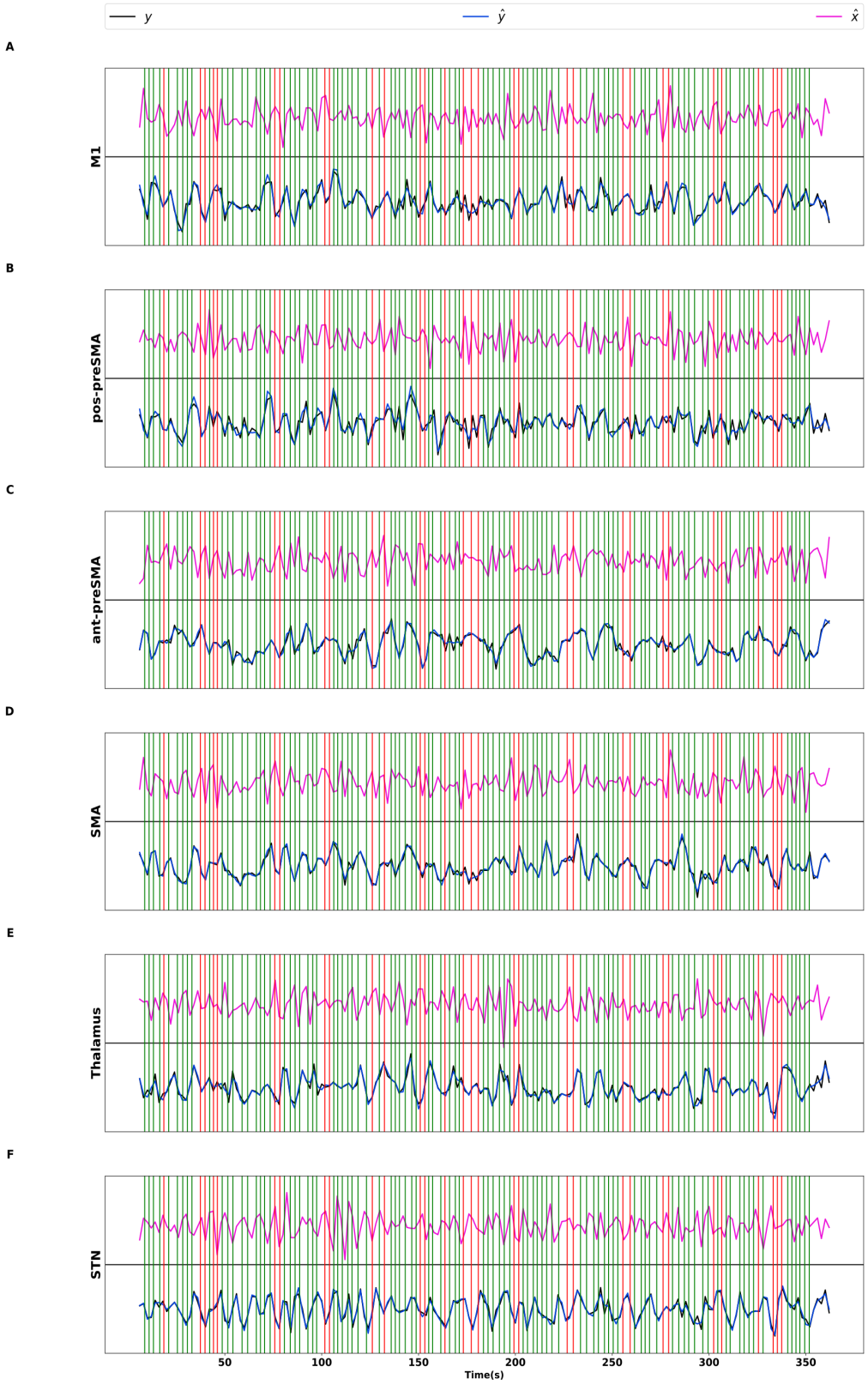

Figure 2.9: Real BOLD signals (y), estimated BOLD signals ($\hat{y}$) and neuronal activations ($\hat{x}$) from the STOP/GO experiment. Green/red vertical lines represent the GO and STOP stimuli respectively.

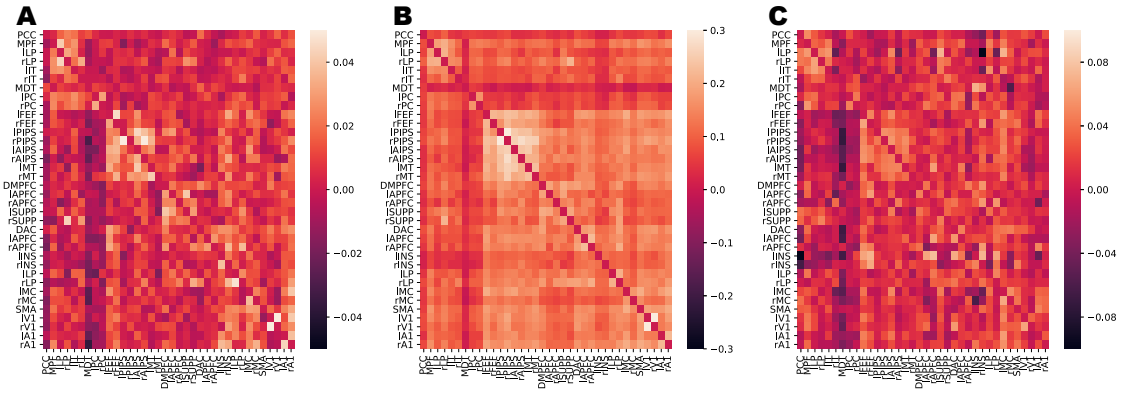

Figure 2.10: Averaged effective (A by CDN, C by [Razi et al., 2017]), functional (B, by correlations) connectivity estimates over 50 subjects. The diagonal elements are plotted as zeros in the figure to better illustrate the inter-regional connections.

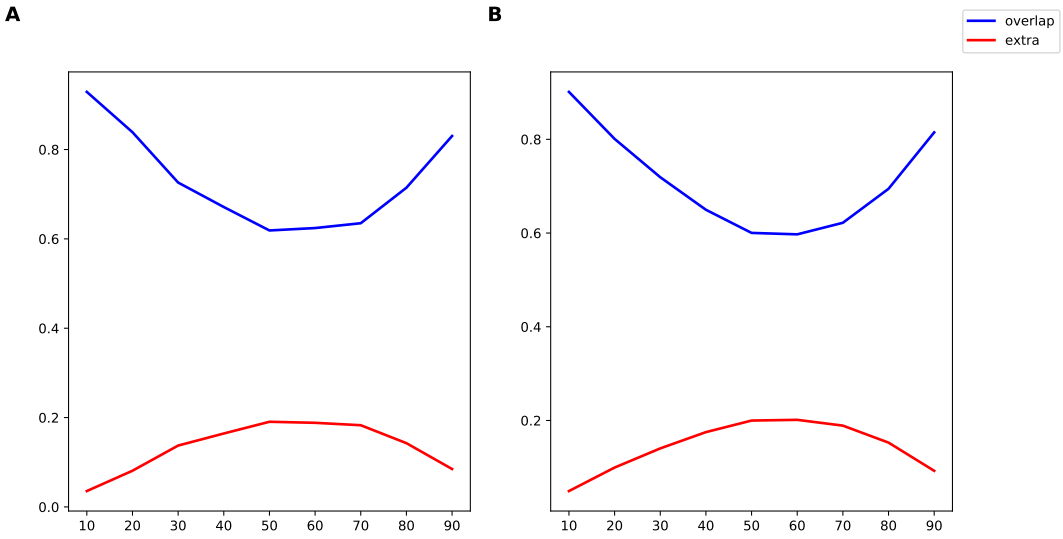

Figure 2.11: Comparison of identified connections recovered by different methods: (A) CDN vs correlations; (B) CDN vs lrDCM [Razi et al., 2017]. Varying thresholds in magnitude are shown on the x axis, blue lines show the percentages of connections shared by both methods compared, and red lines show the percentages of nonzero connections in CDN but not in the other method.

Table 2.3: Brain region names and MNI coordinates of the 36 ROIs used in Section 2.3.2.3.

	Name	Coordinates (in mm)
Default mode network		
1	Posterior cingulate/Precuneus	0 -52 7
2	Medial Prefrontal	-1 54 27
3	Left lateral parietal	-46 -66 30
4	Right lateral parietal	49 -63 33
5	Left inferior temporal	-61 -24 -9
6	Right inferior temporal	58 -24 -9
7	Medial dorsal thalamus	0 -12 9
8	Left posterior cerebellum	-25 -81 -33
9	Right posterior cerebellum	25 -81 -33
Dorsal attention network		
10	Left frontal eye field	-29 -9 54
11	Right frontal eye field	29 -9 54
12	Left posterior IPS	-26 -66 48
13	Right posterior IPS	26 -66 48
14	Left anterior IPS	-44 -39 45
15	Right anterior IPS	41 -39 45
16	Left MT	-50 -66 -6
17	Right MT	53 -63 -6
Control executive network		
18	Dorsal medial PFC	0 24 46
19	Left anterior PFC	-44 45 0
20	Right anterior PFC	44 45 0
21	Left superior parietal	-50 -51 45
22	Right superior parietal	50 -51 45
Salience network		
23	Dorsal anterior cingulate	0 21 36
24	Left anterior PFC	-35 45 30
25	Right anterior PFC	32 45 30
26	Left insula	-41 3 6
27	Right insula	41 3 6
28	Left lateral parietal	-62 -45 30
29	Right lateral parietal	62 -45 30
Sensorimotor network		
30	Left motor cortex	-39 -26 51
31	Right motor cortex	38 -26 48
32	Supplementary motor area	0 -21 48
Visual network		
33	Left V1	-7 -83 2
34	Right V1	7 -83 2
Auditory network		
35	Left A1	-62 -30 12
36	Right A1	59 -27 15

Chapter Three

Estimating High Dimensional ODE Models from Convoluted Observations with an Application to fMRI

3.1 Introduction

We have introduced a general model for effective brain connectivity analysis using functional magnetic resonance imaging (fMRI) in chapter 2. In this chapter, we are interested in extending this model to the high dimensional case. The brain connectivity has been usually categorized into well-known complex networks such as small-world topology which includes small dense clustering between neighboring nodes but lack of long-range connections [Bassett and Bullmore, 2006]. This feature makes it reasonable to assume sparse connections for the brain network. Furthermore, for large-scale networks, the number of model parameters quickly outweighs the number of observations as the number of brain regions increases. Thus a regularization approach is adopted. Least Absolute Shrinkage and Selection Operator (Lasso) was first introduced by Tibshirani [1996]. Lasso is widely used to do variable selection and regularization. There has been some work on sparse brain network modeling with lasso penalty [Lee et al., 2011, Seghouane and Khalid, 2012], group lasso penalty [Bolstad et al., 2011], Potts-based penalty [Zhang et al., 2015]. However, their works either consider only BOLD signal without experimentally-induced stimuli or use different types of data, e.g. electrocorticographic (ECoG) time series which has a high temporal and spatial resolution.

In this work, we propose a method to estimate causal connections of convoluted signals, in particular for our interest, neuronal activities that can be modeled with ordinary differential equations (3.2). We introduce a new optimization problem that combines the data fitting errors, fitting errors of the dynamic system and regularization term. This work proposes iterative algorithms to optimize the loss function to compute efficiently the parameters of differential equations, instead of comparing potentially a huge number of candidate ODE models or adding prior information to

parameters. Our method for a more general convolution-based observation model also adds to the vast statistical literature on ODE parameter estimation.

In section 3.2, we introduce our optimization problem for analyzing brain connectivity. Two algorithms are proposed in section 3.2.1. We will then compare their performances in different settings. Section 3.3 illustrates the numerical performance of our algorithm using extensive simulations. In particular, we will compare our method with DCM in section 3.3.2 with the data generated from DCM.

3.2 Sparse Causal Dynamic Network (SCDN)

Before introducing our optimization problem, we first elaborate the motivation of our formulation. Dynamic systems (3.1)

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{u}(t), \boldsymbol{x}(t), \boldsymbol{\theta}) \tag{3.1}$$

model output change directly by linking the output derivatives $\boldsymbol{x}'$ to $x(t)$ itself, as well as to inputs $\boldsymbol{u}$. For example, Dynamic Causal Modeling models the change of neuronal activities $\boldsymbol{x}$ in time, where each region is represented by a single hidden state. $\boldsymbol{x}$ is the neuronal activities of d regions, $\boldsymbol{u}$ is the input from J stimuli, and $\boldsymbol{\theta}$ is the ODE parameter of interest. It is usually difficult to estimate $\boldsymbol{f}$ in a general form, and instead, we can use a bilinear approximation for this dynamic system as follows

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + \sum_{j=1}^{J} u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t). \tag{3.2}$$

which was introduced in [Friston et al., 2003]. In equation 3.2, the entry A_{mn} in $\boldsymbol{A} = (A_{mn})_{dd}$ denotes the strength of intrinsic causal connection from n-th variable

to m-th variable. $\boldsymbol{B}_j = (B_{mnj})_{dd}$ denotes the influence of the j-th input $u_j(t)$ on the directional connection between these two variables. $\boldsymbol{C} = (C_{mj})_{dJ}$ with C_{mj} denotes the effects of u_j on m-th variable. The signal of interest, which we want to analyze is fMRI. In this setting, different variables come from different brain regions and $\boldsymbol{u}$ are experimental stimuli. The parameters $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ have important biophysical interpretations. $\boldsymbol{A}$ represents the intrinsic effective connectivity between brain regions, $\boldsymbol{B}$ models how these connections are influenced by external stimuli, and $\boldsymbol{C}$ models how brain regions are activated (or suppressed) by stimuli.

In general, equation (3.2) can also model other variations of a system with external influence $\boldsymbol{u}$ when dynamic connections among different variables of the system are of interest. According to equation (3.2), it is natural to add

$$\int \|\frac{d\boldsymbol{x}(t)}{dt} - (\boldsymbol{A}\boldsymbol{x}(t) + \sum_{1\leq j\leq J} u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t))\|^2 dt$$

to our objective function.

As we mentioned before, instead of creating another set of differential equations, we will use the observational model (2.2). This leads us to add

$$\sum_{t_i, 1\leq i\leq T} \|\boldsymbol{y}(t_i) - \beta \star \boldsymbol{x}(t_i))\|^2$$

to our loss function. t_i for $1 \leq i \leq T$ are discrete time points where the data is obtained. For fMRI data discussed in our work, $\beta(s) = h(s)$ where h is canonical hemodynamic function.

In summary, our model (3.3) is different than DCM: we use a simple linear

convolution mapping instead of another set of differential equations.

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{A}\boldsymbol{x}(t) + \sum_j u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t) \tag{3.3a}$$

$$\boldsymbol{y}(t) = \int h(s)\boldsymbol{x}(t-s)ds + \boldsymbol{\epsilon}(t) \tag{3.3b}$$

where $\boldsymbol{y}(t)$ is a $d \times 1$ vector of BOLD signal at time t of d regions ($\boldsymbol{y} = (y_1, ..., y_d)$), $\boldsymbol{\epsilon}(t)$ is the error process, $h(s)$ is a hemodynamic response function. The integration for vectors above should be understood as component-wise. Each dimension of $\boldsymbol{y}(t)$ for a brain region depends only on its corresponding neuronal component in $\boldsymbol{x}(t)$ ($\boldsymbol{x} = (x_1, ..., x_d)$) for the same region.

The mapping from neuronal states to BOLD signals is represented in equation 3.3b, where h is a hemodynamic response function. It is an approximation [Henson and Friston, 2007, Friston et al., 2003] for the differential equation model of DCM which we did not discuss in detail. This approximation is crucial to simply computation when we optimize the loss function. Our approach works for any hemodynamic response functions of fMRI. For simplicity, the canonical hemodynamic response function is employed [Friston et al., 1998] in this work.

From a purely mathematical point of view, without considering the noise (or equivalently $\boldsymbol{\epsilon}(t) = 0$), we are able to show by the following lemma that the latent states $\boldsymbol{x}(t)$ are uniquely identifiable from Equation 3.3b. The proof follows from Titchmarsh's convolution theorem [Titchmarsh, 1926].

Lemma 1. *Given two functions $x_1(t)$ and $x_2(t)$, if $h * x_1(t) = h * x_2(t)$ for $t \in [0, t_T]$, and h is a hemodynamic response function where $h(t) = 0$ for $t < 0$ and $h(t) \neq 0$ a.e. for $t \in [0, t_T]$, then $x_1 = x_2$ a.e. for $t \in [0, t_T]$.*

Proof. To prove this lemma, we simply use Titchmarsh's convolution theorem [Titchmarsh, 1926]: if $\phi(t)$ and $\psi(t)$ are integrable functions, such that $\int_0^x \phi(t)\psi(x-t)dt = 0$ almost everywhere in the interval $0 \leq x \leq k$, then there exists $a_1 > 0$ and $a_2 > 0$ satisfying $a_1 + a_2 > k$ such that $\phi(t) = 0$ almost everywhere in $(0, a_1)$ and $\psi(t) = 0$ almost everywhere in $(0, a_2)$. From this theorem, we can see if we have

$$\int_0^t h(s)(x_1(t-s) - x_2(t-s))ds = 0$$

for $0 \leq t \leq t_T$, $x_1 - x_2$ must be equal to zero a.e. for the entire interval because there is no such interval $[0, t_1]$ in which h is equal to zero a.e.. □

Now we are ready to introduce our optimization problem. By adopting regularity approach, we get our objective function, the one we want to minimize,

$$\begin{aligned} l(\boldsymbol{x}, \boldsymbol{\theta}) = & \sum_{t_i, 1 \leq i \leq T} \|\boldsymbol{y}(t_i) - h \star \boldsymbol{x}(t_i))\|^2 + \\ & \lambda \int \|\frac{d\boldsymbol{x}(t)}{dt} - (\boldsymbol{A}\boldsymbol{x}(t) + \sum_{1 \leq j \leq J} u_j(t)\boldsymbol{B_j}\boldsymbol{x}(t) + \boldsymbol{C}\boldsymbol{u}(t))\|^2 dt \\ & + \lambda\mu(\sum_{i=1}^{d}\sum_{k=1}^{d} \|w_{1k}\|_{L_2}|A_{ik}| + \sum_{i=1}^{d}\sum_{j=1}^{J}\sum_{k=1}^{d} \|w_{2jk}\|_{L_2}|B_{jik}| + \\ & \sum_{i=1}^{d}\sum_{k=1}^{J} \|w_{3k}\|_{L_2}|C_{ik}|) \end{aligned}$$

where $\boldsymbol{\theta}$ includes $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, $\lambda > 0$ and $\mu > 0$ are tuning parameters, and $\|\cdot\|$ is the ℓ_2 norm if not mentioned otherwise. We use $\lambda\mu$ as the tuning parameter of the regularization term for the ease of notation. $w_{1k} = x_k$, $w_{2jk} = u_j x_k$ and $w_{3k} = u_k$ The first part of the loss function is the data fitting error and the second part is regarded as the fidelity to our dynamic system. The third part is essentially the lasso penalty with different normalization weights corresponding to their "features"

(e.g. A corresponding to x).

3.2.1 Estimation

It is common to represent a function with a set of basis. This idea comes from both finite difference methods of solving partial differential equations and functional data analysis [Ramsay, 2006]. We represent $\boldsymbol{x}(t)$ using truncated basis function expansions. This technique allows us to reduce the original problem to a finite-dimensional optimization problem. Let $\boldsymbol{x}(t) = \boldsymbol{\Gamma}\boldsymbol{\Phi}(t)$ where $\boldsymbol{\Phi}(t)$ is a $p \times 1$ vector with entries denoting the basis function value at t and p is the number of basis functions. We choose p to be reasonably large. Usually, we should select a large p if there exists stimulus changing rapidly so the approximation of $\boldsymbol{x}$ is capable to catch the variations of stimuli. We will use a piece-wise linear basis for box-car stimuli (a popular one in fMRI experiments). There exist many other possible choices such as a cubic spline basis. As stimuli are generally not continuous or smooth, it will be more reasonable if $\boldsymbol{x}$ is approximated by piece-wise linear bases.

