Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Shuo Chen Date

New Statistical Techniques for High-dimensional Neuroimaging Data

By

Shuo Chen

Doctor of Philosophy

Biostatistics

F. DuBois Bowman, Ph.D.
Advisor

Lance A. Waller, Ph.D.
Committee Member

Ying Guo, Ph.D.
Committee Member

Hu, Xiaoping, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

New Statistical Techniques for High-dimensional Neuroimaging Data

By

Shuo Chen
M.S., Emory University, 2012
M.S., East Tennessee State University, 2004
B.S., Harbin Institute of Technology, 2003

Advisor: F. DuBois Bowman, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

Abstract

In this dissertation, we focus on developing novel statistical methodology for high-dimensional neuroimaging data (HND) to yield insights about the complex neural processing characteristics associated with mental diseases and furthermore to provide clinically relevant predictive information to aid with treatment selection and prognosis. Our proposed methods would extract new information content from neuroimaging data, while coping with the many analytical challenges posed by these massive data sets. Specifically, we propose three new statistical frameworks: (i) to predict disease progression and therapeutic treatment response based on temporal trends in longitudinal neuroimaging data, e.g. repeatedly collected over weeks or months; (ii) to determine population-level brain networks revealed by functional (or structural) connectivity properties within regions of interest (ROI) and between ROIs, while characterizing whole-brain properties of these identified networks, such as the 'small world' property; (iii) to analyze resting-state functional magnetic resonance imaging (fMRI) data by constructing methodology to determine localized estimates of resting-state brain activity and simultaneously yield functional connectivity information based on spatial correlations in the data.

Recent technological advances have made it possible for many studies to collect HND repeatedly over time. Such studies may yield temporal changes in selected features that, when incorporated with machine learning methods, are able to predict clinical outcomes. However, current methods, such as the support vector machine (SVM), for HND analysis typically consider cross-sectional data collected during one time period. We propose a novel support vector classifier for longitudinal HND that allows simultaneous estimation of the SVM separating hyperplane parameters and temporal trend parameters, which determine the optimal means to combine the longitudinal data for classification and prediction. Our approach is based on an augmented kernel function in reproducing kernel Hilbert space and uses quadratic programming for optimization. The results of a simulation study and a data example indicate that our proposed method leverages the additional longitudinal information to achieve higher accuracy than methods using only cross-sectional data and methods that naively combine longitudinal data by simply stacking the data to expand the feature space.

Currently, the brain connectivity study methods are either based on seed voxel analysis or region representative analysis, which may both lead to substantial information loss due to partial selection or ignoring regional variation. We propose a comprehensive whole-brain voxel pair level and region pair level connectivity method by using a Bayesian hierarchical model which simultaneously accounts for brain network graph theory properties, while adjusting clinical covariate effects such as age,

gender, and medical history. This method can be applied to both functional and structural connectivity analysis, and yield brain connectivity inference on both voxel pair and region pair level as well as brain network graph theory metrics based on the posteriors. We illustrate the application of our method using functional connectivity analysis from a example fMRI data and simulation study.

Resting-state fMRI as a type of functional neuroimaging data is usually collected when individuals are left to think for themselves rather than engaging in a particular task. Since the traditional two-stage approach is not applicable as no stimuli events are included to detect task-related changes, the localized frequency band descriptors are used to detect localised brain activity. In addition, functional connectivity analysis is often used to identify the brain networks showing coherence during resting status. We propose a unified hierarchical Bayesian framework to jointly quantify localized resting-state brain activity as well as the functional connectivity brain network by allocating each region to a latent network cluster. Particularly, we utilize infinite mixture model and Dirichlet process for modeling the latent cluster. We conduct massive parameter estimation by using Markov Chain Monte Carlo (MCMC) techniques. The results based on the posteriors can yield inferences about the brain activity at voxel level and region level as well as the network of brain region parcellation. We apply the proposed method to depression study and the results reveal the level and region level frequency descriptor difference between groups as well as connectivity networks.

New Statistical Techniques for High-dimensional Neuroimaging Data

By

Shuo Chen
M.S., Emory University, 2012
M.S., East Tennessee State University, 2004
B.S., Harbin Institute of Technology, 2003

Advisor: F. DuBois Bowman, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

Acknowledgement

First and foremost, I would like to express my heartfelt gratitude to my advisor Dr. F. DuBois Bowman for his encouragement, guidance, and inspiration for the past years. As a nationally recognized and leading expertise in neuroimaging statistics, Dr. Bowman has guided me since my first year in the Department of Biostatistics at Emory, and he has constantly provided me timely suggestion, encouragement, and support to overcome the challenges during my Ph.D. study. His broad knowledge as well as insights and enthusiasm of neuroimaing statistics research have fostered my great interest in my research and made me determine to pursue scientific research as my lifetime career. He not only taught me how to become a statistician equipped with versatile and sophisticated analytical techniques, but also more importantly as a scientist with comprehensive and practical understanding of his area. I feel very fortunate to have him as my mentor and advisor.

I would also like to thank my dissertation committee members, Dr. Ying Guo, Dr. Xiaoping Hu, and Dr. Lance Waller, for their insightful comments and helpful discussions. Their expertise in neuroimaging statistics, neuroimaging technology, and Bayesian Statistics have led to substantial improvement in this dissertation. I owe special thanks to Dr. Waller for his foreseeing and insightful suggestions to solve the challenges of my Bayesian models and computational algorithms development .

I am also extremely grateful to be enrolled in and supported by the neuroimaging track of the Biostatistics in Genetics, Immunology, and Neuroimaging (BGIN) training program at Emory University led by Dr. Lance Waller. I feel very fortunate to be a member of the Center for Biomedical Imaging statistics (CBIS) at Emory led by Dr. Bowman. I would like to thank all the CBIS members especially to Dr. DuBois

Bowman, Dr. Ying Guo, Dr. Jian Kang, Dr. Lijun Zhang and Dr. Gordana Derado, as I I have learned so much from them through collaboration and discussion on numerous research projects. My thanks go also out to Dr. Helen Mayberg and her lab members for providing me rotation opportunity to learn neuroimage preprocessing techniques and expertise comments about brain anatomy of patients depression diseases. I would also to extend my thanks to Kirk Easley for his guidance on statistical consulting and Dr. Andrew Miller for the opportunity to participate in a psychiatric clinical trial as a statistical consultant. In addition, I am grateful to my dear friends at Emory for making my Ph.D. experience so enjoyable and memorable.

Last but not least, I would like to thank my family: my parents Lujun Chen and Hui Wang, my wife Zhen Zhang and my son Dongyu (Alex) Chen for their love, encouragement, patience, and belief in me. I also want to thank to my in-laws for their unconditional support. Without them, it would have been impossible for me to complete this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The emergence of in-vivo neuroimaging technology has provided a pathway to help improve our understanding of both the neurophysiology of healthy individuals and the pathophysiology of patients suffering from mental illnesses such as major depressive disorder (MDD). Currently, several neuroimaging techniques have been developed to reveal different perspectives of brain neural process. Those brain imaging techniques include the magnetic resonance imaging (MRI) based technology such as structural T1 image, functional MRI (fMRI), and diffusion tensor imaging (DTI), positron emission tomography (PET), and computed tomography (CT). All types of brain imaging data are subject to common properties such as high-dimensionality and complex spatial correlation structure, which pose challenges for statistical modeling. Therefore, our main objective is to develop new statistical methods to draw inference about neurophysiology from neuroimaging data and apply them for clinical purpose of disease diagnosis and treatment selection.

The dissertation is organized as follows: the remainder of Chapter 1 provides 2

background information on the human brain, various neuroimaging technologies, and current methods of neuroimaging data preprocessing and statistical analyses as well as outlines our proposed research objectives. Chapter 2 presents a novel classifier for longitudinal high-dimensional neuroimaging data. Chapter 3 presents methods to evaluate region level brain connectivity based on voxel pair connectivity for group study, and Chapter 4 presents a novel Bayesian hierarchical framework for resting-state fMRI data.

1.2 Basic Knowledge of Human Brain

The brain is the most complex organ in the human body, and it perceives the outside world, produces feeling, emotion, and memory, and coordinates human body actions. The average weight of adult human brain is about 1.4 kg, which contains by 80 or 90 billion non-neural cells (glial) and 80 or 90 billion neurons. In traditional brain anatomy, two brain tissue types are defined gray matter and white matter. The grey matter (cerebrum cortex) consists of neuronal cell bodies, neuropil, glial cells and capillaries. In contrast, the white matter mostly contains myelinated axons as tracts to interconnect different regions of the cerebral cortex and supporting structure.

Brain connectivity and signal transmitting. Neurons communicate with up to tens of thousands of other neurons, by passing signal via synapses. The pattern and strength of such connections keep changing for every second of our lives, based on updated experience, learning, and reinforcement. The changes of connectivities determine the brain functions such as memory, personality, and habit. Therefore, the brain structure is not only shaped by genes but also even more by experience.

There are various ways to pass signals between neurons, for example by electronic, magnetic, and chemical pathways. Among them, the neurotransmitter/receptor model

has been well studied. In synapse, signals are passed between neurons by releasing and capturing neurotransmitters such as dopamine, acetylcholine, and serotonin. The neurotransmitters are very important for brain activity, and abnormality of them is related to diseases. For example, a deficiency in serotonin in limbic system is linked to depression or mood disorders.