With the basis representation, our loss function becomes

$$
\begin{aligned}
l(\boldsymbol{\Gamma}, \boldsymbol{\theta}) = \sum_{t_i} \|\boldsymbol{y}(t_i) - \boldsymbol{\Gamma}[h \star \boldsymbol{\Phi}(t_i)]\|^2 + \\
\lambda \int \|\boldsymbol{\Gamma}\boldsymbol{\Phi}'(t) - (\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\Phi}(t) + \sum_j u_j(t)\boldsymbol{B_j}\boldsymbol{\Gamma}\boldsymbol{\Phi}(t) + \boldsymbol{C}\boldsymbol{u}(t))\|^2 \\
+ \lambda\mu(\sum_i \sum_k \|w_{1k}\||A_{ik}| + \sum_i \sum_j \sum_k \|w_{2jk}\||B_{jik}| + \\
\sum_i \sum_k \|w_{3k}\||C_{ik}|)
\end{aligned}
$$

where $w_{1k} = \mathbf{\Gamma}(k)\mathbf{\Phi}(t)$, $w_{2jk} = u_j\mathbf{\Gamma}(k)\mathbf{\Phi}(t)$ and $\mathbf{\Gamma}(k)$ is the k-th row of $\mathbf{\Gamma}$.

The loss function $l(\mathbf{\Gamma}, \boldsymbol{\theta})$ is not jointly convex for $(\mathbf{\Gamma}, \boldsymbol{\theta})$ which makes it impossible to find a global minimum for the problem. This motivates us to introduce our first algorithm SCDN-1 to minimize the loss function using iterative block-wise updates for $\mathbf{\Gamma}$ and each column of $\boldsymbol{\theta}$ respectively, summarized in Algorithm 2.

Instead of using popular coordinate descent method for lasso penalty update [Wu and Lange, 2008], we use block coordinate descent to update θ. By observing every column of $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]$ can be optimized together with explicit form, we update every column of $\boldsymbol{\theta}$ which is much faster than traditional coordinate descent, i.e. updating one entry every time. This estimation method allows us to estimate the brain network with hundreds of nodes and hundreds of subjects. As solving the minimum for the separate matrix $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ involves inversing high dimensional matrix, which might be numerically unstable, we choose not to go further to speed up our estimation by updating whole matrix $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$.

Algorithm 2 SCDN-1

Input: BOLD signal $\boldsymbol{y}$, stimulus $\boldsymbol{u}$, basis function $\mathbf{\Phi}$, tuning parameter λ
Initialize $\mathbf{\Gamma}$, $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{B_1}, .., \boldsymbol{B_J}, \boldsymbol{C}]$
repeat
  Update $\mathbf{\Gamma}$ of $l(\mathbf{\Gamma}, \boldsymbol{\theta})$ given $\boldsymbol{\theta}$ by gradient descent with backtracking line search
  repeat
    Randomly select one column of parameters from $\boldsymbol{\theta}$ and update the parameters by block coordinate descent method with Gauss Seidel style
  until convergence
until convergence
Return $\mathbf{\Gamma}$, $\boldsymbol{\theta}$

As we did simulations with Algorithm 2, we found it was really difficult to get an estimation of the parameter $\boldsymbol{B}$ with high accuracy. The first reason for this problem is that the duration of each stimulus is relatively shorter than the duration of the

experiment which limits the relevant data to calculate $\boldsymbol{B_i}$ for i-th stimulus. Another reason is neuronal activities $\boldsymbol{x}$ are highly correlated. To conquer these problems, we propose another algorithm: SCDN-2. In this algorithm, we provide a simple remedy to improve our estimation. We first estimate $\boldsymbol{A}$ and $\boldsymbol{C}$ with $\boldsymbol{B} = 0$ fixed. Then we update $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{B}$ together with $B_{ijk} = 0$ if $A_{ij} = 0$. The motivation behind this approach is if we fix $\boldsymbol{B} = 0$, the influence of $\boldsymbol{u}$, the stimuli can be reflected on $\boldsymbol{A}$. The SCDN-2 algorithm is summarized in Algorithm 3.

Algorithm 3 SCDN-2

Input: BOLD signal $\boldsymbol{y}$, stimulus $\boldsymbol{u}$, basis function $\boldsymbol{\Phi}$, tuning parameter λ
$(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ = SCDN-1(Input)
Initialize $\boldsymbol{\Gamma}$, $\boldsymbol{\theta} = [\boldsymbol{A}, \boldsymbol{B_1}, .., \boldsymbol{B_J}, \boldsymbol{C}]$, $\boldsymbol{A_{init}} = \boldsymbol{A}$
repeat
  Update $\boldsymbol{\Gamma}$ of $l(\boldsymbol{\Gamma}, \boldsymbol{\theta})$ given $\boldsymbol{\theta}$ by gradient descent with backtracking line search
  repeat
    Randomly select one column of parameters from $\boldsymbol{\theta}$ and update the parameters by block coordinate descent method with Gauss Seidel style. If the column in $\boldsymbol{B}$ is chosen, update the entry if the corresponding value in $\boldsymbol{A_{init}}$ is nonzero.
  until convergence
until convergence
Return $\boldsymbol{\Gamma}$, $\boldsymbol{\theta}$

3.2.2 Algorithm updates and analysis

As mentioned before, the iterative algorithm has an explicit form which we will state in this section. First we introduce the following notations in order to illustrate our updating procedure. In the following definition, P_i is a matrix with entry $P_i[m, n]$ or $P_i[m, n, j]$. P' is the transpose of P. $\Phi = (\phi_1, ..., \phi_p)$ is our selected basis for

estimating neuronal activity and ϕ_i' is the first order derivative.

$$P_1[n,m] = \int_{t_0}^{t_T} \phi_n' \phi_m' dt \quad P_2[n,m] = \int_{t_0}^{t_T} \phi_n(t)\phi_m'(t)dt$$
$$P_3[n,m,j] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)\phi_m'(t)dt \quad P_4[j,n] = \int_{t_0}^{t_T} u_j(t)\phi_m'(t)dt$$
$$P_6[n,m,j] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)\phi_m(t)dt \quad P_5[n,m] = \int_{t_0}^{t_T} \phi_n(t)\phi_m(t)dt$$
$$P_7[j,n] = \int_{t_0}^{t_T} u_j(t)\phi_n(t)dt \quad P_8[,n] = \int_{t_0}^{t_T} \phi_n' dt$$
$$P_9[,n] = \int_{t_0}^{t_T} \phi_n(t)dt \quad P_{11}[n,j] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)dt$$
$$P_{12}[t,n] = \int_{t_0}^{t_T} h(s) * \phi_n(t-s)ds \quad P_{14}[n,m] = \int_{t_0}^{t_T} u_n(t)u_m(t)dt$$
$$P_{13}[l,n,j] = \int_{t_0}^{t_T} u_l(t)\phi_n(t)u_j(t)dt \quad P_{15}[,n] = \int_{t_0}^{t_T} u_n(t)dt$$
$$P_{10}[n,m,j,k] = \int_{t_0}^{t_T} \phi_n(t)u_j(t)\phi_m(t)u_k(t)dt$$

Theorem 1. *Given Γ, $A[:,-n]$, B and C, the minimizer of l is given by*

$$\tilde{A}[,n] = \frac{\text{sign}(T_{1_A})\max(|T_{1_A}| - T_{2_A}, 0)}{T_{3_A}}$$

where $-n$ represents the matrix without n-th column.

$$T_{1_A} = \Gamma P_2' \Gamma[n,]' - A\Gamma P_5 \Gamma[n,]' - \sum_j B_j \Gamma P_{6j} \Gamma[n,]' - C P_7 \Gamma[n,]' + A[,n]\Gamma[n,]P_5\Gamma[n,]'$$

$$T_{2_A} = \frac{\mu}{2}\sqrt{\Gamma[n,]P_5\Gamma[n,]'}$$

$$T_{3_A} = \Gamma[n,]P_5\Gamma[n,]'$$

Theorem 2. *Given Γ, A, B and $C[:,-j]$, the minimizer of l is given by*

$$\tilde{C}[,j] = \frac{\text{sign}(T_{1_C})\max(|T_{1_C}| - T_{2_C}, 0)}{T_{3_C}}$$

where $-j$ represents the matrix without j-th column.

$$T_{1_C} = \Gamma P_4[j,]' - A\Gamma P_7[j,:]' - \sum_j B_j \Gamma P_{13j}[j,] - CP_{14}[j,]'$$

$$T_{2_C} = \frac{\mu}{2}\sqrt{P_{14}[j,j]}$$

$$T_{3_C} = P_{14}[j,j]$$

Theorem 3. *Given Γ, A, B_1, ...,$B_j[:,-n]$, ..., B_J and C, the minimizer of l is given by*

$$\tilde{B}[,n,j] = \frac{\text{sign}(T_{1_B})\max(|T_{1_B}| - T_{2_B}, 0)}{T_{3_B}}$$

where $-n$ and $-j$ indicates B_j without n-th column.

$$T_{1_B} = \Gamma P_{3j}\Gamma[n,]' - A\Gamma P'_{6j}\Gamma[n,]' - \sum_l B_l \Gamma P_{10lj}\Gamma[n,]' - CP_{13j}\Gamma[n,]' + B_j[,n]\Gamma[n,]P_{10jj}\Gamma[n,]'$$

$$T_{2_B} = \frac{\mu}{2}\sqrt{\Gamma[n,]P_{10jj}\Gamma[n,]'}$$

$$T_{3_B} = \Gamma[n,]P_{10jj}\Gamma[n,]'$$

3.3 Simulations

We use two different sets of simulated data to study our method.

- Data simulated from convolution model 3.3

- Data simulated from DCM

3.3.1 Simulations with model 3.3

We first investigate the accuracy of our estimation method by using extensive simulations from the convolution model. We randomly simulate $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ of the dynamic system. These parameters are also chosen to ensure that the solution of the system will not blow up, which means we add the constraints to the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$. These constraints are reasonable as the brain connectivity will not diverge to infinity during any time period. In all simulation settings, stimuli are box-car type, $T = 300$ and $TR = 0.72$s. They are similar settings as those in our real data examples analyzed in section 2.3.2. The form of $\boldsymbol{A}$ is chosen to be similar to those employed in Smith et al. [2011] in which for large brain networks, all nodes form several small clusters and there exist some sparse connections between different clusters.

We study our estimation under the case that experimental stimuli have an influence on brain connectivity ($\boldsymbol{B} \neq 0$). For each simulation, we randomly generate data with 50 independent subjects, given the same parameters $\boldsymbol{\theta}$. We should notice that by studying the brain network with 50 or more brain regions, we essentially need to estimate thousands of parameters. We conducted our simulation with different numbers of stimuli (1, 3) and different SNRs (1, 3) to test the stability and accuracy of our estimation method. AUC is chosen as our evaluation and comparison criteria. The score of one causal connection is evaluated by the frequency of corresponding edge presence across different trails/subjects.

To compare our two algorithms SCDN-1 and SCDN-2, we introduce two experimental settings: (i) nonzero entries of $\boldsymbol{B}$ are included in those of $\boldsymbol{A}$, (ii) nonzero

Table 3.1: Simulation settings for task-related Brain Network of 50 nodes

Index of Simulations	SNR	# of Stimuli	# of Nodes
sim_0	1	1	50
sim_1	3	1	50
sim_2	1	3	50
sim_3	3	3	50

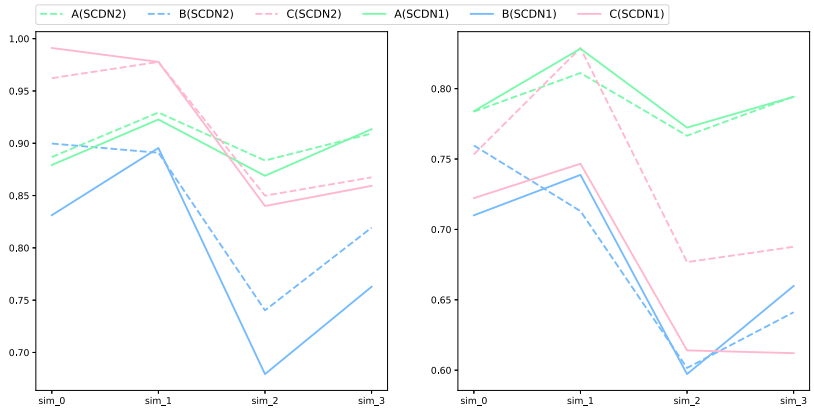

Figure 3.1: Two figures show the comparison of two algorithms SCDN-1 and SCDN-2. Left panel shows the case that nonzero entries of $\boldsymbol{B}$ are included in those of $\boldsymbol{A}$ while right panel shows the other case. See Table 3.1 for setting of simulations

entries of $\boldsymbol{B}$ are not included in those of $\boldsymbol{A}$. Other simulation settings are summarized in Table 3.1. A comparison of our algorithms are shown in Figure 3.1. Figure 3.1 shows that SCDN-2 performs better than or similar to SCDN-1. Even in the setting (ii), nonzero entries of $\boldsymbol{B}$ are not included in those of $\boldsymbol{A}$, the advantages of SCDN-2 can still be recognized. Furthermore, in general, our algorithms show a better estimation accuracy for setting (i) comparing to the setting (ii). This can be explained by that in setting (1), $\boldsymbol{B}$ can be treated as an enhancement of $\boldsymbol{A}$ when corresponding stimuli are created and thus the information of $\boldsymbol{A}$ can be passed to $\boldsymbol{B}$.

Figure 3.1 shows our estimation is very robust to different SNRs. It is critical

to the real-world application of our model and method. However, we can see the estimation for $\boldsymbol{B}$ is much more difficult than that for other parameters due to the number of entries in $\boldsymbol{B}$. Another possible solution to this problem is to add some constraints of the non-zero pattern of parameter $\boldsymbol{B}$. In real-world applications, $\boldsymbol{B}$ is usually much more sparse than $\boldsymbol{A}$ and $\boldsymbol{C}$ which makes it easier to specify the prior pattern of $\boldsymbol{B}$. It is shown in the figure that an increasing number of stimuli will decrease the estimation accuracy, which is as expected because the number of parameters of $\boldsymbol{B}$ increases dramatically as we include more stimuli.

For these simulations, we use tuning parameters: $\lambda = [0.1, 0.01, 1, 10, 100]$ and $\mu = [1, 3, 5, 10, 30, 50, 100]$. It is also possible to include different μ for $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ because they may have different sparsity. The formula can be derived similarly as in our derivation in 3.2.2. However, this can potentially increase the computation time for tuning parameter selection. Tuning parameter selection usually costs most of the computation time.

3.3.2 Simulations with DCM

To test the robustness of our proposed convolution model, we also generate data from DCM. The detailed simulation setting and code are provided by Smith et al. [2011]. Similar to the previous chapter, we only consider a five-node brain network with shared inputs. To make this network complicated enough to penalize the performance of our method, we create a network with inputs to all nodes with the setting of $\boldsymbol{B} = 0$ and 5 different box-car stimuli. To make the simulation close to common fMRI experiments, we fix $TR = 3$s, $SNR = 3$ and simulate 100 independent subjects with the same stimuli and underlying dynamic system. Although our convolution model 3.3 is a different model than DCM, we expect that our estimates from SCDN

for $\boldsymbol{x}$ and $\boldsymbol{\theta}$ should share the same patterns with their counterparts in DCM because of their biophysical interpretations.

Table 3.2: Comparison between estimation results of SCDN and DCM, Both computations are conducted in a cluster with 16 Intel CPU cores (at 2.6 GHz)

Method	# of Nodes	SNR	# of Stimuli	AUC (A)	Computation cost
SCDN	5	3	5	**0.81**	**2 minutes**
DCM	5	3	5	0.25	6 hours

We compute the estimation with both SCDN and DCM methods. To compare these two methods with the same conditions, we provide no additional information for the prior of $\boldsymbol{A}$. The experiment was conducted in SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/). As the estimation of DCM does not provide a sparse estimation as our method, we use the mean magnitude of estimated $\boldsymbol{A}$ across different subjects/trials to compute our evaluation criteria. Summary of computation time and estimation results are listed in Table 3.2. It is clear that our method provides a more accurate estimation and without any prior information, estimation of DCM is no better than a random guess.

In summary, the estimation of our algorithm achieves high accuracy with varying factors including different SNRs and number of stimuli using data simulated from convolution model 3.3. Estimation for $\boldsymbol{B}$ is much harder considering large number of entries in $\boldsymbol{B}$ and other reasons mentioned previously. Combining dimension reduction technique and our method is a potential way to further address the issue of relative low estimation accuracy of B. According to section 3.3.1, we show our estimation method is reliable for our own convolution model. Moreover, we simulated data from DCM, a different but well-accepted model. Based on the comparison with DCM in 3.3.2, our model estimation still performs well without any prior information which shows our convolution model for BOLD signals is reasonable and realistic.

3.4 DISCUSSION

We have developed a novel and simple sparse causal dynamic model to study brain activations and causal connections simultaneously from fMRI. SCDN is the first model that combines an ordinary differential equation model with a functional convolution observation model. Unlike DCM, our model is estimated by an optimization method and overcomes the requirements for prior knowledge and the size of networks. We show that our SCDN approach can recover most of the causal interaction parameters from large-scale brain networks efficiently. Moreover, because SCDN models the neuronal level connectivity, it provides a better biophysical interpretation than GCA and many other functional connectivity methods, without expensive computational cost. We illustrate that our estimation method is robust to varying factors, including model specification, SNR and the scale of brain networks.

It should be noted that it is still a challenging problem to estimate the influence of stimuli on the directional connection between these two regions, i.e. $\boldsymbol{B}$ in our model. It is thus an interesting future research question about how to address the issue of estimation of $\boldsymbol{B}$. Furthermore, We hope our model and estimation method will open up new avenues for research of large-scale brain causal connections which potentially help researchers to understand the underlying mechanism of the brain.

Part III

Canonical Correlation Analysis for Time Series

Chapter Four

Time-dependent Canonical Correlation Analysis for Multilevel Time Series

4.1 Introduction

Canonical Correlation Analysis (CCA) [Hotelling, 1936] is a powerful tool to analyze the relationship between two sets of variables. CCA can be regarded as an extension of ordinary correlation analysis, the difference being that CCA deals with multidimensional variables. It finds two linear transformations, one for each set of variables, that are optimal based on their correlations. It is an especially useful technique in data analysis as a dimensional reduction strategy that reduces the complexity of model space by calculating the combinations of variables that are maximally correlated. In an attempt to increase the flexibility for large dimensional date, several extensions of CCA have been proposed, including kernel Canonical Correlation Analysis (KCCA) [Akaho, 2001, Melzer et al., 2001], Sparse Canonical Correlation Analysis (SCCA) [Waaijenborg et al., 2008, Witten et al., 2009, Witten and Tibshirani, 2009, Hardoon and Shawe-Taylor, 2011, Chu et al., 2013]. Together, CCA-type methods have various applications including analysis of neuroimage, genomic data and information retrieval [Friman et al., 2001, Witten and Tibshirani, 2009, Hardoon et al., 2004].

For multilevel data, CCA has been studied extensively by multi-view CCA [Rupnik and Shawe-Taylor, 2010] and tensor CCA [Kim et al., 2007]. However, canonical correlation analysis of multivariate longitudinal data with multiple observations has received considerably less attention, despite its importance for practical data analysis. This setting arises from a wide range of applications, for instance, functional magnetic resonance imaging (fMRI) and the financial market contain multivariate time-varying observations. To understand the coupling dynamics of two sets of variables with time-stamped observations or to incorporate temporal structures, a few methods have been proposed. For instance, [Friman et al., 2002], maximized the

auto-correlation of fMRI time series, and [Bießmann et al., 2010] proposed to use KCCA to maximize the correlation between the two data sources over a certain time window. Although these approaches do incorporate the temporal dependencies to some extent, the temporal dynamics of consecutive linear transformations (vectors) have not been considered explicitly. Furthermore, these studies focus on performing canonical correlation analysis on two sets of variables with time lags.

A simple method to obtain the dynamic coupling between two sets of variables with time-stamped observation is to apply sparse CCA (for high dimensional data) or KCCA for each timestamp and then compare the vectors. However, in this way, we will lose temporal information and possibly reach the wrong conclusion about temporal dynamics. One such example will be discussed in detail later in this paper. Although one can adopt fused lasso penalty [Tibshirani et al., 2005] directly, more challenges will be raised in this case. First, CCA for high dimensional data with multiple time-varying observations will become computational expensive (with hundreds of non-convex constraints and many possible local optimums). Temporal incoherence [Bießmann et al., 2010] is another severe problem. For example, if we have w_1 and w_2 as our two canonical vectors, according to the definition of CCA problem, the correlation is also maximized by another two vectors $-w_1$ and $-w_2$. There is no guarantee that we get the same absolute sign for canonical vectors of two adjacent timestamps even if the data from these two timestamps are the same, especially when optimizing non-convex and non-smooth objective functions for these problems.