When building such electronic/chemical signal passing channels energy is needed, which is provided by glucose and oxygenated-haemoglobin (generating APT). Therefore, high level brain region activity is usually synchronized with higher metabolite rate and glucose/oxygenated-haemoglobin concentration.

Human brain parcellation and atlases. The human brain contains the brainstem, cerebellum, and cerebrum (neocortex). The cerebrum responds to higher-order reasoning, learning, and personality and is our major research interest. The cerebrum consists of two hemispheres (right and left) connected by white matter commissural fibers (e.g. corpus callosum). Each cerebral hemisphere is conventionally divided into four lobes: frontal, parietal, temporal, and occipital. The two hemispheres and four lobes provide us a general map of human brain, however, finer cerebral cortex parcellation is desired for in-depth study. One fine cerebral cortex parcellation is defined by Brodman areas (48 regions), which are based on cytoarchitecture, or organization of cells. The Automated Anatomical Labeling (AAL) regions (116 regions) were constructed through the identification of major and minor sulci/gyri on a T1 MRI with subsequent labeling based on anatomical location (Tzourio-Mazoyer *et al.*, 2002), which is more used for functional neuroimaging-based research.

For population brain imaging studies, the individual brain images are usually normalized into a common coordinate space to accommodate the between subject variation of brain size and orientation. The Talairach space and the Montreal Neurological Institute (MNI) space are the two most widely used atlas spaces. The Talairach coor-

dinate system is based on a 4 stereotaxic atlas of the human cerebral cortex published by Talairach and Tournoux (Talairach and Tournoux, 1988). Each brain location is defined by three dimensional coordinates with its distance from the midpoint of a brain white matter structure called the anterior commissure (AC). The atlas is built based on a single brain of a 60-year-old French woman with mental disorder. Despite of its popularity, the Talairach space is quite different from normal brains. Later, the MNI defined a new standard brain by using a large series of MRI scans of healthy normal controls (Evans *et al.*, 1993). These atlases differ in shape and size, and have been installed in common neuroimaging processing software.

1.3 Neuroimaging Techniques

Neuroimaging techniques measure brain structure and function non-invasively. There are two major categories of neuroimaging: *structural* and *functional* imaging. Structural imaging maps the brain anatomy at static status, for example diffusion tensor imaging (DTI) measures neural axons of white matter in the brain, T1 MRI provides high quality 3D images of brain. Functional imaging is designed to measure brain activity at dynamic status, for example, fMRI measures signal changes in the brain that are due to changing neural activity and PET measures various brain metabolism by using different radioactive tracers.

In this dissertation, we develop new statistical methods mainly for fMRI/MRI and PET data analysis. Therefore, in the following two subsections we give a brief description of these two neuroimaging techniques.

1.3.1 MRI and fMRI

Currently, MRI/fMRI bas become the most widely used neuroimaging technique due to its low invasiveness, low radiation exposure (radio level), and relatively wide availability. Through different MRI sequence parameters setups, MR signal can measure different tissue types (structural MRI) or neural metabolic changes (functional MRI). We first introduce the fundamental MRI technique.

Basic Principles of MRI

Magnetic Resonance Imaging rooted from the chemical technique known as Nuclear Magnetic Resonance (NMR) that is an effect whereby magnetic nuclei in a magnetic field absorb and re-emit electromagnetic energy with a specific frequency to the atom. The human body is largely composed of water molecules, and MRI is based on NMR of hydrogen protons. MRI provides good contrast between the different soft tissues of the body based on its water concentration (Higgins *et al.*, 1996).

The current MRI machines usually generate a magnetic field from 0.5 to 15 Tesla. As a compass aligns with the earth's magnetic field, the spinning protons placed near a large external magnetic field (B_0) align with or against the external field. The protons that align with the field and those that align against the field cancel each other out. A slight excess will align with the field so that the net result is an alignment with the external field. For instance at 1.5 Tesla, for every 2 million protons, there are only 9 more protons that align with than against the external field, but in a voxel of size 0.02 ml ($1.338 \times 10^2 1$ protons) it will include 6.02×10^{15} excess protons. We denote the magnetic field of the excess protons as M_0, accordingly we define the 3D space coordinates with M_0 is along the Z axis transverse to the XY plane.

At the quantum level, the spinning protons aligned with the external magnet field

are considered in the low energy state. If an electromagnetic radio frequency (RF) pulse is applied at the resonance (Larmor) frequency, then the protons can absorb that energy and jump to a higher energy state. Given B_0, one type of nuclei can only process at certain frequency which is determined by the Larmor equation. The magnetization vector M_0 of the processed protons spirals down toward the XY plane. The flip angle (FA) is a function of the strength and duration of the RF pulse. If the transmitted frequency does not match the natural resonance of the atom with given magnet, the proton will neither resonate nor send a signal.

Once the RF excitation is turned off, three events happen simultaneously: 1. the absorbed RF energy is retransmitted back at the resonance frequency and the excited protons fall back to the low energy status; 2. The excited spinning protons begin to return to align with the Z axis (M_0 direction) which relates to T_1 recovery; 3. The flipped magnetization vector (M_{xy}) of the excited protons begin to dephase (T2 and T2* relaxation) because of the spin-spin interaction and variation of magnet distribution.

T_1 weighted image

The longitudinal recovery rate is characterized by the time constant T_1, which is unique to every tissue. At a time $t = T_1$ after the excitation pulse, 63.2% of the magnetization has recovered. It is the uniqueness in M_z recovery rates that enables MRI to differentiate between different types of tissue (contrast). For example, water has less longitudinal magnetization prior to a RF pulse than fat due to its higher mobility, therefore it yields lower signal and appear dark on a T1 weighted image.

T_2/T_2^* weighted image

When the spins are first tilted down to the XY plane, they are all in phase. Because of the spin-spin interaction and local magnetic field inhomogeneity, the transverse magnetization tend to dephase and the signal decays. T_2 decay refers to the

exponential loss of signal resulting only from spin-spin interactions in the transverse or plane, while T_2^* decay is caused by both spin-spin interaction and local magnetic field inhomogeneity. The value of T_2 (and T_2^*) is unique for every kind of tissue and is determined primarily by its chemical environment with little relation to field strength. Using the the same example above, fat has a shorter T_2 time than water and decays faster than water, therefore, the amount of transverse magnetization left is less for fat than water, fat generates very little signal on a strong T2 weighted contrast image and appears darker than water.

MRI image construction

The 3D MRI image is usually composed by a series of 2D slices. Each 2D slice image is acquired by phase and frequency encoding with the magnet gradient, thus each voxel in a slice is localized by its own phase and frequency code. The readout signal is stored in temporary space called K-space. Then double inverse Fourier transform is applied to the K-space to transform the data into spatial image space. All subsequent analysis is performed on the transformed image.

Functional MRI

fMRI measures hemodynamic response of neural activity, as a surrogate of neural activity toward brain functions. The biological background of fMRI is based on the metabolism/energy consumption rate of neural activity. The neuron activity (firing) builds electronic and chemical gradients to pass signals, which consumes a great amount of energy in the form Adenosine-5'-triphosphate (ATP). Such metabolism increases the need for oxygen conveyed by oxyhemoglobin, and the human body will provide more oxyhemoglobin than consumed to the neural activity area, which leads to the increase of oxyhemoglobin concentration and decrease concentration of deoxyhemoglobin in the blood. The deoxyhemoglobin is paramagnetic and the higher

concentration will increase the inhomogeneity of the local magnet and lower the T_2^* signal. (Paramagnetism is the ability of an otherwise nonmagnetic material to exhibit magnetic properties in the presence of magnetic field.) Therefore, the areas with activated neurons have higher T_2^* signal that is measured by the MRI scans. The image intensity that varies with concentration of oxyhemoglobin and cerebrum blood flow (CBF) has been termed Blood Oxygenation Level Dependent (BOLD) and was first used in functional study of the brain by Ogawa *et al.*, 1990.

BOLD signal

For 3T MRI scan the BOLD signal increases 4% of the baseline signal at the activated area. The BOLD signal increases roughly as the square of the magnetic field strength. Also, there is time lag between the stimuli and BOLD signal, which is summarized by hemodynamic response function (HRF).

Pulse sequence design affect the image acquisition greatly, the two major parameters are repetition time (TR) and echo time (TE). TR is the time, in milliseconds (ms), between successive applications of RF pulses to a particular volume of tissue. It is impossible to measure the signal immediately after the RF is applied, due to hardware limitations. The short waiting time (also measured in ms's) during which the peak signal is obtained is called TE. The most common imaging sequence used in fMRI is the fast method of echo planar imaging (EPI), which allows collection of whole brain data in a few seconds or less.

Experiment design

There are two basic types of fMRI experiment design: event design and block design. In block design, the subjects maintain cognitive engagement in a task by presenting stimuli sequentially within a condition, and alternate with other moments (epochs) when a different condition is presented. In contrast, for the event design, the task is transient. If the two types of design are mixed, it is called mixed design.

fMRI studies yield large amounts of noisy data sets with complex spatio-temporal correlation structure, reflecting sophisticated neurophysiology and aspects of typical experimental designs. Such complexity poses analytical challenges for statistical modeling. Statistics plays a crucial role in understanding the underlying mechanism of the data both at individual and population level.