In order to solve the aforementioned problems, we propose a novel method for inferring the dynamic dependence between two sets of variables. This method integrates the SVD characterized formulation of all solutions of CCA [Chu et al., 2013] and the fused lasso regularization [Tibshirani et al., 2005] in a unified optimization

framework. We introduce a convex optimization problem which can be solved efficiently. There still exists a time incoherence problem in this formulation which we will discuss how to solve later in this paper. We note that since the focus of the paper is on introducing temporal structure in the CCA framework, we will only consider experiments of the first pair of canonical vectors. TDCCA and our algorithm can also be applied in the situations of multiple pairs of canonical vectors.

We summarize our contributions as follows:

- We incorporate temporal dependencies with (sparse) canonical correlation analysis using a convex formulation.
- We propose a fast parallelizable algorithm (Alternating Direction Method of Multipliers) and derive a closed-form ADMM updates to solve the non-smooth objective function.
- Our proposal provides a heuristic method to solve the time incoherence problem that exists in canonical vectors of adjacent timestamps.
- Experimental results on both two different simulated datasets show the effectiveness and accuracy of our method compared with static CCA. The experiments on the real dataset illustrate potential applications of our method for analyzing longitudinal data [Verbeke et al., 2014].

4.2 Canonical Correlation Analysis

We first review the standard canonical correlation analysis problem as

$$\begin{aligned} \text{minimize} \quad & \|XW_x - YW_y\|_F \\ \text{subject to} \quad & W_x^T X^T X W_x = I \\ & W_y^T Y^T Y W_y = I \end{aligned} \tag{4.1}$$

where $X \in R^{n \times d_1}$, $Y \in R^{n \times d_2}$, $W_x \in R^{d_1 \times l}$ and $W_y \in R^{d_2 \times l}$. Let $r = rank(X)$, $s = rank(Y)$ and $t = min(r, s)$. l is the number of pairs of canonical vectors we attempt to compute. d_1 and d_2 are the dimension of features for X and Y. n is the number of observations. We assume both X and Y are column centered. Under our temporal setting, X and Y are the observations at the same time point, see more details later.

Theorem 4 characterizes the solution of (4.1) by a SVD approach. Let us consider the SVD of X and Y,

$$X = Q_1[\Sigma_1, 0][U_1, U_2]^T = Q_1 \Sigma_1 U_1^T$$

$$Y = Q_2[\Sigma_2, 0][V_1, V_2]^T = Q_2 \Sigma_2 V_1^T$$

where $U_1 \in R^{d_1 \times r}$, $U_2 \in R^{d_1 \times (d_1 - r)}$, $\Sigma_1 \in R^{r \times r}$, $Q_1 \in R^{n \times r}$, $V_1 \in R^{d_2 \times s}$, $V_2 \in R^{d_2 \times (d_2 - s)}$, $\Sigma_2 \in R^{s \times s}$ and $Q_2 \in R^{n \times s}$. Furthermore, we consider the SVD of $Q_1^T Q_2$ as $Q_1^T Q_2 = P_1 \Sigma P_2^T$ where $P_1 \in R^{r \times r}$, $P_2 \in R^{s \times s}$ and $\Sigma \in R^{r \times s}$. Denote the distinct eigenvalues of $Q_1^T Q_2$ as $\sigma_1 > \sigma_2 > ... > \sigma_q > 0$ with multiplicity for these q eigenvalues being m_1, ..., m_q. $\alpha_k = \sum_{i=1}^{k} m_i$.

The following theorem from [Chu et al., 2013] shows the conditions for (W_x, W_y)

which will be used later.

Theorem 4. *[Chu et al., 2013] If $l = \sum_{i=1}^{k} m_i$ for some $1 \leq k \leq q$, then (W_x, W_y) is a solution of optimization problem (4.1) if and only if*

$$\begin{aligned} W_x &= U_1 \Sigma_1^{-1} P_1(:, 1:l) W + U_2 F_1 \\ W_y &= V_1 \Sigma_2^{-1} P_2(:, 1:l) W + V_2 F_2 \end{aligned} \tag{4.2}$$

where $W \in R^{l \times l}$ is orthogonal, $F_1 \in R^{(d_1 - r) \times l}$ and $F_2 \in R^{(d_2 - s) \times l}$ are arbitrary.

4.3 Methodology

Before we introduce our proposed method TDCCA, we provide some motivation for our framework, which is related to sparse CCA. If $l = \sum_{i=1}^{k} m_i$ for some $1 \leq k \leq q$, the sparse canonical correlation analysis can be stated as solving the following problem,

$$\begin{aligned} \text{minimize} \quad & \|W_x\|_{l_1} + \|W_y\|_{l_1} \\ \text{subject to} \quad & W_x = U_1 \Sigma_1^{-1} P_1(:, 1:l) W + U_2 F_1 \\ & W_y = V_1 \Sigma_2^{-1} P_2(:, 1:l) W + V_2 F_2 \end{aligned} \tag{4.3}$$

where $\| * \|_{l_1}$ is defined with element-wise l_1 penalty. This formulation is non-convex. A simplified formulation of the above is

$$\begin{aligned} \text{minimize} \quad & \|W_x\|_{l_1} + \|W_y\|_{l_1} \\ \text{subject to} \quad & U_1^T W_x = \Sigma_1^{-1} P_1(:, 1:l) \\ & V_1^T W_y = \Sigma_2^{-1} P_2(:, 1:l) \end{aligned} \tag{4.4}$$

There is a connection between the simplified formulation of sparse CCA problem (4.4) and (4.3). Denote the optimal value (feasible region) of problem (4.3) and

(4.4) as $M_s(\Omega_s)$ and $M(\Omega)$ respectively.

Theorem 5. *If* $l = \sum_{i=1}^{k} m_i$ *for some* $1 \le k \le q$, *let* $C = \frac{1}{\sqrt{l}}$, *then* $CM \le M_s \le M$

Proof. First it is easy to see

$$W_x = U_1 \Sigma_1^{-1} P_1(:, 1:l) W + U_2 F_1$$
$$W_y = V_1 \Sigma_2^{-1} P_2(:, 1:l) W + V_2 F_2$$

is equivalent to

$$U_1^T W_x = \Sigma_1^{-1} P_1(:, 1:l) W$$
$$V_1^T W_y = \Sigma_2^{-1} P_2(:, 1:l) W$$

By taking $W = I$, we get $M_s \le M$.

On the other hand, for W_f and W_x^f which satisfy (4.5),

$$\begin{aligned} U_1^T W_x^f &= \Sigma_1^{-1} P_1(:, 1:l) W_f \\ V_1^T W_y^f &= \Sigma_2^{-1} P_2(:, 1:l) W_f \end{aligned} \tag{4.5}$$

we get

$$U_1^T W_x^f W_f^{-1} = \Sigma_1^{-1} P_1(:, 1:l)$$
$$V_1^T W_y^f W_f^{-1} = \Sigma_2^{-1} P_2(:, 1:l)$$

because W_f is orthogonal. This implies $(W_x^f W_f^{-1}, W_y^f W_f^{-1}) \in \Omega$. Furthermore, for all $W_x, W_y \in \Omega$ and W orthogonal,

$$U_1^T W_x = \Sigma_1^{-1} P_1(:, 1:l)$$
$$V_1^T W_y = \Sigma_2^{-1} P_2(:, 1:l)$$

$\Longrightarrow$

$$U_1^T W_x W = \Sigma_1^{-1} P_1(:, 1:l) W$$
$$V_1^T W_y W = \Sigma_2^{-1} P_2(:, 1:l) W$$

This means $(W_x W, W_y W, W) \in \Omega_s$. Specifically, we can take $W_x = W_x^f W_f^{-1}$, $W_y = W_y^f W_f^{-1}$ and $W = W_f$ for all $(W_x^f, W_y^f, W_f) \in \Omega_s$. This implies $\{(W_x W, W_y W, W) | (W_x, W_y) \in \Omega \, , W \text{ is orthogonal}\} = \Omega_s$ and thus problem (4.3) is equivalent to (4.6)

$$\begin{aligned} \text{minimize} \quad & \|W_x W\|_{l_1} + \|W_y W\|_{l_1} \\ \text{subject to} \quad & U_1^T W_x = \Sigma_1^{-1} P_1(:, 1:l) \\ & V_1^T W_y = \Sigma_2^{-1} P_2(:, 1:l) \\ & W \in R^{l \times l} \quad \text{orthogonal} \end{aligned} \tag{4.6}$$

Based on the equivalences of norm of finite dimensional spaces and the orthogonality of W, we have

$$\begin{aligned} \frac{1}{\sqrt{l}} \|W_x\|_{l_1} = \frac{1}{\sqrt{l}} \sum_i \|W_x(i,:)\|_{l_1} \leq \sum_i \|W_x(i,:)\|_{l2} = \sum_i \|W_x(i,:) W\|_{l_2} \\ \leq \sum_i \|W_x(i,:) W\|_{l_1} = \|W_x W\|_{l_1} \end{aligned} \tag{4.7}$$

We can get similar result for W_y. Inequality (4.7) implies $CM \leq M_s$ and we already know $M_s \leq M$, thus $CM \leq M_s \leq M$ where $C = \frac{1}{\sqrt{l}}$ $\square$

4.3.1 Problem Formulation

Let's now consider time-dependent views of column centered data $X_t \in R^{n_t \times d_1}$ and $Y_t \in R^{n_t \times d_2}$ for $t \in [1, 2, ..., T]$. We attempt to analyze dynamic coupling canonical vectors $W_{xt} \in R^{d_1 \times l}$ and $W_{yt} \in R^{d_2 \times l}$ in a CCA framework which incorporate temporal information of these time-stamped observations. It is worth mentioning that our data X_t or Y_t does not have to be the observations from the same time point. They can be selected using a sliding window approach. In particular, a temporal window with length W, is chosen, and within the temporal interval that it spans (from time t=1 to time t=W), the first set of data are selected as X_1 and Y_1. Then, the window is shifted by a step T, and the same data extraction procedure is repeated over the time interval $[1+T, W+T]$. This process is iterated until the window spans the end part of the time series. Motivated by Theorem 4 and Theorem 5, we can formulate the problem as

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_t(\|W_{xt}\|_{l1} + \|W_{yt}\|_{l1}) + \lambda \cdot \sum_t(\|W_{xt} - W_{x,t-1}\|_{l1} + \|W_{yt} - W_{y,t-1}\|_{l1}) \\ \text{subject to} \quad & U_{1t}^T W_{xt} = \Sigma_{1t}^{-1} P_{1t}(:, 1:l) \\ & V_{1t}^T W_{yt} = \Sigma_{2t}^{-1} P_{2t}(:, 1:l) \end{aligned} \tag{4.8}$$

where

$$X_t = Q_{1t}[\Sigma_{1t}, 0][U_{1t}, U_{2t}]^T = Q_{1t}\Sigma_{1t}U_{1t}^T$$

$$Y_t = Q_{2t}[\Sigma_{2t}, 0][V_{1t}, V_{2t}]^T = Q_{2t}\Sigma_{2t}V_{1t}^T$$

$$Q_{1t}^T Q_{2t} = P_{1t}\Sigma_t P_{2t}^T$$

are SVDs for each t.

By allowing the relaxation of the constraints, we propose the TDCCA by opti-

mizing the following objective function (4.9).

$$\frac{1}{2}\sum_t(\|U_t^T W_{xt} - \Sigma_{1t}^{-1}P_{1t}(:,1:l)\|_F^2 + \lambda\sum_t\|W_{xt}\|_{l1} + \mu\sum_t\|W_{xt} - W_{x,t-1}\|_{l1} +$$
$$\frac{1}{2}\sum_t(\|V_t^T W_{yt} - \Sigma_{2t}^{-1}P_{2t}(:,1:l)\|_F^2 + \lambda\sum_t\|W_{yt}\|_{l1} + \mu\sum_t\|W_{yt} - W_{y,t-1}\|_{l1} \quad (4.9)$$

It is clear that the problem we formulate is a convex problem, which avoids the constraints $W_{xt}^T X_t^T X_t W_{xt} = I$ and $W_{yt}^T Y_t^T Y_t W_{yt} = I$ in CCA framework.

4.3.2 Optimization

We present an algorithm to optimize the objective function of TDCCA in (4.9). The estimation of W_x and W_y can be separated which makes it possible for parallel computing. In addition, the estimation of different pairs of canonical vectors can also be computed in parallel. For the ease of notation, we will ignore the dimension number inside of P_1 which represents $P_1(:, ll)$ if we try to estimate ll-th pair of canonical vectors. Without loss of generality, we will only discuss the algorithm for $W_x \in R^{d_1 \times T}$, the first pair of canonical vectors.

By separating (4.9), we get

$$l_x = \frac{1}{2}\sum_t(\|U_t^T W_x(:,t) - \Sigma_{1t}^{-1}P_{1t}\|_F^2 + \lambda\|W_x\|_{l1} + \mu\|W_x D\|_{l_1} \quad (4.10)$$

where D is the time differencing operator, i.e.

$$D = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \dots & -1 \end{bmatrix} \in R^{T \times (T-1)}$$

Replacing W_x with $\widehat{W}_x$ and $\widetilde{W}_x$, (4.10) is equivalent to (4.11)

$$l_x = \frac{1}{2} \sum_t (\|U_t^T W_x(:,t) - \Sigma_{1t}^{-1} P_{1t}\|_F^2 + \lambda \|\widetilde{W}_x\|_{l_1} + \mu \|\widehat{W}_x D\|_{l1} \tag{4.11}$$

subject to $W_{xt} = \widehat{W}_{xt}$ and $W_{xt} = \widetilde{W}_{xt}$.

Now we can adopt ADMM [Boyd et al., 2011] to optimize (4.11). We first write down the augmented Lagrangian for (4.11),

$$\begin{aligned} l_x = \frac{1}{2}(\sum_t \|U_{1t}^T W_x(:,t) - \Sigma_{1t}^{-1} P_{1t}\|_F^2 + \lambda \|\widetilde{W}_x\|_{l_1} + \mu \|\widehat{W}_x D\|_{l_1} + tr(\Theta^T(W_x - \widetilde{W}_x)) \\ + tr(\Phi^T(W_x - \widehat{W}_x)) + \frac{\nu}{2}(\|W_x - \widetilde{W}_x\|_F^2 + \|W_x - \widehat{W}_x\|_F^2) \end{aligned} \tag{4.12}$$

It can be solved by alternatively updating the five variables W_x, $\widetilde{W}_x$, $\widehat{W}_x$, Θ and Φ.

1. Fix $\widetilde{W}_x$, $\widehat{W}_x$, Θ and Φ, we get

$$\begin{aligned} \text{minimize} \quad & (\tfrac{1}{2} \textstyle\sum_t (\|U_{1t}^T W_x(:,t) - \Sigma_{1t}^{-1} P_{1t}\|_F^2 + + \tfrac{\nu}{2}(\|W_x - \widehat{W}_x\|_F^2 \\ & + \|W_x - \widetilde{W}_x\|_F^2) + tr(\Phi^T(W_x - \widehat{W}_x)) + tr(\Theta^T(W_x - \widetilde{W}_x)) \end{aligned}$$

Simply by setting the derivative with respect to W_x to zero, we have

$$W_x(:,t) = (U_{1t}U_{1t}^T + 2\nu I)^{-1}(-\Phi(:,t) - \Theta(:,t) + U_{1t}\Sigma_{1t}^{-1}P_{1t} + \nu(\widehat{W}_i^k(:,t) + \widetilde{W}_i^k(:,t))) \quad (4.13)$$

2. Fix W_x, $\widehat{W}_x$, Θ and Φ, the problem is transformed to

$$\text{minimize} \quad \lambda\|\widetilde{W}_x\|_{l_1} + tr(\Theta^T(W_x - \widetilde{W}_x)) + \tfrac{\nu}{2}\|\widetilde{W}_x - W_x\|_F^2$$

we thus get

$$\widetilde{W}_x = \text{sign}(W_x + \frac{\Theta}{\nu})\max(|W_x + \frac{\Theta}{\nu}| - \frac{\lambda}{\nu}, 0) \quad (4.14)$$

3. Fix W_x, $\widehat{W}_x$, Θ and Φ, (4.12) becomes

$$\text{minimize} \quad \mu\|\widehat{W}_x D\|_{l_1} + tr(\Phi(W_x - \widehat{W}_x)) + \tfrac{\nu}{2}\|W_x - \widehat{W}_x\|_F^2)$$

which is equivalent to

$$\text{minimize} \quad \mu\|\widehat{W}_x D\|_{l_1} + \tfrac{\nu}{2}(\|\widehat{W}_x - W_x - \tfrac{\Phi}{\nu}\|_F^2 \quad (4.15)$$

It is a combination of 1-d fused lasso problems which can be solved exactly using dynamic programming method [Johnson, 2013] or a taut string principle [Davies and Kovac, 2001] (both linear time algorithm) in parallel.

The similar process can be applied to Θ and Φ.

4.

$$\Theta = \Theta + \nu(W_x - \widetilde{W}_x) \quad (4.16)$$

Algorithm 4 Algorithm of TDCCA method (TDCCA-1)

Input: X_t, U_{1t}, P_{1t}, Σ_{1t}, λ, μ, ν
Initialize W_x, $\widetilde{W}_x$, $\widehat{W}_x$, Θ and Φ
repeat
  Update W_x with (4.13)
  Update $\widehat{W}_x$ with (4.14)
  Update $\widetilde{W}_x$ with (4.15)
  Update Θ with (4.16)
  Update Φ with (4.17)
until convergence
Return W_x, $\widetilde{W}_x$, $\widehat{W}_x$

5.

$$\Phi = \Phi + \nu(W_x - \widehat{W}_x) \tag{4.17}$$

We summarize our algorithm in Algorithm 4 (TDCCA-1). It is worth noting that in our proposed algorithm, step 2 and 3 can run in parallel due to that fact that the computation of $\widehat{W}_x$ and $\widetilde{W}_x$ only depends on W_x.

4.3.3 Time Incoherence

In (4.8), there exists a problem called time incoherence. The reason for this problem is that the original constraint in Theorem 4 is $U_1^T W_x = \Sigma_1^{-1} P_1(:, 1:l)W$ where W is orthogonal. For $l = 1$, $W \in R^{1\times 1}$ can be either 1 or -1, which causes the sign ambiguity. In previous section, we ignored the constraint on W, i.e the sign for the case $l = 1$. The problem could be tackled by adding integer variable $b_t \in \{-1, 1\}$

and we get a new optimization problem,

$$\frac{1}{2}\sum_t(\|U_t^T W_{xt} - \Sigma_{1t}^{-1} P_{1t}(:,1:l)b_t\|_F^2 + \lambda\sum_t \|W_{xt}\|_{l1} + \mu\sum_t \|W_{xt} - W_{x,t-1}\|_{l1} + \\ \frac{1}{2}\sum_t(\|V_t^T W_{yt} - \Sigma_{2t}^{-1} P_{2t}(:,1:l)b_t\|_F^2 + \lambda\sum_t \|W_{yt}\|_{l1} + \mu\sum_t \|W_{yt} - W_{y,t-1}\|_{l1} \tag{4.18}$$

One naive way to solve (4.18) is to compute the optimal value for every choice of sequence $[b_1, ..., b_T]$. The computational burden will increase exponentially. Problem (4.18) is a non-convex mixed integer problem which is generally hard to solve.

Instead of diving into the non-convex problem, we propose a three-step approach. First, we will use Algorithm 4 with a very small (1e-10 chosen in our experiments) μ and thus the temporal difference is not penalized. This step allows us to obtain an initial estimation of W_x. Then we will change the sign of P_1 and P_2 according to whether the condition (4.19) is satisfied. Finally, we run Algorithm 4 again with μ chosen by grid search and obtain the final estimations. This method allows us to detect those SVD results with the incoherent sign and thus the temporal consistency is achieved. The intuition behind this approach is that the optimal value of (4.9) continuously depends on μ. The final algorithm is summarized in Algorithm 5.

$$\|W_x(:,t-1) - W_x(:,t)\|_{l_1} + \|W_y(:,t-1) - W_y(:,t)\|_{l_1} \\ > \|W_x(:,t-1) + W_x(:,t)\|_{l_1} + \|W_y(:,t-1) + W_y(:,t)\|_{l_1} \tag{4.19}$$

As the Algorithm 4 is computationally efficient, Algorithm 5 is still efficient, considering we will fix μ in the first step. Deflation method [Witten and Tibshirani, 2009, Shawe-Taylor and Cristianini, 2004, Hardoon and Shawe-Taylor, 2011] can also

Algorithm 5 Algorithm of TDCCA method

Input: X_t, Y_t, U_{1t}, V_{1t}, P_{1t}, Σ_{1t}, P_{2t}, Σ_{2t}, λ, μ, ν
First step: $\widehat{W}_x$, $\widehat{W}_y$ = TDCCA-1(Input(with $\mu = e^{-10}$))
Second step: Revise the sign of P_1 and P_2
for $t \in [2, ..., T]$ **do**
  if (4.19) is True **then**
    $P_{1t} = -P_{1t}$ and $P_{2t} = -P_{2t}$
  end if
end for
Third step: W_x, W_y = TDCCA-1(Input)

be easily combined with our algorithm after calculating each pair of canonical vectors to acquire multiple pairs of canonical vectors.

4.3.4 Tuning Parameter Selection

In our method, TDCCA contains two tuning parameters λ and μ which determine the sparsity and continuity (along temporal dimension) of canonical vectors. We propose a cross-validation approach as follows: we partition our data as training and validation data, and then we select the tuning parameters that maximize the canonical correlation on the validation data, plugging the canonical vectors solved from the training data. We also apply the grid search to determine the optimal values of λ and μ.

4.3.5 Algorithm Analysis

The convergence of ADMM under certain conditions has been analyzed and proven in previous papers [He and Yuan, 2012, Chen et al., 2016, Hong and Luo, 2017]. Our optimization problem satisfies the conditions and thus our algorithm is guaranteed

to converge to a non-empty solution set if it exists. In each iteration, the major computation burden is in step 1. One iteration of step 2, 3, 4, 5 is $O(dT)$ where d is dimension of feature space. For the first step, the inverse of $(U_{1t}U_{1t}^T + 2\nu I)$ can be precomputed before the iterations. For each iteration, time complexity of step 1 (matrix multiplication) is $O(d^2T)$. Another feature of our algorithm is both step 1 and 3 can run in parallel along the temporal dimension.

4.4 Experiments

In this section, we evaluate the performance of the proposed method on two simulated datasets and real functional magnetic resonance imaging (fMRI) data. We compare our method with static sparse CCA [Witten and Tibshirani, 2009]. For the static sparse CCA, we treat each t as an independent problem. We use the R package called PMA for SCCA, which is publicly available at https://cran.r-project.org/web/packages/PMA/index.html. The package for our method is available at https://github.com/xuefeicao/tdcca.

4.4.1 Simulations

Table 4.1: Simulation 1 results from 50 independent trials

n	d	T	Method	CDR	F1	TDR	Cosine of Angle
100	40	100	TDCCA	0.0021	1.0000	97.7767	0.9800
			SCCA	0.0025	0.9488	5.5884	0.9693
100	100	100	TDCCA	0.0008	1.0000	89.8201	0.9730
			SCCA	0.0021	0.8192	2.1969	0.8439
100	400	100	TDCCA	0.0002	0.9977	67.5984	0.9460
			SCCA	0.0025	0.5771	1.6001	0.6164

Table 4.2: Simulation 2 results from 50 independent trials

n	d	T	Method	CDR	F1	TDR	Cosine of Angle
100	40	100	TDCCA	0.0599	0.9896	74.8900	0.9297
			SCCA	0.0398	0.8177	2.2071	0.9039
100	100	100	TDCCA	0.0339	0.9647	53.1478	0.9642
			SCCA	0.0412	0.7317	1.3389	0.7692

We introduce four metrics to measure the accuracy of our estimates in simulations.

- **Correlation Deviation Ratio (CDR)**: This evaluates the capability of our method to recover the true correlation between two sets of variables. It is defined as the ratio of l_1 distance between estimated correlation and true correlation to the true correlation.

- **F1 score**: This measures the ability of our method to capture the true pattern of related variables.

- **Cosine of Angle between estimated and real canonical vectors**: This measures the similarity of our estimation and real canonical vectors. It is defined as the absolute value of cosine angle between two vectors.

- **Temporal Deviation Ratio (TDR)**: The temporal deviation defined as $\|W_x(:,t) - W_x(:,t-1)\|_{l_1} + \|W_y(:,t) - W_y(:,t-1)\|_{l_1}$, illustrates how much the estimation changes at each time step. This value is the ratio of temporal deviation at change point (in our simulation, for simplicity, only one change point is included) to the average temporal deviation value of all time points. This metric serves the purpose of testing the ability of our method to detect temporal dynamics. We notice that SCCA method does not distinguish the absolute sign of the canonical vectors ($-W$ or W can both be solutions). To achieve a fair comparison, we alter the canonical vector W_t to $-W_t$ obtained

in SCCA method at time t if (4.19) is satisfied.

All the results reported in this section are the averages along temporal dimension. We denote $W_{xt} = W_x(:,t)$ and $W_{yt} = W_y(:,t)$. $X_t \in R^{d_1}$ and $Y_t \in R^{d_2}$. We fix $T = 100$ and $d_1 = d_2$ for simplicity. Let $d = d_1 + d_2$

4.4.1.1 Simulation 1

In this simulation, we generate our data according to the following model.

$$\begin{aligned} X_t &= (W_{xt} + \epsilon_{1t})u_t + \eta_{1t} \\ Y_t &= (W_{yt} + \epsilon_{2t})u_t + \eta_{2t} \end{aligned} \tag{4.20}$$

where $u_t \sim N(0,1)$, $\epsilon_{it} \sim N(0, 0.1^2)$ and $\eta_{it} \sim N(0, 0.1^2)$ for $i = 1, 2$. For $t \leq 50$

$$W_{xt} = (\underbrace{1, ..., 1}_{d_1/4}, \underbrace{-1, ..., -1}_{d_1/4}, 0, ..., 0)$$

$$W_{yt} = (0, ..., 0, \underbrace{1, ..., 1}_{d_2/4}, \underbrace{-1, ..., -1}_{d_2/4})$$

For $t > 50$,

$$W_{xt} = (\underbrace{0, ..., 0}_{d_1/4}, \underbrace{-1, ..., -1}_{d_1/4}, \underbrace{1, ..., 1}_{d_1/4}, 0, ..., 0)$$

$$W_{yt} = (\underbrace{1, ..., 1}_{d_2/4}, 0, ..., 0, \underbrace{-1, ..., -1}_{d_2/4})$$

From the model, we can see that for $t \leq 50$, the first half variables of X_t and second half variables of Y_t are correlated, while for $t > 50$, variables of X_t located in

$[\frac{1}{4}d_1, \frac{3}{4}d_1]$ are correlated with variables of Y_t located in $[1, \frac{1}{4}d_2]$ and $[\frac{3}{4}d_2, d_2]$. To test our algorithm, we conducted the estimation with different settings:

- $n = 100$, $d = 40$
- $n = 100$, $d = 100$
- $n = 100$, $d = 400$

where n is the number of samples. We summarize our results of 50 independent trials in Table 5.1.

4.4.1.2 Simulation 2

In this section, we employ data generated from a more complicated model called the single canonical pair model [Chen et al., 2013]. The model is described in (4.21). We used the same W_{xt} and W_{yt} as in the first simulation.

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \lambda \Sigma_1 W_{xt} W_{yt}^T \Sigma_2 \\ \lambda \Sigma_2 W_{yt} W_{xt}^T \Sigma_1 & \Sigma_2 \end{pmatrix} \right] \tag{4.21}$$

where $0 < \lambda \leq 1$, $W_{xt}^T \Sigma_1 W_{xt} = 1$ and $W_{xt}^T \Sigma_2 W_{xt} = 1$. Let $\lambda = 0.9$. We define $\Sigma_1 = \Sigma_2 = (\sigma_{ij})_{ij}$ where $\sigma_{ij} = c \times 0.3^{|i-j|}$ which indicates covariance has a certain rate of decay. The scaling factor c is obtained by normalization. In addition, we add independent noise ($\epsilon \sim N(0, \frac{c}{100})$) to the generated data. It is easily to verify that for model (4.21), W_{xt} and W_{yt} are first pair of canonical vectors, which maximizes the correlation between $X_t^T W_{xt}$ and $Y_t^T W_{yt}$. Furthermore, the corresponding correlation

is λ. We note that our method TDCCA does not depend heavily on the Gaussian covariance assumption. However, in [Chen et al., 2013], their Sparse CCA method utilizes the model structure (4.21) ($\Sigma_{12} = \Sigma_1 W_x W_y^T \Sigma_2$) explicitly and then get an estimation of W_x and W_y directly where Σ_{12} is cross-covariance between X and Y. We used the following settings in this simulation,

- $n = 100$, $d = 40$
- $n = 100$, $d = 100$

where n is the number of samples. Table 5.2 showed the averaged results of 50 independent trials.

4.4.1.3 Simulation Results

We show results for both models in Table 5.1 and Table 5.2. In terms of F1 score, Temporal Deviation Ratio and Cosine of Angle between estimated and real canonical vectors, our TDCCA approach significantly outperforms SCCA method (i.e the CCA method without considering time series structure). The F1 score of TDCCA stays above 0.9 in different settings of two simulations while the F1 score of SCCA can be less than 0.6 when the ratio $\frac{\log(d)}{n}$ increases. The cosine of the angle of our approach is close to 1 which indicates the high similarity between estimated and real canonical vectors. In addition, the temporal deviation ratio of TDCCA is up to 40 times higher than the SCCA method, which shows a big advantage of TDCCA in detecting change points. Furthermore, our estimated value of correlation is closer to the true correlation of simulated data than the SCCA method.

4.4.2 Canonical Correlation Variation for resting-state fMRI

In this section, we apply our method on the resting-state fMRI data and we propose that the canonical correlation variation (CCV), a new metric obtained from our method can provide clues for the connectivity patterns that transfer and present aging features. Canonical correlation variability (CCV) is defined as the standard deviation of time-dependent canonical correlation from our method.

Recent work has shown that functional connectivity is temporally dynamic and functional connectivity fluctuates across shorter time-windows for resting-state fMRI [Chang and Glover, 2010, Handwerker et al., 2012b]. Unlike conventional FC analysis, which assumes static connectivity over several minutes, the dynamic functional connectivity variation (FCV) is calculated as the standard variation of the dynamic FC series. In this approach, the stability of the FC fluctuation over time is quantitatively measured and compared between brain region pairs. The basic sliding window framework has been used widely and is repeatedly applied by researchers to investigate how functional brain dynamics relate to our cognitive abilities [Preti et al., 2017]. The age-related dynamic pattern of functional connectivity has been also explored in [Madhyastha and Grabowski, 2014, Chen et al., 2016, 2018].

Canonical correlation is another way to characterize the strength of the functional connectivity for each region pair which has been used to construct a region-level functional connectivity network for predicting major depressive disorder [Kang et al., 2016]. In our experiment, two groups of individuals (N = 156, ages 22–25 for the first group; N = 226, ages 31–35 for the second group) were recruited from the public data of the Human Connectome Project (`http://www.humanconnectomeproject.org/`). In particular, we are interested in the connectivity patterns inside of the default mode network (DMN) which contains Precuneus (pC), Posterior cingulate (PCC),

Ventral anterior cingulate (vACC), and Medial prefrontal cortex (mPFC). For each subject, we used a fixed-length rectangle window (width = 60 TRs) and the window was shifted by 2 TRs. These parameters are chosen based on the rule of choosing parameters for dynamic functional connectivity [Preti et al., 2017]. Thus we can obtain the time-dependent canonical correlation estimated for each pair of ROIs from our method for every subject from two groups.

Additionally, we use two popular measurements of connectivity pattern: static functional connectivity (FC), functional connectivity variation (FCV). As a baseline method, we compute the sparse canonical correlation for each rectangle window and calculate its standard deviation which we will call it CCVB in the remaining paper. The canonical correlation coefficient (CCC) for the entire time series is also included for each subject. These features are summarized in table 4.3.

Table 4.3: Features of connectivity pattern used in our experiment

Method	Description
CCV	Canonical Correlation Variation calculated from TDCCA
FC	Static Functional Connectivity (Fisher-transformed correlations) of entire time series
FCV	Functional Connectivity Variation using sliding window approach
CCVB	Canonical Correlation Variation calculated from SCCA for each sliding window
CCC	Canonical Correlation Coefficient (Fisher-transformed correlations) of entire time series

Table 4.4: Adjusted p–value for each pair of ROIs chosen from default mode network. Values in bold represent significant p–value with threshold 0.05.

ROI 1	ROI 2	CCV	FC	FCV	CCVB	CCC
vACC	pC	0.5197	0.8092	0.9452	0.9498	0.7419
vACC	PCC	0.0138	0.9725	0.9452	0.4273	0.7419
PCC	pC	0.5197	0.8092	0.9452	0.2464	0.7419
mPFC	vACC	0.4224	0.9725	0.9452	0.4273	0.7419
mPFC	pC	0.1593	0.8092	0.9452	0.2464	0.7419
mPFC	PCC	0.5197	0.9725	0.9452	0.2464	0.7419

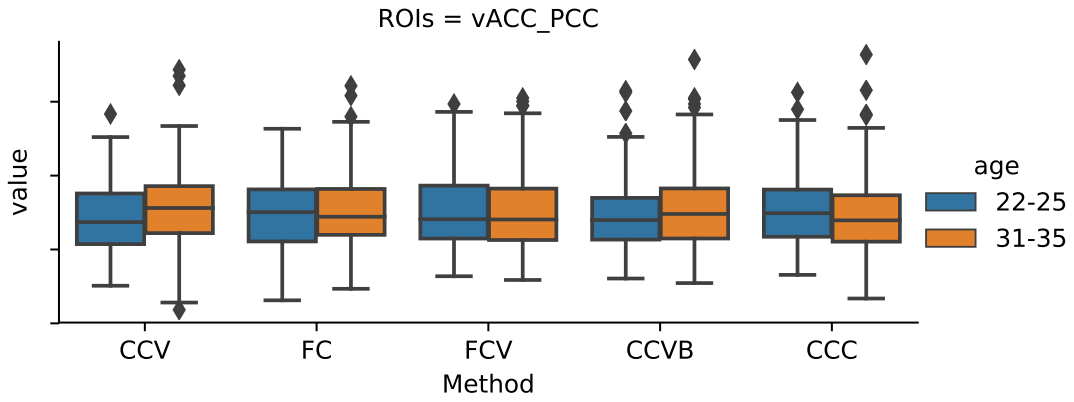

Figure 4.1: Group differences of different features for the ROI pair with a significant p–value.

Table 4.4 shows the p–value of two sample t-test of different features for each pair of ROIs compared between two different groups. Multiple testing correction can be performed using the FDR method [Benjamini and Hochberg, 1995, Benjamini and Yekutieli, 2001]. In this work, we followed the previous work [Zhao et al., 2018] and applied the method proposed in [Benjamini and Hochberg, 1995]. From table 4.4, we can see the only one significant difference for metrics calculated based on two different groups are from our proposed canonical correlation variation measurement. It is between vACC and PCC. Figure 4.1 illustrates the group differences of different features for the ROI pair with a significant p–value. The values are scaled for better

visualization. It shows CCV (between vACC and PCC) in the age group 31–35 is higher than the age group 22–25. This example shows promising results by applying CCV as a novel way to measure dynamic functional connectivity patterns using resting-state fMRI.

4.4.3 Task-based fMRI: motor task detection and feature extraction

In this section, we apply the TDCCA method to analyze task-related fMRI motor data obtained from the Human Connectome Project [Van Essen et al., 2013]. This example will show how our method can spot change points due to different tasks.

This motor task fMRI is composed of five most basic motor tasks including tapping left/right fingers, squeezing left/right toes and moving tongue. Participants were presented with visual cues which asked them to either tap their left or right fingers, or squeeze their left or right toes, or move their tongue to map motor areas. Each block of a movement type lasted 12 seconds (5 movements), and was preceded by a 3-second cue.

Based on the prior scientific findings on the motor task experiment [Thomas Yeo et al., 2011], we select six brain regions corresponding to these different tasks according to MNI coordinates: left/right hand coordinates ($\pm$41, 20, 62), left/right foot coordinates($\pm$6, 26, 76), tongue coordinates ($\pm$55, 4, 26), thalamus (MNI: -12, -13, 7). We extracted voxels around these coordinates, depending on the availability of voxels centered around these coordinates. Thus we combine data from these six regions as our one set of data. We set the length of the sliding window as 20 TRs according to the length of each task and the window was shifted by 1 TR (0.72 s).

Our TDCCA method is applied to a pair of subjects.

We should mention that brain electrical activity is not directly measured, instead, the human hemodynamic responses to the short period of neural activity are delayed in time. Thus fMRI measures the subsequent demand for oxygenated blood that follows about several seconds after the neuronal activations [Liao et al., 2002].

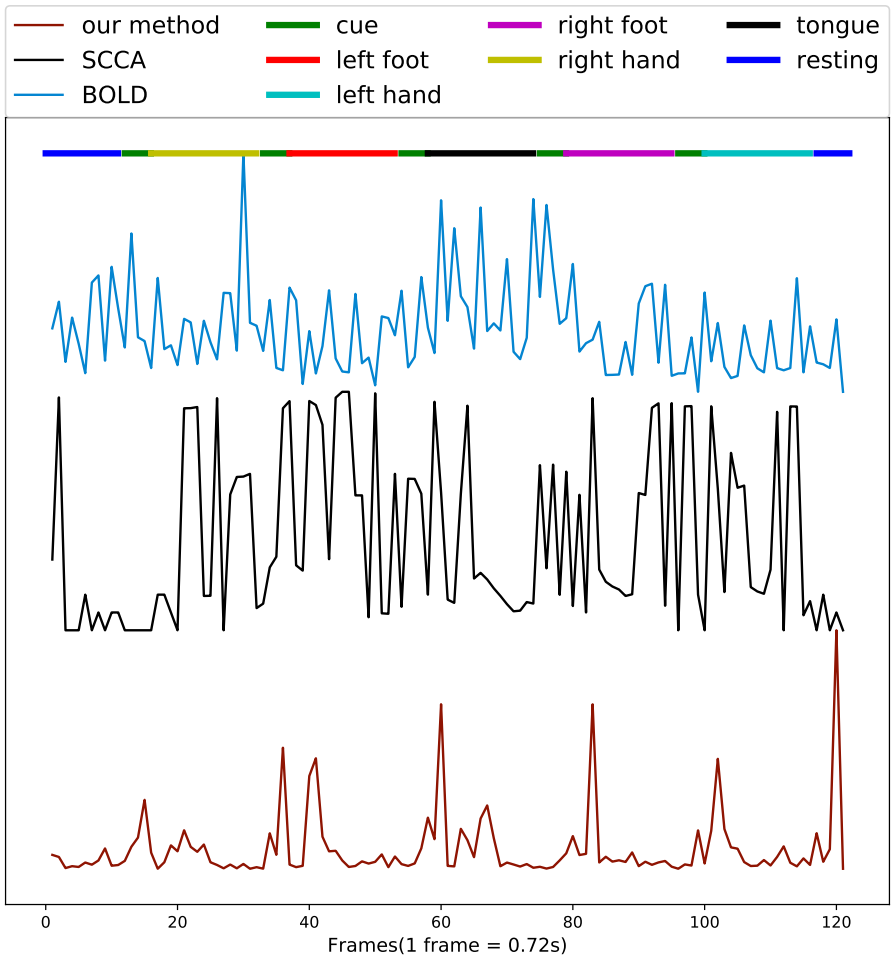

Figure 4.2: Plot of the temporal deviation of TDCCA, SCCA and original BOLD signal, in which TDCCA detects six significant shift of tasks. The straight line above the plot uses different colors to indicate the different tasks during the experiment.

We estimate the leading canonical vectors and elaborate on how the W_x varies with time periods of the different task activation. We also compared the SCCA (sparse CCA method) with TDCCA. Figure 4.2 shows the scaled temporal deviation $\|W(:,t)-W(:\,t-1)\|_{l_1}$ estimated from two methods as well as the BOLD variations. Figure 4.2 elaborates the ability of our approach in detecting temporal dynamics. From the results of SCCA and the original BOLD signal, one can barely see the dynamic change point for different motor tasks. However, our method detects six significant shift of tasks and a clear time delay from task commands to the peak of temporal deviation.

4.5 Conclusions and future work

In this paper, a convex framework for combining temporal structure with canonical correlation analysis is proposed. The proposed framework incorporates temporal information explicitly. Furthermore, our algorithm is computationally efficient with guaranteed convergence and has the advantage of parallel computing. Finally, we introduce a heuristic method to solve the time incoherence problem without using a mixed-integer optimization algorithm. The proposed method outperforms the (static) sparse CCA algorithm both in accuracy and ability to recover temporal variations. Our proposed canonical correlation variation (CCV) can also provide clues for brain connectivity patterns. Our method introduces an additional tool to determine change points and extract critical features in multivariate analysis. In future work, we will explore the theoretical property of our proposed algorithm. It would be also promising to apply our method to analyze multivariate longitudinal data from medical images.