1.3.2 PET

PET is an analytical imaging technique, in which tracer compounds labeled with positron-emitting radionuclides are injected into the subject of the study. These tracer compounds can then be used to track biochemical and physiological processes *in vivo*. Using positron-emitting isotopes of elements such as carbon, nitrogen, oxygen and fluorine can create a range of tracer compounds which are similar to naturally occurring substances in the body.

PET physical principle

When the radioisotope undergoes positron emission decay, it emits a positron, which is an antiparticle of the electron and has opposite charge. The emitted positron tend to fly to interact with a electron, which annihilates both electron and positron and produces a pair of annihilation (gamma) photons moving in approximately opposite directions. The photons are detected when they reach a scintillator in the PET scanning device, creating a burst of light which is detected by photomultiplier tubes. Since the emitted positron travels in tissue for a short distance (generally less than 1 mm, the spatial resolution of PET image is lower than MRI (about 4 mm).

PET tracer mechanism

Before the PET scanning, a radiotracer is injected into the subject's bloodstream and its circulation within the human body. Usually, the uptake of the radiotracer

is correlated with desired physiology and the radiotracer is absorbed across capillary membrane to the cell. Unlike natural substances, the radiotracer will move out of the cell until radioactive decay due to certain chemical change. For example, the FDG radiotracer is correlated with glucose metabolism, and the high cell activity is associated with FDG uptake. During metabolism, FDG-6-phosphate is formed within the cell and it cannot move out of the cell before radioactive decay. Therefore, the scanner can detect the distribution of glucose uptake and activated area. The more trapped radiotracers correspond to higher activity level and leads to higher image intensity. Therefore, in brain studies functional activity can be recorded.

Provided with various radiotracer compounds, the PET is versatile to detect a number of physiology processes in the human body such molecular diffusion, protein synthesis, and receptor systems. For brain study it can not only measure CBF and glucose metabolism but also detect neuronal signal transmitter/receptor systems such as Dopamine. For functional brain study, the block design is usually employed as experiment design method.

1.4 Preprocessing and Data Analysis Methods for Neuroimaging Data

In this section, we give a brief review of current preprocessing and data analysis methods for neuroimaging data (mainly for fMRI and PET). In the data analysis pipeline, the raw neuroimaging data first go through preprocessing steps to become available for statistical analysis. The statistical analysis for neuroimaging data generally include: (i) activation studies attempt to identify brain areas that are source(s) of task-related neural processing, (ii) connectivity studies to detect what brain areas are activated synchronously over time to accomplish certain brain functions and (iii)

neuroimaging based biomarker selection and classification/prediction for the purpose of disease diagnosis and treatment selection.

1.4.1 Preprocessing Procedures

Several preprocessing steps for fMRI preprocessing are usually applied (in order): 1. brain extraction (BET) that strips the skull and non-brain tissues and generates a mask for the brain; 2. motion correction that realigns all scans (3D images) to a common reference (because subject will move during the fMRI experiment around 20 mins), usually by using rigid body transformation using 6 degrees of freedom; 3. slice timing correction (because each slice 2D image is scanned at a slightly different time and they are expected to be measured simultaneously) which moves each voxel's time course by interpolation and resampling; 4. spatial smoothing by convolving the 3D image with Gaussian kernel, which can increase signal to noise ratio if size of the blurring is less than size of activation; 5. temporal smoothing/band filtering that removes low frequency drifts and high frequency noise; 6. registration which attempts to register each subject's brain to a standard template brain atlas for example MNI space or Talairach space, using linear/affine transformation with 12 degree of freedom or nonlinear matrix transformation; 7. global intensity normalization which scales each subject's 4D dataset by a single value to get the overall 4D mean to be the same for all subjects. Preprocessing steps and the order in which they are performed are important since they affect both the spatial and temporal correlation structure of the data.

For PET data, Woods *et al.*, 1998 and Woods *et al.*, 1998 provide details regarding the preprocessing steps.

1.4.2 Statistical Modeling for Activation Studies

For activation analysis, usually a two-stage general linear model is employed. At the first level, the 4D datasets for a single subject are summarized by a 3D activation coefficient data set for each experiment condition; then the second level conducts group level analysis on the summarized coefficients while adjusting for the covariates such as age, gender, and race. In the following, we present methods for fMRI data, and PET analyses follow similar procedures. All the formulae in this section use the following notation: total number of voxels in the brain is V and each voxel is indexed by $v = 1 \ldots, V$, and the temporal series length is T and each time points is indexed by $t = 1, \ldots, T$, the number of subjects involved in the experiment is N and each subject is indexed by $i = 1, \ldots, N$.

Single-Subject Analysis

The first level models each voxel independently for one subject. For subject i and voxel v,

$$\mathbf{y}_i(v) = \mathbf{X}_i\mathbf{B}_i(v) + \mathbf{H}_{iv}\boldsymbol{\nu}_i(v) + \boldsymbol{\varepsilon}_i(v), \tag{1.1}$$

where where $\mathbf{y}_i(v) = (y_{iv1}, \ldots, y_{ivT})'$ is the temporal profile of the fMRI (T by 1 vector) and $\mathbf{X}_i$ represents the convolution of stimuli of l-th condition and HRF (T by L matrix) with the coefficients for all conditions as $\mathbf{B}_i(v) = (\beta_{iv1}, \ldots, \beta_{ivL})'$, $\mathbf{H}_{iv}$ represents the nuisance parameters such as high-pass filtering parameters if using a pth order polynomial function, and $\boldsymbol{\nu}_i(v) = (\nu_{vi1}, \ldots, \nu_{vip})'$; since the dependent variable is a time series and autocorrelated temporally, $\boldsymbol{\varepsilon}_i(v)$ is usually assumed to follow an AR(2) correlation structure (Worsley *et al.*, 1995, Friston *et al.*, 2002).

In (1.1), HRF is very important to fit the voxel activation coefficients toward

a stimuli. A canonical HRF is used by default which typically includes a gamma function with its first derivative and dispersion term. The secondary parameters could also be estimated by using a contrast vector $\mathbf{c}$ and $\theta_i = \mathbf{c}'\mathbf{B}_i$. For example, we can test for the increased brain activity on reading emotional words to describe selves and others for depression patients. The hypothesis tests and test statistics could be provided as in generalized linear models (GLM).

Group Level Analysis

The second level model is based on the stage 1 analysis result $\mathbf{B}_i(v)$ from (1.1), with a goal to estimate population level effects such treatment group, age, and gender for brain activation. The conventional model treats each voxel independently, and thus for the brain of total V voxels the second level analysis build V independent GLM to estimate the effects. The model for voxel v is

$$\mathbf{B}_i(v) = \mathbf{W}_{iv}\boldsymbol{\beta}(v) + \mathbf{e}_i(v) \tag{1.2}$$

where $\mathbf{W}_{iv}$ is the design matrix and $\boldsymbol{\beta}(v)$ the vector of parameters of interest. The whole brain inference is based the concatenation of the voxel level inference.

Although the independence between voxels assumption is convenient for computation and modeling, one key drawback for doing such is ignoring the spatial correlation structure of $(\mathbf{B}_i(v))$ impacting the efficiency of the estimation, since brain areas work interactively even for simple tasks. Bowman *et al.*, 2005 proposed a second level model to estimate localized activity that accounts for spatial dependencies between voxels within the same neural processing cluster defined by a data-driven clustering analysis. Derado *et al.*, 2010 extended it to a model that simultaneously accounts for spatial dependencies between voxels within the same anatomical region and for tem-

poral dependencies between a subject's estimates from multiple sessions (repeated measures over time in terms of $\mathbf{B}_i(v)$). Bowman *et al.*, 2008 developed a flexible Bayeisian hierarchical model that accounts for and estimates simultaneously the spatial correlations between voxels in the same anatomical regions as well as between distant regions. At the cost of complexity, those models achieve more efficiency and accuracy.

1.4.3 Connectivity Analysis

There are two types of connectivity based neuroimaging data: functional connectivity (mainly fMRI and PET based) and structural connectivity (mainly DTI based). Functional connectivity is defined as functional correlations between spatially remote neurophysiological events. Functional connectivity can include correlation of neural activity in resting status or based on stimuli. Functional connectivity is very important to study, since most brain functions are realized by functional connectivity of different neurons; and during deep sleep the functional connectivity is much weaker than during awake status. For fMRI data, functional connectivity analysis is based on a time series of BOLD signals. On the other hand, structural connectivity measures the white matter fibre tracts based on DTI imaging. The estimation is based on the path of water molecules' motion direction. It has intrinsic relation with functional connectivity because the neural signal is passed through axons in the white matter (Honey *et al.*, 2009). In the following, we mainly focus on the models for functional connectivity.

Seed Voxel Approach

The seed voxel approach first selects a voxel or a set of voxels (or ROI) based on functional or anatomical knowledge gained previously and then correlates its average time-course with the remaining voxels. Thus, the whole brain connectivities are built

based on the those reference voxels. The seed voxel approach is easy to calculate but the choice of the seed voxel could be subjective and it ignores network relationship between all voxels (or ROIs). Greicius*et al.*, 2003 applied this method to resting-state fMRI data for functional connectivity analysis, and revealed the connectivity between the posterior cingulate cortex and the ventral anterior cingulate cortex and provided evidence of a default mode networks (DMN) of brain function.