Part IV

Covariance Regression Models for Brain Connectivity Analysis

Chapter Five

Common Vector Projection Regression for Covariance Analysis

5.1 Introduction

Estimating variances has played an important role in many fields, including finance, risk management and statistics [Chen and Conley, 2001, Xue et al., 2012]. Classic sample covariance matrix estimator has been used extensively in the statistical literature and is proven to be consistent when the number of features is fixed and the number of samples goes to infinity. More recently, some research efforts have been devoted to reducing the number of parameters to be estimated in the covariance matrix because the number of parameters grows rapidly with the increasing number of features. One such direction is to add sparse constraints on parameters of the covariance matrix or the inverse of the covariance matrix [Huang et al., 2006, Bickel et al., 2008, Friedman et al., 2008, Cai and Liu, 2011, Cai et al., 2011]. However, these methods focus on estimating covariance matrices instead of linking covariance matrices with auxiliary information (a vector of covariates, collected from each subject in the biostatistics area).

Hence, regression models for covariance matrices have been explored to make the connection between covariates and covariance matrices. In [Anderson et al., 1973, Szatrowski et al., 1980, Zwiernik et al., 2017], the covariance matrix has a linear structure and the resulting estimate of covariance is a function of covariates. Another direction of study connects covariances with covariates using specific model structures. For example, [Chiu et al., 1996] introduced a model where the elements of the logarithm of the covariance matrix is a linear combination of covariates. [Pourahmadi, 1999] proposed to use the Cholesky decomposition of the inverse of a covariance matrix to associate covariates. [Hoff and Niu, 2012] introduced a covariance regression model that parameterizes the covariance matrix of a multivariate response vector as a parsimonious quadratic function of explanatory

variables. [Fox and Dunson, 2015] reduced dimensionality and induced a flexible Bayesian nonparametric covariance regression model by using a lower-dimensional subspace representation. Later, [Franks and Hoff, 2016] proposed to explore a spiked covariance model for each group while sharing information about the space spanned by the group-level eigenvectors. However, it is unclear how this model can be extended to consider continuous covariates. [Zou et al., 2017] integrated the similarity concept and then explored a covariance regression model to directly quantify the relationship between the covariance and a linear combination of matrices induced by corresponding covariates. It is worth noting that to link with covariates, these methods treat the whole covariance matrix as outcomes. More recently, [Zhao et al., 2018] instead aims to find the projections for covariance matrices that are associated with the covariates, where the covariance matrix is not estimated.

Utilizing eigenvectors of the covariance matrix has been a common practice in different domains. In social networks, if all the entries of a covariance matrix are positive, its leading eigenvector gives the relative centrality score of the vertex in the network [Newman, 2016]. This has also been explored in the neuroimaging domain. [Lohmann et al., 2010] showed that eigenvector centrality is modulated by the state that the subjects were in and they also demonstrated that eigenvector centrality is a computationally efficient tool for capturing intrinsic neural architecture on a voxel-wise level. [Zuo et al., 2011] used resting-state functional magnetic resonance imaging data from 1003 healthy adults, and investigated multiple network centrality measures including eigenvector centrality. Another application is principal component analysis (PCA), which is usually used to reduce model dimension and create a more explainable model. PCA constructs relevant features by linearly transforming correlated variables (e.g. raw voxels in a brain scan) into a smaller number of uncorrelated variables, i.e. principal components. It has been successfully utilized in

extracting relevant features in neuroimaging domain [Caprihan et al., 2008, Mwangi et al., 2014]. To incorporate group-level differences, [Flury, 1984, 1986, 1988] introduced common principal components models, where the group-level covariances are simultaneously diagonalizable. [Boik, 2002] proposed a generalized version of these models using spectral decompositions.

Modeling the relationship between covariances and covariates is critical in different areas. In the financial domain, many researchers have shown that a covariance matrix of returns of stocks is affected by firms' fundamentals [Chan et al., 1998], which suggests that the covariance matrix can be partially explained by its associated relevant explanatory variables. Modeling the covariate-related covariance matrices is also an important topic in neuroimaging analysis. The variations in functional connectivity (correlation matrices after standardization) are associated with various subject-level explanatory variables (e.g. clinical status, disease information, cognitive behaviors, and mental health capacities) [Grady et al., 2003, Wang et al., 2007, Finn et al., 2015]. In the context of geostatistics, the influence of local effects in the correlation structure makes it unreliable to model partially distributed phenomena through an underlying stationary and isotropic spatial process. Thus it has been of interest to model nonstationarity in the spatial effects e.g. the latitude, longitude and altitude of the observation sites [Reich et al., 2011].

In this work, we propose a Common Vector Projection Regression (CVPR) model for multiple covariance matrix outcomes. Instead of focusing on estimating covariances directly in most of the existing works, we are interested in linking covariances and subject-level covariates or explainable variables. Similar to the recent Covariate Assisted Principal (CAP) regression model [Zhao et al., 2018], our model aims to identify linear projections to connect covariance matrices with auxiliary information. However, unlike the previous approach, our method targets to find the projection

that links eigenvectors and covariates directly. We can study the covariates-induced variations not only in covariance matrices but also in their eigenvectors which are important features of the overall network structure.

We organize our work as follows. In section 5.2 and 5.3, we introduce our model and investigate different objective functions. In section 5.4 and 5.5, we discuss how to optimize our objective function and show accuracy of our proposed method in different simulation settings. In section 5.6, we apply our method to a resting-state fMRI to investigate the relationship between brain networks and some common demographic information. Finally, in section 5.7 and 5.8, we illustrate the algorithmic and asymptotic properties of our model and optimization method.

5.2 Common Vector Projection Regression Model

Let $y_{it} \in R^p$ for $t = 1, 2, 3, \ldots, T_i$, $T = \max(T_1, ..., T_n)$ and $i = 1, 2, 3, \ldots, n$. $\{y_{it}\}_{t=1,...,T_i}$ are independent and identically distributed random variables with mean 0 and variance Σ_i. $\{y_{it}\}_{i,t}$ are independent. In our application to fMRI, $y_i = [y_{i1}, ..., y_{iT_i}]$ is the brain response of p regions with length of T_i. Σ_i depends on $x_i \in R^{l+1}$, the covariates, e.g. the age of the subjects through $v_i \in R^p$, the leading eigenvectors of Σ_i. We assume that there exists a vector $e \in R^p$ such that

$$|\langle v_i, e \rangle| = x_i^T \beta \tag{5.1}$$

where $\|v_i\|_{l2} = 1$, $\|e\|_{l2} = 1$ and we assume x_i are bounded such that $1 > ub > x_i^T\beta > lb > 0$. The value of the first dimension of x_i is fixed as 1. β is the model coefficient. We show one simple visualization of our model in Figure 5.1. Here we

give some intuition of our proposed model using an fMRI example. Suppose we have $e = (1, 0, \ldots, 0)$ and the leading eigenvectors calculated from the positive adjacency matrices of BOLD signals from multiple subjects. Then the first dimension of these eigenvectors [Newman, 2016] indicates the relative centrality (influence) of the first brain region to the brain network. If the brain network satisfies our model 5.1, the first brain region's influence can be represented by a linear model of the covariates of these subjects. This model is also motivated by previous work [Lin et al., 2012] where they introduced a projection regression model to assess the relationship between a multivariate phenotype and a set of covariates.

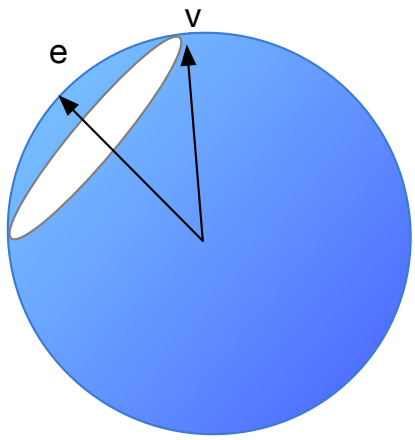

Figure 5.1: A Geometric Visualization of Our Model

5.3 Method

We propose to estimate the model parameters $V = [v_1, ..., v_n]^T$, e and β by optimizing the following loss function,

$$
\begin{aligned}
\text{minimize}_{e,\beta,v_i} \quad & \||Ve| - X\beta\|^2 - \lambda \textstyle\sum_i v_i^T \tilde{\Sigma}_i v_i \\
\text{subject to} \quad & \|e\|_{l^2}^2 = 1 \\
& \|v_i\|_{l^2} = 1 \\
& i = 1, .., n
\end{aligned}
\tag{5.2}
$$

where $\tilde{\Sigma}_i$ is the sample covariance of y_{it} for $i = 1, 2, ..., n$. However, the absolute value function used in (5.2) is non-smooth and thus makes it hard to optimize. It is to see this formulation is equivalent to (5.3).

$$
\begin{aligned}
\text{minimize}_{e,\beta,v_i} \quad & \|Ve - X\beta\|^2 - \lambda \textstyle\sum_i v_i^T \tilde{\Sigma}_i v_i \\
\text{subject to} \quad & \|e\|_{l^2}^2 = 1 \\
& Ve \geq 0 \\
& \|v_i\|_{l^2} = 1 \\
& i = 1, .., n
\end{aligned}
\tag{5.3}
$$

Although in (5.3), we don't have $|Ve|$ in the objective function, we are forced to add another coupled constraint $Ve \geq 0$, which contains two variables V and e. It is generally hard to optimize. One solution is to solve a relaxed problem (5.4),

$$
\begin{aligned}
\text{minimize}_{e,\beta,v_i,b_i} \quad & \|Ve * b - X\beta\|^2 - \lambda \textstyle\sum_i v_i^T \tilde{\Sigma}_i v_i \\
\text{subject to} \quad & \|e\|_{l^2}^2 = 1 \\
& \|v_i\|_{l^2} = 1 \\
& b_i \in -1, 1 \\
& i = 1, .., n
\end{aligned}
\tag{5.4}
$$

The global minimum of this formulation is smaller than that of (5.3) because we don't have the condition that $Ve \geq 0$. However, this formulation involves an integer variable and thus the common algorithm has no guaranteed convergence. To conquer

these problems, we find under the assumption that the global minimum of (5.3) satisfies $X\beta \geq 0$, formulation (5.5) is equivalent to (5.3).

$$\begin{aligned} \text{minimize}_{e,\beta,v_i} \quad & \|Ve - X\beta\|^2 - \lambda \textstyle\sum_i v_i^T \tilde{\Sigma}_i v_i \\ \text{subject to} \quad & \|e\|_{l^2}^2 = 1 \\ & X\beta \geq 0 \\ & ||v_i||_{l^2} = 1 \\ & i = 1, .., n \end{aligned} \tag{5.5}$$

Thus we will focus on optimizing (5.5). For simplicity, in the rest of our work, we define $\mathcal{L}$ and $\mathcal{F}$ in (5.6) and (5.7).

$$\mathcal{L} = \|Ve - X\beta\|^2 - \lambda \sum_i v_i^T \tilde{\Sigma}_i v_i \tag{5.6}$$

$$\mathcal{F} = \mathcal{L} + \mathbb{1}_{\{\|e\|_{l^2}^2 = 1, X\beta \geq 0, ||v_i\|_{l^2} = 1\}} \tag{5.7}$$

5.4 Algorithm

Problem (5.5) is non-convex. We propose to solve these optimization problems by block coordinate updates. In particular, for problem (5.5), we explore block prox-linear (BPL) method which updates a block of variables at each iteration by minimizing a prox-linear surrogate function [Xu and Yin, 2017].

In general, for an optimization problem (5.8),

$$\text{minimize}_x \quad f(x_1, x_2, ..., x_s) + \textstyle\sum_i r(x_i) \tag{5.8}$$

where f is continuously differentiable, functions r_i, i = 1, $\cdots$, s, are proximable, At

each iteration k, a block b_k is chosen. BPL algorithm updates x_i^k by (5.9) if $i = b_k$

$$x_i^k = \underset{x_i}{\operatorname{argmin}} \langle \nabla_{x_i} f(x_{\neq i}^{k-1}, x_i^{k-1}), \quad x_i - x_i^{k-1} \rangle + L_k \|x_i^k - x_i^{k-1}\| + r(x_i) \tag{5.9}$$

where L_k is the Lipshitz constant of $\nabla_{x_i} f(x_{\neq i}^{k-1}, x_i)$ about x_i

The detailed algorithm applied for our problem (5.5) is illustrated in Algorithm 6 where $\tilde{\Sigma}_i$ is the sample covariance of y_{it} for $i = 1, 2, ..., n$.

Algorithm 6 Algorithm of CVPR method (5.5)

Input: X, $\tilde{\Sigma}_i$, λ
Initialize e^0, β^0, V^0 such that $X\beta^0 \geq 0$
repeat
  Update e^k with (5.10)
  Update β^k with (5.11)
  Update v_i^k with (5.12) for $i = 1, 2, ..., n$
until convergence
Return V^k, e^k, β^k

$$\begin{aligned} e^k = \operatorname{argmin}_e \quad & \langle 2{V^{k-1}}^T V^{k-1} e^{k-1} - 2{V^{k-1}}^T X \beta^{k-1}, \quad e - e^{k-1} \rangle + L_{ek} \|e - e^{k-1}\|^2 \\ \text{subject to} \quad & \|e\|_{l^2}^2 = 1 \end{aligned} \tag{5.10}$$

where L_{ek} is the Lipschitz constant of $2{V^{k-1}}^T V^{k-1} e - 2{\beta^{k-1}}^T X^T V^{k-1}$ with respect to e.

$$\begin{aligned} \beta^k = \operatorname{argmin}_\beta \quad & \langle 2X^T X \beta^{k-1} - 2X^T V^{k-1} e^k, \quad \beta - \beta^{k-1} \rangle + L_{\beta k} \|\beta - \beta^{k-1}\|^2 \\ \text{subject to} \quad & X\beta \geq 0 \end{aligned} \tag{5.11}$$

where $L_{\beta k}$ is the Lipschitz constant of $2X^T X\beta - 2X^T V^{k-1} e^k$ with respect to β. More details on exact solution for solving (5.11) can be found in Rutkowski [2017].

$$
\begin{aligned}
v_i^k = \text{argmin}_{v_i} \quad & \langle 2e^k e^{k^T} v_i^{k-1} - 2x_i^T \beta^k e^k - 2 * \lambda \tilde{\Sigma}_{ii} v_i^{k-1}, v_i - v_i^{k-1} \rangle + L_{v_i k} \|v_i - v_i^{k-1}\|^2 \\
\text{subject to} \quad & ||v_i||_{l^2} = 1 \\
& i = 1, .., n
\end{aligned}
\tag{5.12}
$$

where $X = [x_1, ..., x_n]^T$ and $L_{v_i k}$ is the Lipschitz constant of $2e^k e^{k^T} v_i - 2x_i^T \beta^k e^k - 2 * \lambda \tilde{\Sigma}_{ii} v_i$ with respect to v_i.

5.5 Simulation Studies

In this simulation study, we generate data from a multivariate normal distribution with $p = 5, 10, 20$, $l + 1 = 3, 4$, $n = 100$, $T_i = T = 600$. We repeat our experiment 10 times and report their average.

5.5.1 Implementation

Here we introduce how we simulate our data.

1. Simulate the covariates matrix $X \in R^{n \times (l+1)}$. Let $X[:, 1] = 1$ and $X[:, 2 : l + 1] \sim$ Bernoulli(0.5). x_i are iid samples.

2. Simulate $e \in R^{p \times 1}$ from standard normal distribution. $\beta \in R^{(l+1) \times 1}$ from uniform distribution, e is scaled to have unit norm.

3. Simulate independent v_i from $v_i^T e = x_i^T \beta$ by utilizing projected normal distribution [Hernandez-Stumpfhauser et al., 2017] for $i = 1, 2, ..., n$.

4. Simulate Σ_i with v_i as the eigenvector corresponding the largest eigenvalue (we call it leading eigenvectors). We assume the dimension of corresponding eigenspace is 1.

5. Simulate $Y \sim N(0, \Sigma_i)$ and calculate the sample covariance $\tilde{\Sigma}_i$ for each i.

As our optimization problem is non-convex, the initialization is critical for achieving low objective value. One option is to set the initial V with a random perturbation of sample covariances' leading eigenvectors. Then we set initial value of e as one of v_i and initialize β by minimizing $\||Ve| - X\beta\|$ under the constraint $X\beta > 0$. We use 5 different initializations (i.e. run 5 times) for each tuning parameter and choose the one with lowest loss. We use two metrics to measure the accuracy of our model $\beta_{nmse} = \frac{\|\beta-\beta^*\|^2}{\|\beta^*\|^2}$ and $e_{nmse} = \frac{\|e-e^*\|^2}{\|e^*\|^2}$.

5.5.2 Simple Case

In this section, we show the result of our proposed algorithm using simulation procedure illustrated in section 5.5.1 in Table 5.1. We also run our algorithm with fixed Σ_i ($p = 10$, $l+1 = 4$) and X. for each $T = 500$, 1000, 2000, 4000, 6000, 8000, 10000, 20000, 50000. We show the trend of errors in Figure 5.2.

5.5.3 Multi-dimensional Case

In practice, we may not be able to know which eigenvectors are associated with the covariates. Here we set $q = 3$ and $l+1 = 4$. We modify step 4 in the simulation procedure so we have v_i as the eigenvector corresponding to the q-th largest eigenvalue.

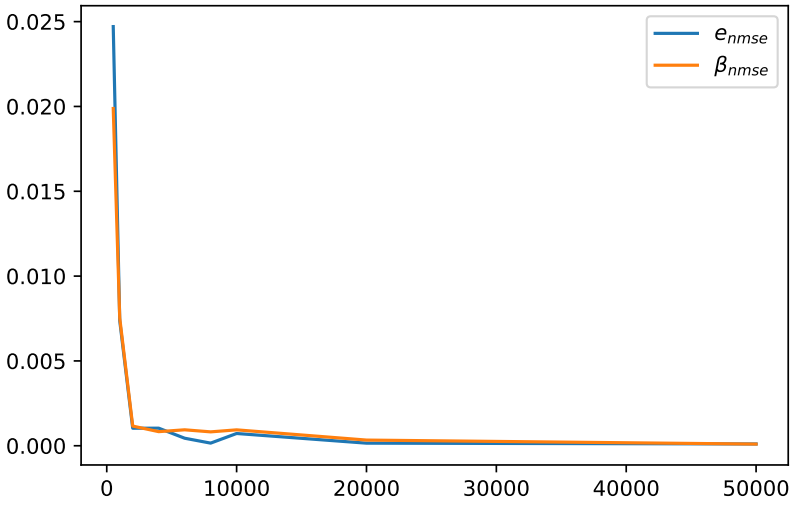

Figure 5.2: With increasing T, e_{nmse} and β_{nmse} obtained have a trend of decreasing.

Table 5.1: Experiments are repeated 10 times for one-dimensional case. We show the averaged results here.

p	$l+1$	β_{nmse}	e_{nmse}
5	3	0.0146	0.0070
10	3	0.0067	0.0104
20	3	0.0068	0.0601
5	4	0.0185	0.0105
10	4	0.0162	0.0114
20	4	0.0102	0.0420

Besides, we require the eigenspace of the three largest eigenvalues has dimension 1. After we simulate the data, we run the algorithm 7 with $R = 3$. where the `Deflation` function is defined as $\texttt{Deflate}(\tilde{\Sigma}_i) = \tilde{\Sigma}_i - \lambda_{i1}\tilde{v}_i\tilde{v}_i^T$ and λ_{i1} is the largest eigenvalue of input matrix and $\tilde{v}_i$ is the corresponding eigenvector. We list the results in Table 5.2. From the table, we can see the our algorithm obtains good results and we are able to discover eigenvectors that are associated to the covariates.

Algorithm 7 Algorithm of CVPR method - multi-dimensional case (5.5)

Input: X, $\tilde{\Sigma}_i$, $\lambda_1, ..., \lambda_m$, R
for $r = 1, 2, ...R$ **do**
  $\tilde{\Sigma}_i = \texttt{Deflate}(\tilde{\Sigma}_i)$
  for $\lambda = \lambda_1, ..., \lambda_m$ **do**
    Run Algorithm 6
    Return $V_{r,\lambda}$, $e_{r,\lambda}$, $\beta_{r,\lambda}$
  end for
end for
Select (V, e, β) from $\{(V_{r,\lambda}, e_{r,\lambda}, \beta_{r,\lambda})\}_{r,\lambda}$ based on permutation test.

Table 5.2: Experiments are repeated 10 times for multi-dimensional case. We show the averaged results here.

p	β_{nmse}	e_{nmse}
5	0.0331	0.0065
10	0.0134	0.0136
20	0.0087	0.0448

5.5.4 Tuning Parameter Selection

Throughout our simulation studies, we choose λ from $[0.01, 0.05, 0.1, 0.5, 1]$ based on permutation tests [Witten and Tibshirani, 2009].

5.6 Application

We apply our proposed method to Human Connectome Project (HCP) [Smith et al., 2013] resting-state fMRI (rs-fMRI) data. Our dataset includes $n = 97$ healthy young adults (36 with age in 22-25 and 61 in 26-30; 60 male and 37 female). The rs-fMRI dataset was preprocessed following the minimal preprocessing pipeline in [Glasser et al., 2013]. The blood-oxygen-level dependent (BOLD) signals were extracted from the default mode network (DMN) (with $p = 20$ functional brain regions) [Power et al.,

Table 5.3: FDR-corrected p-values using our algorithm for two different fMRI sessions

Session	Age	Gender	Age-Gender
1	0.0005	0.0119	0.0006
2	0.0000	0.3222	0.0000

2011] and averaged over voxels within the 5 mm radius.

In this work, we include age and gender (both as categorical variables), together with their interaction (age-gender) as our model covariates. For age, age = 1 represents age in 22-25 while age = 0 represents age in 26-30; for gender, gender = 1 for male and 0 for female. BOLD time series tend to be correlated across successive, which means they can no longer be treated as independent. Thus we calculate the effective sample size (ESS) defined by [Kass et al., 1998], which also follows the procedure used in [Zhao et al., 2018]. We subsample 424 time points from $T = 1200$, the original whole sequence.

We compare our method with the element-wise correlation method [Wang et al., 2007]. For the element-wise correlation regression, each off-diagonal element in the correlation matrix is Fisher z-transformed and multiple testing adjustment is performed to control the false discovery rate (FDR). See Figure 5.4 for the corrected p-values. None of the FDR-corrected p-values are significant at level 0.01.

Applying our algorithm 6, we show the FDR-corrected p-values in Table 5.3. Our method shows the p-values of age and age-gender are significant at level 0.01. We illustrate the projection vector e in Figure 5.4. Those with high magnitude are the regions which show a strong connection with the given covariates. We also apply the estimated projections to the second scanning session of resting-state fMRI data acquired from the same subjects. From table 5.3, we can see the significance of two different sessions are very similar, which shows the reliability of our proposed

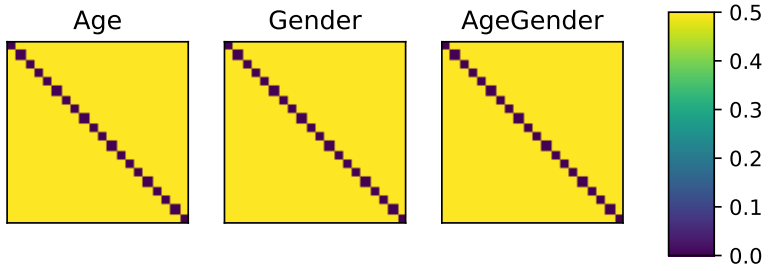

Figure 5.3: FDR corrected p-values of regression coefficients from element-wise regression. None of the entries are significant.

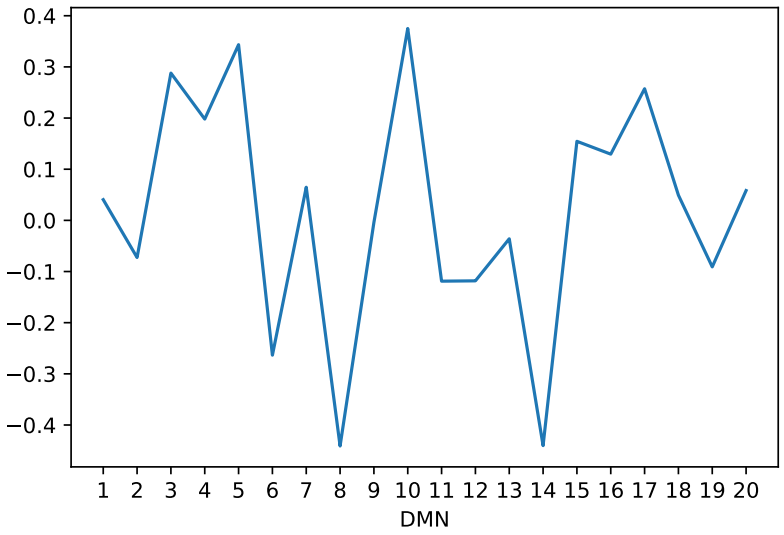

Figure 5.4: Illustrations of e in DMN

method.

5.7 Algorithmic Properties

Now, let's discuss the property our proposed algorithm for Problem (5.5). We assume X is full rank in this section.

Theorem 6. *Let $S = \{S_k\}_{k\geq 1} = \{e_k, \beta_k, v_1^k, ..., v_n^k\}_{k\geq 1}$ be a sequence generated from Algorithm 6, then any limit point $\{e_0, \beta_0, v_1^0, ..., v_n^0\}_{k\geq 1}$ of this sequence is a critical point of (5.5).*

Proof. First, we can see our proposed objective function $\mathcal{L}$ defined in (5.6) is continuously differentiable. $\mathbb{1}_{\|v_i\|=1}$ for $i = 1, 2, ..., n$, $\mathbb{1}_{\|e\|=1}$ and $\mathbb{1}_{X\beta>0}$ is lower semi-continuous. From the constraints, V^k and e^k are bounded. If $\|\beta^k\|$ goes to infinity, $\|X\beta^k\| \to \infty$ because X is full rank. Thus $\mathcal{F}$ is proper and lower bounded. The value of objective function $\mathcal{F}$ in (5.5) is monotonically decreasing during the optimization. Each L_k listed in Algorithm 6 is also upper bounded. Applying Theorem 1 [Xu and Yin, 2017], we get the stated results.

□

Theorem 7. *Let $S = \{S_k\}_{k\geq 1} = \{e_k, \beta_k, v_1^k, ..., v_n^k\}_{k\geq 1}$ be a sequence generated from Algorithm 6, Then S has a limit point $\{e_0, \beta_0, v_1^0, ..., v_n^0\}_{k\geq 1}$ and S converge to $\{e_0, \beta_0, v_1^0, ..., v_n^0\}_{k\geq 1}$.*

Proof. The proposed algorithm satisfies $L(S_k) \leq L(S_{k-1})$ [Xu and Yin, 2017], from

which we get S are bounded and any limit point of S is finite.

Now let's introduce the definition of semi-algebraic functions. A set $D \subset R^n$ is called semi-algebraic [Bochnak et al., 2013] if it can be represented as

$$D = \cup_i \cap_j \{x \in R^n : p_{ij}(x) = 0, q_{ij}(x) > 0\} \tag{5.13}$$

where p_{ij} ,q_{ij} are real polynomial functions for $1 \leq i \leq s$, $1 \leq j \leq t$. Indicator functions of semi-algebraic sets are semi-algebraic [Xu and Yin, 2013]. Thus we get $\mathbb{1}_{\|v_i\|=1}$ for $i = 1, 2, ..., n$, $\mathbb{1}_{\|e\|=1}$ and $\mathbb{1}_{X\beta>0}$ are semi-algebraic. Finite sums and products of semi-algebraic functions are semi-algebraic. Therefore, we obtain $\mathbb{1}_{\{\|e\|_{l^2}^2=1, X\beta\geq 0, \|v_i\|_{l^2}=1\}}$ is semi-algebraic. In addition, $\mathcal{L}$ is polynomial function and thus is real analytic. The sum of a real analytic function and a semi-algebraic function satisfies Kurdyka- Lojasiewicz property [Xu and Yin, 2013], so $\mathcal{F}$ satisfies Kurdyka- Lojasiewicz property [Łojasiewicz, 1993]. Lastly, $\nabla_e \mathcal{L}(V, e, \beta)$, $\nabla_{v_i} \mathcal{L}(V, e, \beta)$ for $i = 1, 2, ..., n$ and $\nabla_\beta \mathcal{L}(V, e, \beta)$ is Lipschitz continuous about (V, e, β) on any compact set. Applying Theorem 2 [Xu and Yin, 2017] completes the proof. □

5.8 Asymptotic Properties

5.8.1 Identifiability

Identifiability is critical for building a well-posed model. Here we provide sufficient conditions for our model to achieve identifiable projection vectors before establishing theoretical properties. Let's review our problem (5.2), from the objective function, we can easily see the sign of e and V is not determined due to the absolute value

function in the objectives. One simple fix is to revise the sign according to $\text{sum}(e) \geq 0$ and $\langle v_i, e\rangle \geq 0$ for $i = 1, 2, ..., n$, which can be done after applying our algorithm. We acknowledge that the above choice is arbitrary and one can utilize other ways to solve this issue. WLOG, we assume $\text{sum}(e^*) > 0$, $\text{sum}(e) > 0$ and $\langle v_i^*, e^*\rangle > 0$ in the rest of our work.

Assumption 1. *We consider the following problem*

$$\begin{aligned} &\textit{minimize}_{e,\beta} && \tfrac{1}{n}\||V^*e| - X\beta\|^2 \\ &\textit{subject to} && \|e\|_{l^2}^2 = 1 \end{aligned} \tag{5.14}$$

where V^* is the model ground truth, i.e. V^* is the leading eigenvectors for $\{\Sigma_1, ..., \Sigma_n\}$ and $V^*e^* = X\beta^*$ where e^* and β^* are true parameters. Then (5.14) has unique global minimum $(e = e^*, \beta = \beta^*)$.

Before going further to prove the identifiability of our model, we first give one sufficient condition of V^* which makes our model satisfy 1. In the rest of section, WLOG, we assume $e^* = (1, 0, 0, ..., 0)$. As $v_i^{*T}e^* = x_i^T\beta^*$, we can rewrite $v_i^* = x_i^T\beta^*e^* + \sqrt{(1 - (x_i^T\beta^*)^2)}r_i$ where r_i is perpendicular to e^* $(r_i \perp e^*)$. Because $e^* = (1, 0, 0, ..., 0)$, we can write $r_i = (0, r_{i0})$ where $r_{i0} \in R^{p-1}$

Theorem 8. *Assume x_i are iid samples from a distribution with invertible $E(x_ix_i^T) = \Omega$. If r_{i0} is iid random variables sampled from projected normal distribution [Hernandez-Stumpfhauser et al., 2017] on R^{p-1} and r_{i0} is independent of x_i, then $\mathcal{P}(o \in O : \exists N_o, \forall n > N_o, \textit{Assumption 1 is satisfied}) = 1$ where O is the sample space.*

Proof. First, we note when $n \to \infty$, (e^*, β^*) is the unique global minimum of (5.14). When $n \to \infty$, $\frac{1}{n}\||V^*e| - X\beta\|^2 \to E(\||v_i^{*T}e| - x^i\beta\|^2)$ if $o \notin O_1$ where $\mathcal{P}(O_1) = 0$. We know $E(\||v_i^{*T}e^*| - x^i\beta^*\|^2) = 0$ and $E(\||v_i^{*T}e| - x^i\beta\|^2) = E(E(\||v_i^{*T}e| - x_i\beta\|^2|x_i))$.

When x_i is fixed, $|v_i^{*T}e|$ must be constant a.s.. From the distribution of r_{i0}, we get e^* is the only e that makes $|v_i^{*T}e|$ = constant a.s provided sum(e) >= 0. Given $Ex_ix_i^T$ is invertible, β is thus uniquely determined.

Let $\frac{X^TX}{n} \to \Omega$ for $o \notin O_2$. Now, let's prove the proposition by contradiction. Fix $o \notin O_1 \cup O_2$. If there $(e_{n_w}, \beta_{n_w}) \neq (e^*, \beta^*)$ such that $\frac{1}{n_w}\||V^*e_{n_w}| - X\beta_{n_w}\|^2 = 0$ and $n_w \to \infty$. Because $(e_{n_w}, \beta_{n_w} = (\frac{X^TX}{n_w})^{-1}\frac{X^T|V^*e_{n_w}|}{n_w})$ are bounded, we have a subsequence $(e_{n_{w_j}}, \beta_{n_{w_j}})$ such that $(e_{n_{w_j}}, \beta_{n_{w_j}}) \to (e_0, \beta_0)$. Denote $f_j = \frac{\||V^*e_0| - X\beta_0\|^2 - \||V^*e_{n_{w_j}}| - X\beta_{n_{w_j}}\|^2}{n_{w_j}}$.

$$|f_j| \leq C(|e_{n_{w_j}} - e_0| + |\beta_{n_{w_j}} - \beta_0|) \tag{5.15}$$

where C is a constant. This gives us $\lim_{i\to\infty} f_i \to 0$.

$$\begin{aligned} &E(\||v_i^{*T}e_0| - x_i^T\beta_0\|^2) - \lim_{j\to\infty} \frac{\||V^*e_{n_{w_j}}| - X\beta_{n_{w_j}}\|^2}{n_{w_j}} \\ &\quad = E(\||v_i^{*T}e_0| - x^i\beta_0\|^2) - \lim_{j\to\infty} \frac{\||V^*e_0| - X\beta_0\|^2}{n_{w_j}} - \lim_{j\to\infty} f_j = 0 \end{aligned} \tag{5.16}$$

$\Rightarrow$

$$E(\||v_i^{*T}e_0| - x_i^T\beta_0\|^2) = 0 \tag{5.17}$$

As we discussed before, when $n \to \infty$, (e^*, β^*) is the unique global minimum of (5.14). We get $(e_0, \beta_0) = (e^*, \beta^*)$. Because $0 < lb < X_i^T\beta^* < ub < 1$, we know there exists N_1 such that for any $j > N_1$, $V^*e_{n_{w_j}} > 0$. This means when $j > N_1$, (5.18) has at least two global minimum.

$$\begin{aligned} &\text{minimize}_{e,\beta} \quad \tfrac{1}{n_{w_j}}\|V^*e - X\beta\|^2 \\ &\text{subject to} \quad \|e\|_{l^2}^2 = 1 \end{aligned} \tag{5.18}$$

To achieve the global minimum, we must have $\beta = (X^T X)^{-1} X^T V^* e$. Thus we can rewrite the objective function as

$$\frac{1}{n_{w_j}}(e^T V^{*T} V^* e - e^T V^{*T} X (X^T X)^{-1} X^T V^* e) = e^T (\frac{V^{*T} V^*}{n_{w_j}} - \frac{V^{*T} X}{n_{w_j}} (\frac{X^T X}{n_{w_j}})^{-1} \frac{X^T V^*}{n_{w_j}}) e$$

We denote $\Gamma_j = \frac{V^{*T} V^*}{n_{w_j}} - \frac{V^{*T} X}{n_{w_j}} (\frac{X^T X}{n_{w_j}})^{-1} \frac{X^T V^*}{n_{w_j}}$, $\Omega = E x_i x_i^T$ and ${\beta^*}^T \Omega \beta^* = C < 1$.

$$\lim_{j \to \infty} \frac{V^{*T} V^*}{n_{w_j}} = \lim_{j \to \infty} (\sum_{i=1}^{n_{w_j}} \frac{(x_i^T \beta^*)^2}{n_{w_j}} e^* e^{*T} + \sum_{i=1}^{n_{w_j}} \frac{1 - (x_i^T \beta^*)^2}{n_{w_j}} r_i r_i^T) = C e^* e^{*T} + (1 - C)\Gamma \text{ a.s}$$

where the second equation comes from r_i and x_i are independent. Γ is a diagonal matrix with $\Gamma[1,1] = 0$ and $\Gamma[m,m] > 0$ for $m \neq 1$. We also have

$$\lim_{j \to \infty} \frac{V^{*T} X}{n_{w_j}} = e^* {\beta^*}^T \Omega \text{ a.s}$$

$$\lim_{j \to \infty} \frac{V^{*T} X}{n_{w_j}} (\frac{X^T X}{n_{w_j}})^{-1} \frac{X^T V^*}{n_{w_j}} = C e^* e^{*T} \text{ a.s}$$

Combining these equations, we get $\lim_{j \to \infty} \Gamma_j = (1 - C)\Gamma$ if $o \notin O_3$ where $\mathcal{P}(O_3) = 0$. The rank of Γ is $p - 1$ and thus we get there exists N_2, for any $j > N_2$, Γ_j has rank of $p - 1$. So we can not have two global minimum with objective value 0 and we reach a contradiction. Thus we have for any $o \notin O_1 \cup O_2 \cup O_3$, there exists N_o such that for any $n > N_o$, assumption 1 is satisfied. □

Now we are ready to give a sufficient condition for our model to be identifiable. Note, in the rest of the work, we may omit the constraints of the considered optimization problem for simplicity. First, we give the second assumption on covariances' eigenvalues.

Assumption 2. *$\lambda_{i1} > \lambda_{i2}$ where $\lambda_i = (\lambda_{i1}, ..., \lambda_{ip})$ is the eigenvalues of Σ_i with*

descending order.

Theorem 9. *Suppose that Assumption 1 and Assumption 2 are satisfied. When* $T \to \infty$, $\||Ve| - X\beta\|^2 - \lambda \sum_i v_i^T \tilde{\Sigma}_i v_i$ *has unique global minimum* (V^*, e^*, β^*) *a.s.. Our model is identifiable.*

Proof. When $T \to \infty$, the objective function becomes $\||Ve| - X\beta\|^2 - \lambda \sum_i v_i^T \Sigma_i v_i$ a.s.. We note $\||V^*e^*| - X\beta^*\|^2 = 0$ and V^* achieves the minimum of $-\lambda \sum_i v_i^T \Sigma_i v_i$. Because $\lambda_{i1} > \lambda_{i2}$ for $i = 1, 2, ..., n$, we thus have $\||V^*e^1| - X\beta^1\|^2 = 0$ if the objective function has a second minimum $(e^1, \beta^1) \neq (e^*, \beta^*)$. The result then follows from from Assumption 1. Note that the sign of V can be determined by $\langle v_i, e \rangle \geq 0$. □

5.8.2 Consistency

In this section, we study the consistency of our proposed model given X. Throughout this section, we assume Assumption 2 is satisfied. We first rewrite our model using (5.19).

$$\begin{aligned} \text{maximize}_{e,\beta,v_i} \quad & \sum_i v_i^T \tilde{\Sigma}_i v_i - \lambda_T \||Ve| - X\beta\|^2 \\ \text{subject to} \quad & \|e\|_{l^2}^2 = 1 \\ & \|v_i\|_{l^2} = 1 \\ & i = 1, .., n \end{aligned} \tag{5.19}$$

We will discuss the model consistency when $T \to \infty$ and $\lambda_T \to 0$. We follow the sign convention discussed in section 5.8.1 ($\text{sum}(v_i, \tilde{e}) \geq 0$ and $\text{sum}(\tilde{e}) \geq 0$ where $\tilde{e}$ can be obtained by optimizing (5.2) with a constraint similar to (5.20)). All c_i's with are assumed to be non-negative constants. Denote $L_T(e, \beta, V) = \sum_i v_i^T \tilde{\Sigma}_i v_i -$

$\lambda|||Ve|-X\beta||^2$, $l_t(V,Y)=\sum_i v_i^T\tilde{\Sigma}_i v_i$ and $l_t(V,y_t)=\sum_{i=1}^n v_i^T y_{it}y_{it}^T v_i$. And we assume $T_1=T_2=...=T$. Suppose we have global maximum (e_T^*,β_T^*,V_T^*) for problem (5.19) where we omit T in $\tilde{\Sigma}_i$. We have a sequence of estimations (e_T,β_T,V_T) which satisfies (5.20).

$$L_T(e_T,\beta_T,V_T)\geq L_T(e_T^*,\beta_T^*,V_T^*)-a_T \tag{5.20}$$

where $\lim_{T\to\infty}a_T=0$. We define $J(V)=\min_{\|e\|=1,\beta}|||Ve|-X\beta||^2$, then we get

$$L_T(e_T^*,\beta_T^*,V_T^*)=l_t(V_T^*,Y)-\lambda_T J(V_T^*)$$
$$L_T(e_T,\beta_T,V_T)\leq l_t(V_T,Y)-\lambda_T J(V_T)$$

Combining these equations, we get (5.21)

$$l_t(V_T,Y)-\lambda_T J(V_T)\geq L_T(e_T,\beta_T,V_T)\geq L_T(e_T^*,\beta_T^*,V_T^*)-a_T=\max_V(l_t(V,Y)-\lambda_T J(V))-a_T \tag{5.21}$$

Let $l_\Delta=l_t(V^*,y_t)-l_t(V,y_t)$ and $K(V^*,V)=E_Y(l_\Delta)=\rho(V^*,V)=\sum_i v_i^{*T}\Sigma_i v_i^*-\sum_i v_i^T\Sigma_i v_i\geq 0$. We rewrite $v_i=(r_i*v_i^*+\sqrt{1-r_i^2}*\tilde{v}_i)$ where $v_i^*\perp\tilde{v}_i$, $\|\tilde{v}_i\|=1$ and $|r_i|\leq 1$. Therefore we get

$$K(V^*,V)=\sum_i(1-r_i^2)(v_i^{*T}\Sigma_i v_i^*-\tilde{v}_i^T\Sigma_i\tilde{v}_i)$$

, $\|v_i-v_i^*\|=\sqrt{2(1-r_i)}$ and $\|V-V^*\|=\sqrt{2\sum_i(1-r_i)}$. By assumption 2, we get there exists c_0, $(v_i^{*T}\Sigma_i v_i^*-\tilde{v}_i^T\Sigma_i\tilde{v}_i)>c_0$ for all i. Recall that $\text{sum}(v_i,\tilde{e})\geq 0$ and $\text{sum}(v_i^*,e^*)>0$, by Theorem 9, $1+r_i>c_1$ for some $c_1>0$, for large T. We then have

$$K(V^*,V)\geq c_0*c_1\sum_i(1-r_i)\geq\frac{c_0*c_1}{2}\|V-V^*\|^2$$

Let $W(V^*, V) = \mathrm{Var}(l_\Delta)$, then

$$W(V^*, V) = \sum_i \mathrm{Var}((1 - r_i^2)(\tilde{v_i}^T y_{it}^T y_{it} \tilde{v_i} - v_i^{*T} y_{it}^T y_{it} v_i^*) + 2r_i \sqrt{1 - r_i^2} \tilde{v_i}^T y_{it}^T y_{it} v_i^*)$$

Following the similar argument as before, we have

$$c_3 \|V^* - V\|^2 \geq W(V^*, V) \geq c_2 \|V^* - V\|^2$$

Since, in practice, the observations Y are usually finite, we assume Y are sampled from some distribution with bounded domain.