Distance-based Clustering

Distance-based clustering is an approach that can classify the V voxels in an image into G groups, with each cluster consisting of V_g voxels, where $g = 1, \ldots, G$ and $V = \sum_{g=1}^{G} V_g$. There are many distance (dissimilarity) metrics to measure correlation between the time series of a pair of voxels (or ROIs), such as Euclidean and correlation distance for continuous variables and Mahalonobis for categorical variables. The voxels within the same cluster are expected to have more coherent performance.

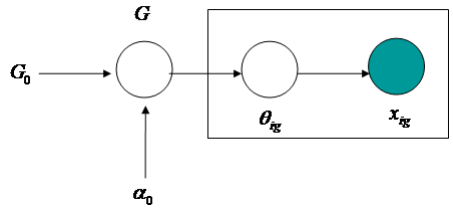
Based the distance matrix of all voxels, two types of algorithms usually are used for grouping: hierarchical clustering algorithms and partitioning algorithms. Hierarchical clustering starts from the closest two voxels then merges voxels and clusters with the shortest distance. Hierarchical clustering procedures vary due to the difference of the cluster distance function (linkage function), the common linkage functions are Ward's criterion, centroid linkage, variable linkage, median linkage, single linkage, and complete linkage algorithms (Bowman *et al.*, 2004a). Partitioning algorithms mainly refer to the K-means algorithm and its derivatives. The steps are 1. pre-specify the number of clusters at initial step, 2. randomly generate k clusters for all voxels and calculate the centers of the clusters, 3. reallocate voxels to clusters with the shortest voxel-centroid distance and generate new clusters and centroids, 4. repeat 2 and 3 until clusters are stable (Fadili *et al*, 2001). Rencher *et al.*, 2002 review current available clustering algorithms, and Bowman *et al.*, 2004b incorporates

several of these algorithms in neuroimaging data connectivity analysis.

Model-based Clustering

Traditional distance-based methods such as hierarchical clustering, k-means clustering, and self organizing map are simple and intuitive, but perform poorly with large datasets and are sensitive to noise. They may also require the number of clusters, K to be pre-specified. Model-based clustering methods are robust to the complications mentioned above, and do not require prespecification of K. Infinite mixture models (also known as the Dirichlet Process or Chinese Restaurant Process) allows simultaneous assignment of cluster membership along with optimization of the number of clusters, K. In the Chinese Restaurant Process (CRP) scheme, the brain regions are analogous to "customers" and the clusters to "tables". We devised a model-based clustering algoritm based on the CRP.

We used the nonparametric Bayesian infinite mixture model (see illustration below) and assumed the clustering of brain regions followed a Dirichlet process. ϕ_k is defined from G_0, which is a probability measure, $\pi_k \sim Dir(\frac{\alpha_0}{K}, \cdots, \frac{\alpha_0}{K})$, where α_0 is the predefined tuning parameter, $G = \Sigma_{k=1}^{G} \pi_k \delta_{\phi_k}$ is a random measure, and $\theta_{ig} \sim G$ is the parameter space defined from G, and $x_{ig} \sim P(. \mid \theta_{ig})$.



Independent Component Analysis (ICA)

Independent components analysis attempts to decompose functional neuroimaging data into sets of spatial components that are statistically independent sources with their associated time-courses (Common *et al.*, 1994). We denote the data signal as

a matrix X (V by by T) where V is the number of voxels, and T is the number of repeated scans. Then we assume that the matrix can be modeled as

$$X = AS + E \tag{1.3}$$

where the columns of A represent the component maps and the rows of S represent the corresponding time series of the component. E is a spatiotemporal Gaussian noise process (McKeown *et al.*, 1998). In spatial ICA, we assume that the columns of the mixing matrix A are statistically independent. It is possible to pursue the temporal independence of the target components.

One method of performing ICA analysis is based on the objective function to minimize the mutual information between components, which is an important concept in information theory, by using the 'infomax' algorithm (Calhoun *et al.*, 2000). An alternative is the fixed-point algorithm with the same goal of minimizing the mutual information, but also uses the concept of negentropy which can be described as a measure of non-normality c, and maximizes the negentropy to search directions of maximal-non-normality of X (Hyvarinen, 1999). The likelihood-based ICA algorithm has been proposed, which can also solve group level decomposition (Calhoun *et al.*, 2001,Guo *et al.*, 2008).

Structural Connectivity

For DTI based structural connectivity analysis, we refer to Behrens *et al.*, 2003 to calculate probabilistic tractography based on sampling from Bayesian estimated directions. Briefly the algorithm first estimates the probabilistic distribution of direction for each voxel by Bayesian methods, then applies n times (e.g. n=5000) sampling based on the direction posteriors from a seed voxel to determine where the next step goes by certain path growing rules; the connectivity of two voxels is calculated by the

number of paths from one to the other divided by n. Recently, more methods have been developed to combine information from both functional connectivity and structural connectivity, which leverages the advantages of each connectivity measurement (Honey *et al.*, 2009, Skudlarski *et al.*, 2008).