Theorem 10. *Suppose that Assumption 2 is satisfied. V_T is a sample estimator satisfying (5.20). Then we have*

$$\mathcal{P}(\rho(V_T, V^*) \geq \epsilon_T) \leq 7 \exp(-cT\epsilon_T^2)$$

where $\lambda_T \sim \epsilon_T^2$, $a_T \sim o(\epsilon_T^2)$, $\epsilon_T \sim \frac{1}{T}^{\frac{\eta}{2\eta+1}}$ for any $\eta > 1$ and $c > 0$.

Proof. We first verify several conditions of Corollary 2 in [Shen, 1998]. Let $A(k_1, k_2) = \{\|v_i\| = 1 : k_1 \leq \rho(V^*, V) \leq 2 * k_1, J(V) \leq k_2\}$ and $\mathcal{F} = \{l_\Delta(V|\cdot) : V \in A(k_1, k_2)\}$. We have $\sup_{A(k1,k2)} W(V^*, V) \leq c_4 k_1^2 (1 + (k_1^2 + k_2)^\beta)$ with $\beta = 0$. $|l_\Delta| = |\sum_i (v_i^{*T} y_{it}^T y_{it} v_i^* - v_i^T y_{it}^T y_{it} v_i)| \leq \sum_i |v_i^{*T} y_{it}^T y_{it} v_i^* - v_i^T y_{it}^T y_{it} v_i| \leq c_4 |V - V^*|$. Additionally, $\sup_{A(k_1,k_2)} \|V - V^*\|_{sup} \leq c_5 (k_1^2 + k_2)^\gamma$ where $\gamma = 0$ because V is bounded.

We next verify the assumption on the Hellinger metric entropy. Let

$$S(\epsilon, m) = \{g_1^l, g_1^u, ..., g_m^l, g_m^u\} \in \mathcal{L}_2$$

where $\|g_i^u - g_i^l\|_2 \leq \epsilon$. Suppose for any function $f \in \mathcal{F}(k_1, k_2)$, there exists i such

that $g_i^l \leq f \leq g_i^u$ a.s.. Then the Hellinger metric entropy with bracketing is defined as $H(\epsilon, F) = \log N(\epsilon, \mathcal{F}) = \log(\min\{m : S(\epsilon, m)\})$. As $\mathcal{F}(k_1, k_2) \in W_2^\infty$, according to Theorem 5.2 in [Birman and Solomyak, 1967], for any $\eta > 1$, there exists c_6, such that $H(\epsilon_T, \mathcal{F}) \leq c_6 \epsilon_T^{-\frac{1}{\eta}}$. Define

$$\psi(k_1, k_2) = \int_{L_0}^{U_0} H^{1/2}(u, \mathcal{F}) du / L_0$$

where $L_0 = c_7 \lambda_T (k_1^2 + k_2)$ and $U_0 = c_8 \epsilon_T (k_1^2 + k_2)^{1/2}$. Therefore, we have

$$\psi(k_1, k_2) \leq c_9 \int_{L_0}^{U_0} u^{-\frac{1}{2\eta}} du / L_0$$

To satisfy the condition that $\psi(k_1, k_2) \leq c_{10} T^{1/2}$ and $\lambda_T \sim c_{11} \epsilon_T^2$, we have the best possible rate $\epsilon_T \sim \frac{1}{T}^{\frac{\eta}{2\eta+1}}$ for any $\eta > 1$. We note

$$\begin{aligned}
\mathcal{P}(\rho(V_T, V^*) \geq \epsilon_T) &\leq \mathcal{P}^*(\sup_{\rho(V,V^*) \geq \epsilon_T} (l_t(V, Y) - \lambda_T J(V)) - (l_t(V_T^*, Y) - \lambda_T J(V_T^*)) \geq -a_T) \\
&\leq \mathcal{P}^*(\sup_{\rho(V,V^*) \geq \epsilon_T} (l_t(V, Y) - \lambda_T J(V)) - (l_t(V^*, Y) - \lambda_T J(V^*)) \geq -a_T)
\end{aligned}$$

The result then follows. □

Theorem 11. *Suppose that Assumption 1 and 2 are satisfied. ϵ_T, a_T and λ_T are defined as in Theorem 10. Then we have $e_T, \beta_T \to e^*, \beta^*$ a.s. when $T \to \infty$*

Proof. Based on Theorem 4 in [Shen, 1998], for any $0 < \delta < 1/4$, if $(1-\delta) c_{12} \epsilon_T^2 \leq \lambda_n$, we have

$$\mathcal{P}(J(V_T) \geq O(\delta)) \leq 10 \exp(-c_{13} T \epsilon_T^2)$$

Therefore, $J(V_T) \to 0$ a.s. when $T \to \infty$. Additionally, from

$$L_T(e_T, \beta_T, V_T) + a_T \geq l_t(V_T, Y) - \lambda_T J(V_T),$$

we get $J(V_T) + \frac{a_T}{\lambda_T} \geq \| |V_T e_T| - X\beta_T \|$. Thus $\| |V_T e_T| - X\beta_T \| \to 0$ as $T \to \infty$ a.s.. We note $V_T \to V^*$, and e_T, β_T are bounded a.s.. Therefore, $e_T, \beta_T \to e^*, \beta^*$ a.s.. This concludes our proof. □

Remark: Although our results in this section is proved under the condition that V^* corresponds to the largest eigenvalues of $\{\Sigma_i\}_i$, it is possible to extend our results to multi-dimensional case by adding additional model assumptions, and we leave it to future work. It is worth noting that the consistency for eigenvectors of covariance matrices has been studies in previous work [Johnstone and Lu, 2009].

5.9 Conclusion

In this work, we introduce a Common Vector Projection Regression Model for multiple covariance matrix outcomes. It builds a relationship between the global structure covariance network (eigenvectors) and auxiliary information (covariates). Our approach allows the identification of a projection direction that is associated with external variables. Under mild conditions, we show our proposed algorithm has the property of global convergence. In theory, we demonstrate the estimated parameter achieves asymptotic consistency. It is also worth noting that we do not require the knowledge of observations' distribution in our model and estimation. Lastly, we illustrate the application of our algorithm in both simulations and real fMRI datasets. Our work is an initial attempt to connect eigenvectors and covariates. There are

several future directions of our work. It is interesting to study how to extend our framework to large-scale problems with the settings $p \gg n$ or $l \gg n$. Another direction to study the convergence rate of e and β. Our model can also be applied to the case where we are interested in analyzing a transformed covariance matrices, e.g., $\text{abs}(\Sigma_i)$.

Part V

Conclusions

Chapter Six

Conclusions

6.1 Dynamic Causal Modeling using Functional Data Method

Functional MRI (fMRI) is a popular approach to investigate brain connections and activations when human subjects perform tasks. Because fMRI measures the indirect and convoluted signals of brain activities at a lower temporal resolution, complex differential equation modeling methods (e.g. Dynamic Causal Modeling) are usually employed to infer the neuronal processes and to fit the resulting fMRI signals. However, this modeling strategy is computationally expensive and remains to be mostly a confirmatory or hypothesis-driven approach. One major statistical challenge here is to infer, in a data-driven fashion, the underlying differential equation models from fMRI data.

In section 2, we propose a causal dynamic network (CDN) method to estimate brain activations and connections simultaneously. Our method links the observed fMRI data with the latent neuronal states modeled by an ordinary differential equation (ODE) model. Using the basis function expansion approach in functional data analysis, we develop an optimization-based criterion that combines data-fitting errors and ODE fitting errors. We also develop and implement a block coordinate-descent algorithm to compute the ODE parameters efficiently. We illustrate the numerical advantages of our approach using data from realistic simulations and two task-related fMRI experiments. Comparing with various effective connectivity methods, our method achieves higher estimation accuracy while improving the computational speed by from tens to thousands of times. Though our method is developed for task-related fMRI, we also demonstrate the potential applicability of our method (with a simple modification) to resting-state fMRI, by analyzing both simulated and real data from medium-sized networks. In section 3, built on the popular DCM that

was developed mainly for a hypothesis-validation approach of small-scale data, we introduce a new optimization problem and an efficient iterative algorithm that uses data-driven search from a huge number of ODE models.

6.2 Canonical Correlation Analysis for Time Series

Canonical Correlation Analysis is a technique in multivariate data analysis for finding linear projections that maximize the correlation between two groups of variables. The correlations are typically defined without accounting for the serial correlations between observations, a typical setting for time series data. To understand the coupling dynamics and temporal variations between the two time-varying sources, in section 4, we introduce the time-dependent canonical correlation analysis (TDCCA), a method for inferring time-dependent canonical vectors from multilevel time series data. A convex formulation of the problem is proposed, which leverages the singular value decomposition (SVD) characterization of all solutions of the CCA problem. We use simulated datasets to validate the proposed algorithm. Moreover, we propose a novel measure, canonical correlation variation as another way to assess the dynamic pattern of brain connections and we apply it to a real resting-state fMRI dataset to study the aging effects on brain connectivity. Additionally, we explore our proposed method in a task-related fMRI to detect the temporal dynamics due to different motor tasks. We show that, compared to extant methods, the TDCCA-based approach not only detect temporal changes but also improves feature extraction. Together, this work contributes broadly to new computational methodologies in understanding multilevel time series.

6.3 Covariance Regression Models for Brain Connectivity Analysis

Modeling covariances has been an important top in many areas including financial and neuroimaging analysis. In section 5, we introduce common vector projection regression, an optimization-based method for identifying the relationship between covariates and covariances. In the work, we develop an efficient algorithm to jointly search the common projection direction and regression coefficients of covariates. We also establish algorithmic property of the introduced algorithm and the asymptotic properties of our proposed model with certain underlying model assumptions. Our algorithm shows high accuracy and robustness in coefficient estimation in both single-dimensional and multi-dimensional analysis. In addition, we explore our method using a resting-state functional magnetic resonance imaging study. It shows our approach identifies the human brain network varies associated with external information of subjects.

Bibliography

G. K. Aguirre, E. Zarahn, and M. D'esposito. The variability of human, BOLD hemodynamic responses. *Neuroimage*, 8(4):360–369, 1998.

S. Akaho. A kernel method for canonical correlation analysis. In *International Meeting on Psychometric Society, 2001*, 2001.

T. Anderson et al. Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, 1(1):135–141, 1973.

A. R. Aron. The neural basis of inhibition in cognitive control. *The Neuroscientist*, 13(3):214–228, 2007.

A. R. Aron and R. A. Poldrack. Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. *Journal of Neuroscience*, 26 (9):2424–2433, 2006. doi: 10.1523/JNEUROSCI.4682-05.2006.

M. Asencio, G. Hooker, and H. O. Gao. Functional convolution models. *Statistical Modelling*, 14(4):315–335, 2014. doi: 10.1177/1471082X13508262.

A. Badhwar, A. Tam, C. Dansereau, P. Orban, F. Hoffstaedter, and P. Bellec. Resting-state network dysfunction in alzheimer's disease: a systematic review and meta-analysis. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 8:73–85, 2017.

P. A. Bandettini. What's new in neuroimaging methods? *Annals of the New York Academy of Sciences*, 1156:260, 2009.

O. Banerjee, L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis. Convex optimization techniques for fitting sparse Gaussian graphical models. In *Proceedings of the 23rd International Conference on Machine learning*, pages 89–96. ACM, 2006. doi: 10.1145/1143844.1143856.

D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al. Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80: 169–189, 2013.

Y. Bard. *Nonlinear parameter estimation*. Academic press, 1974.

D. S. Bassett and E. Bullmore. Small-world brain networks. *The Neuroscientist*, 12 (6):512–523, 2006. doi: 10.1177/1073858406293182.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

P. Bhaumik and S. Ghosal. Bayesian two-step estimation in differential equation models. *Electronic Journal of Statistics*, 9(2):3124–3154, 2015. doi: 10.1214/15-EJS1099.

P. J. Bickel, E. Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

F. Bießmann, F. C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. K. Logothetis, and K.-R. Müller. Temporal kernel cca and its application in multimodal neuronal data analysis. *Machine Learning*, 79(1-2):5–27, 2010.

J. R. Binder, J. A. Frost, T. A. Hammeke, R. W. Cox, S. M. Rao, and T. Prieto. Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, 17(1):353–362, 1997.

J. R. Binder, W. L. Gross, J. B. Allendorfer, L. Bonilha, J. Chapin, J. C. Edwards, T. J. Grabowski, J. T. Langfitt, D. W. Loring, M. J. Lowe, et al. Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *Neuroimage*, 54(2):1465–1475, 2011.

M. S. Birman and M. Z. Solomyak. Piecewise-polynomial approximations of functions of the classes w_pˆα. *Matematicheskii Sbornik*, 115(3):331–355, 1967.

J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

R. J. Boik. Spectral models for covariance matrices. *Biometrika*, 89(1):159–182, 2002.

A. Bolstad, B. D. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing*, 59(6):2628–2641, 2011. doi: 10.1109/TSP.2011.2129515.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

K. H. Brodersen, T. M. Schofield, A. P. Leff, C. S. Ong, E. I. Lomakina, J. M. Buhmann, and K. E. Stephan. Generative embedding for model-based classification of fMRI data. *PLoS Computational Biology*, 7(6):e1002079, 2011.

N. J. Brunel, Q. Clairon, and F. d'Alche Buc. Parametric estimation of ordinary differential equations with orthogonality conditions. *Journal of the American Statistical Association*, 109(505):173–185, 2014. doi: 10.1080/01621459.2013.841583.

N. J. Brunel et al. Parameter estimation of ODE's via nonparametric estimators. *Electronic Journal of Statistics*, 2:1242–1267, 2008. doi: 10.1214/07-EJS132.

R. L. Buckner, J. Sepulcre, T. Talukdar, F. M. Krienen, H. Liu, T. Hedden, J. R. Andrews-Hanna, R. A. Sperling, and K. A. Johnson. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *Journal of Neuroscience*, 29(6):1860–1873, 2009.

R. B. Buxton, E. C. Wong, and L. R. Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, 39(6):855–864, 1998. doi: 10.1002/mrm.1910390602.

T. Cai and W. Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

T. Cai, W. Liu, and X. Luo. A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106 (494):594–607, 2011.

J. Cao and J. O. Ramsay. Generalized profiling estimation for global and adaptive penalized spline smoothing. *Computational Statistics & Data Analysis*, 53(7): 2550–2562, 2009. doi: 10.1016/j.csda.2008.12.004.

M. Cappelletti, B. Butterworth, and M. Kopelman. Spared numerical abilities in a case of semantic dementia. *Neuropsychologia*, 39(11):1224–1239, 2001.

A. Caprihan, G. D. Pearlson, and V. D. Calhoun. Application of principal component analysis to distinguish patients with schizophrenia from healthy controls based on fractional anisotropy measurements. *Neuroimage*, 42(2):675–682, 2008.

L. Chaari, T. Vincent, F. Forbes, M. Dojat, and P. Ciuciu. Fast joint detection-estimation of evoked brain activity in event-related fMRI using a variational approach. *IEEE Transactions on Medical Imaging*, 32(5):821–837, 2013.

L. K. Chan, J. Karceski, and J. Lakonishok. The risk and return from factors. *Journal of financial and quantitative analysis*, 33(2):159–188, 1998.

C. Chang and G. H. Glover. Time–frequency dynamics of resting-state brain connectivity measured with fmri. *Neuroimage*, 50(1):81–98, 2010.

C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of admm for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.

M. Chen, C. Gao, Z. Ren, and H. H. Zhou. Sparse cca via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

X. Chen and T. G. Conley. A new semiparametric spatial model for panel time series. *Journal of econometrics*, 105(1):59–83, 2001.

Y. Chen, Y.-n. Liu, P. Zhou, X. Zhang, Q. Wu, X. Zhao, and D. Ming. The transitions between dynamic micro-states reveal age-related functional network reorganization. *Frontiers in Physiology*, 9, 2018.

T. Y. Chiu, T. Leonard, and K.-W. Tsui. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210, 1996.

O. A. Chkrebtii, D. A. Campbell, B. Calderhead, M. A. Girolami, et al. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267, 2016. doi: 10.1214/16-BA1017.

D. Chu, L.-Z. Liao, M. K. Ng, and X. Zhang. Sparse canonical correlation analysis: new formulation and algorithm. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):3050–3065, 2013.

S. J. Crutch and E. K. Warrington. Preserved calculation skills in a case of semantic dementia. *Cortex*, 38(3):389–399, 2002.

J. Daunizeau, O. David, and K. E. Stephan. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage*, 58(2):312–322, 2011.

J. Daunizeau, K. E. Stephan, and K. J. Friston. Stochastic dynamic causal modelling of fMRI data: should we care about neural noise? *Neuroimage*, 62(1):464–481, 2012.

O. David, I. Guillemain, S. Saillet, S. Reyt, C. Deransart, C. Segebarth, and A. Depaulis. Identifying neural drivers with functional MRI: an electrophysiological validation. *PLoS Biology*, 6(12):e315, 2008.

P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Annals of Statistics*, pages 1–48, 2001.

E. L. Dennis and P. M. Thompson. Functional brain connectivity using fmri in aging and alzheimer's disease. *Neuropsychology review*, 24(1):49–62, 2014.

G. Deshpande, X. Hu, R. Stilla, and K. Sathian. Effective connectivity during haptic perception: a study using Granger causality analysis of functional magnetic resonance imaging data. *Neuroimage*, 40(4):1807–1814, 2008.

D. Dima, D. E. Dietrich, W. Dillo, and H. M. Emrich. Impaired top-down processes in schizophrenia: A DCM study of ERPs. *NeuroImage*, 52(3):824–832, 2010.

V. B. Domselaar and P. Hemker. Nonlinear parameter estimation in initial value problems. *Stichting Mathematisch Centrum. Numerieke Wiskunde*, pages 1–49, 1975.

N. F. Dronkers, B. B. Redfern, and R. T. Knight. The neural architecture of language disorders. *The New Cognitive Neurosciences*, 2:949–960, 2000.

N. F. Dronkers, D. P. Wilkins, R. D. Van Valin Jr, B. B. Redfern, and J. J. Jaeger. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1-2):145–177, 2004.

E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, and R. T. Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11): 1664, 2015.

B. Flury. *Common principal components & related multivariate models*. John Wiley & Sons, Inc., 1988.

B. N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.

B. N. Flury. Asymptotic theory for common principal component analysis. *The annals of Statistics*, pages 418–430, 1986.

E. B. Fox and D. B. Dunson. Bayesian nonparametric covariance regression. *The Journal of Machine Learning Research*, 16(1):2501–2542, 2015.

A. Franks and P. Hoff. Shared subspace models for multi-group covariance estimation. *arXiv preprint arXiv:1607.03045*, 2016.

S. Frässle, K. E. Stephan, K. J. Friston, M. Steup, S. Krach, F. M. Paulus, and A. Jansen. Test-retest reliability of dynamic causal modeling for fMRI. *Neuroimage*, 117:56–66, 2015.

S. Frässle, E. I. Lomakina, A. Razi, K. J. Friston, J. M. Buhmann, and K. E. Stephan. Regression DCM for fMRI. *NeuroImage*, 155:406–421, 2017.

S. Frässle, E. I. Lomakina, L. Kasper, Z. M. Manjaly, A. Leff, K. P. Pruessmann, J. M. Buhmann, and K. E. Stephan. A generative model of whole-brain effective connectivity. *NeuroImage*, 179:505–529, Oct. 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson. Detection of neural activity in functional mri using canonical correlation analysis. *Magnetic Resonance in Medicine*, 45(2):323–330, 2001.

O. Friman, M. Borga, P. Lundberg, and H. Knutsson. Exploratory fmri analysis by autocorrelation maximization. *NeuroImage*, 16(2):454–464, 2002.

K. Friston, K. H. Preller, C. Mathys, H. Cagnan, J. Heinzle, A. Razi, and P. Zeidman. Dynamic causal modelling revisited. *NeuroImage*, 2017.

K. J. Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1 (1):13–36, 2011.

K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994. doi: 10.1002/hbm.460020402.

K. J. Friston, O. Josephs, G. Rees, and R. Turner. Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39(1):41–52, 1998. doi: 10.1002/mrm.1910390109.

K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003. doi: 10.1016/S1053-8119(03)00202-7.

K. J. Friston, B. Li, J. Daunizeau, and K. E. Stephan. Network discovery with DCM. *Neuroimage*, 56(3):1202–1221, 2011. doi: 10.1016/j.neuroimage.2010.12.039.

K. J. Friston, J. Kahan, B. Biswal, and A. Razi. A DCM for resting state fMRI. *Neuroimage*, 94:396–407, 2014. doi: 10.1016/j.neuroimage.2013.12.009.

M. Girolami. Bayesian inference for differential equations. *Theoretical Computer Science*, 408(1):4–16, 2008. doi: 10.1016/j.tcs.2008.07.005.

M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124, 2013.

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

J. Gorski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007. doi: 10.1007/s00186-007-0161-1.

C. L. Grady, A. R. McIntosh, and F. I. Craik. Age-related differences in the functional connectivity of the hippocampus during memory encoding. *Hippocampus*, 13(5): 572–586, 2003.

M. M. Grant, K. Wood, K. Sreenivasan, M. Wheelock, D. White, J. Thomas, D. C. Knight, and G. Deshpande. Influence of early life stress on intra-and extra-amygdaloid causal connectivity. *Neuropsychopharmacology*, 40(7):1782, 2015.

S. Gugushvili, C. A. Klaassen, et al. $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. *Bernoulli*, 18(3):1061–1098, 2012. doi: 10.3150/11-BEJ362.

P. Haggard. Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, 9(12):934, 2008.

D. A. Handwerker, J. M. Ollinger, and M. D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651, 2004.

D. A. Handwerker, J. Gonzalez-Castillo, M. D'esposito, and P. A. Bandettini. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *Neuroimage*, 62(2):1017–1023, 2012a.

D. A. Handwerker, V. Roopchansingh, J. Gonzalez-Castillo, and P. A. Bandettini. Periodic changes in fmri connectivity. *Neuroimage*, 63(3):1712–1719, 2012b.

D. R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12): 2639–2664, 2004.

B. He and X. Yuan. On the o(1/n) convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

R. Henson and K. Friston. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*, chapter Convolution models for fMRI, pages 178–192. Academic Press London, UK, 2007. doi: 10.1016/B978-012372560-8/50014-0.

D. Hernandez-Stumpfhauser, F. J. Breidt, M. J. van der Woerd, et al. The general projected normal distribution of arbitrary dimension: modeling and bayesian inference. *Bayesian Analysis*, 12(1):113–133, 2017.

P. D. Hoff and X. Niu. A covariance regression model. *Statistica Sinica*, pages 729–753, 2012.

M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98, 2006.

J. V. Hull, L. B. Dokovna, Z. J. Jacokes, C. M. Torgerson, A. Irimia, and J. D. Van Horn. Resting-state functional connectivity in autism spectrum disorders: a review. *Frontiers in psychiatry*, 7:205, 2017.

S. E. Joel, B. S. Caffo, P. C. Van Zijl, and J. J. Pekar. On the relationship between seed-based and ica-based measures of functional connectivity. *Magnetic Resonance in Medicine*, 66(3):644–657, 2011.

N. A. Johnson. A dynamic programming algorithm for the fused lasso and l 0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104 (486):682–693, 2009.

J. Kang, F. D. Bowman, H. Mayberg, and H. Liu. A depression network of functionally connected regions discovered via multi-attribute canonical correlation graphs. *NeuroImage*, 141:431–441, 2016.

F. I. Karahanoğlu, C. Caballero-Gaudes, F. Lazeyras, and D. Van De Ville. Total activation: fMRI deconvolution through spatio-temporal regularization. *Neuroimage*, 73:121–134, 2013.

R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. Markov chain monte carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

M. S. Koyama, C. Kelly, Z. Shehzad, D. Penesetti, F. X. Castellanos, and M. P. Milham. Reading networks at rest. *Cerebral Cortex*, 20(11):2549–2559, 2010.

E. W. Lang, A. M. Tomé, I. R. Keck, J. Górriz-Sáez, and C. G. Puntonet. Brain connectivity analysis: a short survey. *Computational intelligence and neuroscience*, 2012:8, 2012.

H. Lee, D. S. Lee, H. Kang, B.-N. Kim, and M. K. Chung. Sparse brain network recovery under compressed sensing. *IEEE Transactions on Medical Imaging*, 30 (5):1154–1165, 2011. doi: 10.1109/TMI.2011.2140380.

L. Lee, K. Friston, and B. Horwitz. Large-scale neural models and dynamic causal modelling. *Neuroimage*, 30(4):1243–1254, 2006.

N. Leonardi and D. Van De Ville. On spurious and real fluctuations of dynamic functional connectivity during rest. *Neuroimage*, 104:430–436, 2015.

B. Li, J. Daunizeau, K. E. Stephan, W. Penny, D. Hu, and K. Friston. Generalised filtering and stochastic DCM for fMRI. *Neuroimage*, 58(2):442–457, 2011.

C.-s. R. Li, C. Huang, R. T. Constable, and R. Sinha. Imaging response inhibition in a stop-signal task: neural correlates independent of signal monitoring and post-response processing. *Journal of Neuroscience*, 26(1):186–192, 2006.

K. Li, L. Guo, J. Nie, G. Li, and T. Liu. Review of methods for functional brain connectivity detection using fmri. *Computerized Medical Imaging and Graphics*, 33(2):131–139, 2009.

C. H. Liao, K. J. Worsley, J.-B. Poline, J. A. Aston, G. H. Duncan, and A. C. Evans. Estimating the delay of the fmri response. *NeuroImage*, 16(3):593–606, 2002.

J.-a. Lin, H. Zhu, R. Knickmeyer, M. Styner, J. Gilmore, and J. G. Ibrahim. Projection regression models for multivariate imaging phenotype. *Genetic epidemiology*, 36(6):631–641, 2012.

M. A. Lindquist, J. M. Loh, L. Y. Atlas, and T. D. Wager. Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *Neuroimage*, 45(1):S187–S198, 2009. doi: 10.1016/j.neuroimage.2008.10.065.

G. Lohmann, D. S. Margulies, A. Horstmann, B. Pleger, J. Lepsien, D. Goldhahn, H. Schloegl, M. Stumvoll, A. Villringer, and R. Turner. Eigenvector centrality mapping for analyzing connectivity patterns in fmri data of the human brain. *PloS one*, 5(4):e10232, 2010.

S. Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

X. Luo, S. Zhang, S. Hu, S. R. Bednarski, E. Erdman, O. M. Farr, K.-I. Hong, R. Sinha, C. M. Mazure, and C.-s. R. Li. Error processing and gender-shared and-specific neural predictors of relapse in cocaine dependence. *Brain*, 136(4): 1231–1244, 2013. doi: 10.1093/brain/awt040.

T. M. Madhyastha and T. J. Grabowski. Age-related differences in the dynamic architecture of intrinsic networks. *Brain connectivity*, 4(4):231–241, 2014.

A. C. Marreiros, S. J. Kiebel, and K. J. Friston. Dynamic causal modelling for fMRI: a two-state model. *Neuroimage*, 39(1):269–278, 2008. doi: 10.1016/j.neuroimage.2007.08.019.

G. Marrelec, A. Krainik, H. Duffau, M. Pélégrini-Issac, S. Lehéricy, J. Doyon, and H. Benali. Partial correlation for functional brain interactivity investigation in functional MRI. *Neuroimage*, 32(1):228–237, 2006. doi: 10.1016/j.neuroimage.2005.12.057.

A. Mclntosh and F. Gonzalez-Lima. Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping*, 2(1-2):2–22, 1994. doi: 10.1002/hbm.460020104.

T. Melzer, M. Reiter, and H. Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *International Conference on Artificial Neural Networks*, pages 353–360. Springer, 2001.

M.-M. Mesulam. From sensation to cognition. *Brain*, 121(6):1013–1052, 1998.

B. Mwangi, T. S. Tian, and J. C. Soares. A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, 12(2):229–244, 2014.

M. E. Newman. Mathematics of networks. *The new Palgrave dictionary of economics*, pages 1–8, 2016.

W. D. Penny, K. E. Stephan, A. Mechelli, and K. J. Friston. Modelling functional integration: a comparison of structural equation and dynamic causal models. *Neuroimage*, 23:S264–S274, 2004.

W. D. Penny, K. E. Stephan, J. Daunizeau, M. J. Rosa, K. J. Friston, T. M. Schofield, and A. P. Leff. Comparing families of dynamic causal models. *PLoS Computational Biology*, 6(3):e1000709, 2010. doi: 10.1371/journal.pcbi.1000709.

R. Poldrack, E. Congdon, W. Triplett, K. Gorgolewski, K. Karlsgodt, J. Mumford, F. Sabb, N. Freimer, E. London, T. Cannon, et al. A phenome-wide examination of neural and cognitive function. *Scientific Data*, 3:160110, 2016. doi: 10.1038/sdata.2016.110.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

A. Poyton, M. S. Varziri, K. B. McAuley, P. McLellan, and J. O. Ramsay. Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers & Chemical Engineering*, 30(4):698–708, 2006. doi: 10.1016/j.compchemeng.2005.11.008.

M. G. Preti, T. A. Bolton, and D. Van De Ville. The dynamic functional connectome: state-of-the-art and perspectives. *Neuroimage*, 160:41–54, 2017.

X. Qi and H. Zhao. Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Annals of Statistics*, pages 435–481, 2010. doi: 10.1214/09-AOS724.

J. O. Ramsay. Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society. Series B*, 58:495–508, 1996.

J. O. Ramsay. *Functional data analysis*. Wiley Online Library, 2006. doi: 10.1002/0470013192.bsa239.

J. D. Ramsey, R. Sanchez-Romero, and C. Glymour. Non-Gaussian methods and high-pass filters in the estimation of effective connections. *Neuroimage*, 84:986–1006, 2014.

D. Rangaprakash, G.-R. Wu, D. Marinazzo, X. Hu, and G. Deshpande. Hemodynamic response function (HRF) variability confounds resting-state fMRI functional connectivity. *Magnetic Resonance in Medicine*, 2018.

A. Razi, J. Kahan, G. Rees, and K. J. Friston. Construct validation of a DCM for resting state fMRI. *Neuroimage*, 106:1–14, 2015.

A. Razi, M. L. Seghier, Y. Zhou, P. McColgan, P. Zeidman, H.-J. Park, O. Sporns, G. Rees, and K. J. Friston. Large-scale DCMs for resting state fMRI. *Network Neuroscience*, 2017. doi: 10.1162/NETN_a_00015.

B. J. Reich, J. Eidsvik, M. Guindani, A. J. Nail, and A. M. Schmidt. A class of covariate-dependent spatiotemporal covariance functions. *The annals of applied statistics*, 5(4):2265, 2011.

S. Reyt, C. Picq, V. Sinniger, D. Clarençon, B. Bonaz, and O. David. Dynamic causal modelling and physiological confounds: a functional MRI study of vagus nerve stimulation. *Neuroimage*, 52(4):1456–1464, 2010.

K. R. Ridderinkhof, M. Ullsperger, E. A. Crone, and S. Nieuwenhuis. The role of the medial frontal cortex in cognitive control. *Science*, 306(5695):443–447, 2004.

A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25(1):230–242, 2005. doi: 10.1016/j.neuroimage.2004.11.017.

P. Rossini, R. Di Iorio, M. Bentivoglio, G. Bertini, F. Ferreri, C. Gerloff, R. Ilmoniemi, F. Miraglia, M. Nitsche, F. Pestilli, et al. Methods for analysis of brain connectivity: an ifcn-sponsored review. *Clinical Neurophysiology*, 2019.

J. Rupnik and J. Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4, 2010.

K. E. Rutkowski. Closed-form expressions for projectors onto polyhedral sets in hilbert spaces. *SIAM Journal on Optimization*, 27(3):1758–1771, 2017.

S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage*, 54(2):807–823, 2011.

S. Ryali, T. Chen, K. Supekar, T. Tu, J. Kochalka, W. Cai, and V. Menon. Multivariate dynamical systems-based estimation of causal brain interactions in fMRI: Group-level validation using benchmark data, neurophysiological models and human connectome project data. *Journal of Neuroscience Methods*, 268:142–153, 2016a.

S. Ryali, Y.-Y. I. Shih, T. Chen, J. Kochalka, D. Albaugh, Z. Fang, K. Supekar, J. H. Lee, and V. Menon. Combining optogenetic stimulation and fMRI to validate a multivariate dynamical systems model for estimating causal brain interactions. *Neuroimage*, 132:398–405, 2016b.

Ü. Sakoğlu, G. D. Pearlson, K. A. Kiehl, Y. M. Wang, A. M. Michael, and V. D. Calhoun. A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(5-6):351–366, 2010.

A.-K. Seghouane and M. U. Khalid. Hierarchical sparse brain network estimation. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012. doi: 10.1109/MLSP.2012.6349756.

A. K. Seth. A MATLAB toolbox for Granger causal connectivity analysis. *Journal of Neuroscience Methods*, 186(2):262–273, 2010.

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

X. Shen. On the method of penalization. *Statistica Sinica*, 8(2):337–357, 1998.

S. M. Smith. The future of fMRI connectivity. *Neuroimage*, 62(2):1257–1266, 2012. doi: 10.1016/j.neuroimage.2012.01.022.

S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fMRI. *Neuroimage*, 54(2):875–891, 2011. doi: 10.1016/j.neuroimage.2010.08.063.

S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, et al. Resting-state fMRI in the human connectome project. *Neuroimage*, 80:144–168, 2013. doi: 10.1016/j.neuroimage.2013.05.039.

R. C. Sotero, A. Bortel, R. Martinez-Cancino, S. Neupane, P. O'CONNOR, F. Carbonell, and A. Shmuel. Anatomically-constrained effective connectivity among layers in a cortical column modeled and estimated from local field potentials. *Journal of Integrative Neuroscience*, 9(04):355–379, 2010. doi: 10.1142/S0219635210002548.

K. E. Stephan, L. Kasper, L. M. Harrison, J. Daunizeau, H. E. den Ouden, M. Breakspear, and K. J. Friston. Nonlinear dynamic causal models for fMRI. *Neuroimage*, 42(2):649–662, 2008. doi: 10.1016/j.neuroimage.2008.04.262.

K. E. Stephan, W. D. Penny, R. J. Moran, H. E. den Ouden, J. Daunizeau, and K. J. Friston. Ten simple rules for dynamic causal modeling. *Neuroimage*, 49(4):3099–3109, 2010. doi: 10.1016/j.neuroimage.2009.11.015.

T. H. Szatrowski et al. Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *The Annals of Statistics*, 8(4):802–810, 1980.

B. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–1165, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

E. C. Titchmarsh. The zeros of certain integral functions. *Proceedings of the London Mathematical Society*, 2(1):283–302, 1926. doi: 10.1112/plms/s2-25.1.283.

A. U. Turken and N. F. Dronkers. The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. *Frontiers in Systems Neuroscience*, 5:1, 2011.

M. P. Van Den Heuvel and H. E. H. Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20 (8):519–534, 2010.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3 (1):28–46, 1982. doi: doi.org/10.1137/0903003.

G. Verbeke, S. Fieuws, G. Molenberghs, and M. Davidian. The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research*, 23 (1):42–59, 2014.

T. Vincent, L. Risser, and P. Ciuciu. Spatially adaptive mixture modeling for analysis of fMRI time series. *IEEE Transactions on Medical Imaging*, 29(4):1059–1074, 2010.

T. Vincent, S. Badillo, L. Risser, L. Chaari, C. Bakhous, F. Forbes, and P. Ciuciu. Flexible multivariate hemodynamics fMRI data analyses and simulations with PyHRF. *Frontiers in Neuroscience*, 8:67, 2014.

S. Waaijenborg, P. C. V. de Witt Hamer, and A. H. Zwinderman. Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Statistical applications in genetics and molecular biology*, 7 (1), 2008.

J. Wang, X. Zuo, and Y. He. Graph-based network analysis of resting-state functional mri. *Frontiers in systems neuroscience*, 4:16, 2010.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624.

K. Wang, M. Liang, L. Wang, L. Tian, X. Zhang, K. Li, and T. Jiang. Altered functional connectivity in early alzheimer's disease: A resting-state fmri study. *Human brain mapping*, 28(10):967–978, 2007.

Y. Wang, O. David, X. Hu, and G. Deshpande. Can Patel's τ accurately estimate directionality of connections in brain networks from fMRI? *Magnetic Resonance in Medicine*, 78(5):2003–2010, 2017.

M. D. Wheelock, K. R. Sreenivasan, K. H. Wood, L. Ver Hoef, G. Deshpande, and D. Knight. Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *Neuroimage*, 102:904–912, 2014.

R. Wise, F. Chollet, U. Hadar, K. Friston, E. Hoffner, and R. Frackowiak. Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain*, 114(4):1803–1817, 1991.

D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

K. J. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and A. Evans. A general statistical analysis for fMRI data. *Neuroimage*, 15(1):1–15, 2002.

G.-R. Wu, W. Liao, S. Stramaglia, J.-R. Ding, H. Chen, and D. Marinazzo. A blind deconvolution approach to recover effective connectivity brain networks from resting state fMRI data. *Medical Image Analysis*, 17(3):365–374, 2013.

T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, pages 224–244, 2008.

Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

Y. Xu and W. Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, 2017.

H. Xue, H. Miao, and H. Wu. Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Annals of Statistics*, 38(4):2351, 2010.

L. Xue, S. Ma, and H. Zou. Positive-definite 1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491, 2012.

W. Yang, L. Cun, X. Du, J. Yang, Y. Wang, D. Wei, Q. Zhang, and J. Qiu. Gender differences in brain structure and resting-state functional connectivity related to narcissistic personality. *Scientific reports*, 5:10924, 2015.

B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. doi: doi:10.1152/jn.00338.2011.

A. Zalesky and M. Breakspear. Towards a statistical test for functional connectivity dynamics. *Neuroimage*, 114:466–470, 2015.

S. Zhang, J. S. Ide, and C.-s. R. Li. Resting-state functional connectivity of the medial superior frontal cortex. *Cerebral Cortex*, 22(1):99–111, 2011. doi: 10.1093/cercor/bhr088.

T. Zhang, J. Wu, F. Li, B. Caffo, and D. Boatman-Reich. A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *Journal of the American Statistical Association*, 110(509):93–106, 2015. doi: 10.1080/01621459.2014.988213.

Y. Zhao, B. Wang, S. Mostofsky, B. Caffo, and X. Luo. Covariate assisted principal regression for covariance matrix outcomes. *bioRxiv*, page 425033, 2018.

Y. Zhong, H. Wang, G. Lu, Z. Zhang, Q. Jiao, and Y. Liu. Detecting functional connectivity in fmri using pca and regression analysis. *Brain topography*, 22(2):134–144, 2009.

H. I. Zonneveld, R. H. Pruim, D. Bos, H. A. Vrooman, R. L. Muetzel, A. Hofman, S. A. Rombouts, A. van der Lugt, W. J. Niessen, M. A. Ikram, et al. Patterns of functional connectivity in an aging population: The rotterdam study. *NeuroImage*, 189:432–444, 2019.

T. Zou, W. Lan, H. Wang, and C.-L. Tsai. Covariance regression analysis. *Journal of the American Statistical Association*, 112(517):266–281, 2017.

X.-N. Zuo, R. Ehmke, M. Mennes, D. Imperati, F. X. Castellanos, O. Sporns, and M. P. Milham. Network centrality in the human functional connectome. *Cerebral cortex*, 22(8):1862–1875, 2011.

P. Zwiernik, C. Uhler, and D. Richards. Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1269–1292, 2017.