Graph Theory Property of Connectivity Network

Based on the fMRI, PET, and DTI and using algorithms previously mentioned, we can build the connectivity networks. The brain's structural and functional connectivity network has complex network topological features, such as high clustering, small-worldness, the presence of high-degree nodes or hubs, assortativity, modularity or hierarchy at both the whole-brain scale and local level, which are not presented in a random network.

The two fundamental elements of graph theory are node (vertex) and edge. For neuroimaging data, the nodes could be voxels or ROIs, and the edges represent the two connectivities between nodes (based on certain thresholds). The graph composed by the nodes and edges of brain network could provide an amount of metrics to describe the whole brain complex network.

The degree of a node is quantified by the number of edges that link it to the rest of the network. The distribution of the degrees of brain network generally follow non-Gaussian degree distributions, often with a long tail towards high degrees. In random networks the distribution follows naturally a Gaussian and symmetrically centred degree distribution. Assortativity is defined to describe the correlation between the degrees of connected nodes, and positive assortativity indicates that high-degree (hub) nodes tend to connect to each other.

The clustering coefficient describe the closeness between nodes within a cluster, by using the number of connections that exist between the nearest neighbors of a node

as a proportion of the maximum number of possible connections. Random networks have low average clustering whereas complex networks have high clustering.

Path length is the minimum number of edges that must be traversed for one node to travel to another. Efficiency is related to the inverse of path length and is calculated to estimate topological distances between nodes of disconnected graphs. Connection density is used to describe how busy is the network the actual number of edges in the graph as a proportion of the total number of possible edges from a complete graph.

The centrality of a node measures its importance by counting how many shortest paths between all other node pairs in the network pass through it, and thus a node with high centrality is crucial to efficient communication. Hubs nodes have high degree and high centrality. Provincial hub are connected mainly to nodes within the modules, whereas connector hubs are connected to nodes in other modules.

The small-worldness property combines high levels of within-module local connection density among nodes of and short paths (high efficiency) that connect globally distant nodes the network. This means that all nodes in a complex network can reach each other through relatively few intermediate steps. Small-worldness is calculated as the ratio of the clustering coefficient to the path length after both metrics with reference of their values to those in equivalent random networks (same nodes and edges). The small-worldness property exists widely in complex networks of genetics, communications, computational and neural networks.

Previous studies revealed that all networks found in nature and human-designed systems have non-random/non-regular properties which could be reflected by the above graph properties. Furthermore, those properties could be used for neuropathology analysis.

1.4.4 Classification and Prediction

The two above studies consider imaging measurement as dependent variables to explore brain physiology, but we can also utilize the imaging data as features to predict clinical outcomes such as disease status and treatment response (Evans *et al.*, 2006). To achieve this goal, feature selection and classifier construction are two major tasks.

Feature selection

Feature selection is a technique to choose a subset of most relevant features to the outcomes based on supervised learning models. Due to the "curse" of high-dimensionality, it is not computationally efficient and robust to use all available variables to build the model. Feature selection algorithms typically fall into two categories: filters and wrappers (Guyon *et al.*, 2003). The filter method ranks the features by a statistic and eliminates all features that do not pass the preset thresholds. Many methods have been developed to control the false positive discovery rate for the large scale tests, for example local FDR (Efron *et al.*, 2002).

The wrapper method is prediction/objective function oriented and it searches for the optimum set of possible features to achieve highest prediction accuracy or objective function value, for example the recursive feature elimination algorithm (Guyon *et al.*, 2002). The shrinkage method such Lasso and elastic network could also be categorized as wrappers, since the variables selected are based on the penalized objective function (Hastie and Tibshirani, 2004, Zou *et al.*, 2005). The selected features could be used as inputs of the following classification models.

Classification and prediction

Machine learning classification methods have been successfully applied to neuroimaging data, for example, Support-Vector Machines (SVM), Artificial Neural Network classifiers (ANN), decision tree based algorithm such us CART, C4.5, and ran-

dom forrest, and Bayesian classifiers such as naive Bayes classifiers and Bayesian networks with Markov blankets. Those supervised learning procedures usually include two steps: model training and prediction. In the training step, we use a part of the subjects to build the model (like the regression procedure) with the objective of high classification rate with certain constraints. In the prediction step, we test the rest of the subjects based on the trained model for evaluation and future use. The cross validation procedures such as k-fold and leave one out could be applied. Using such approaches, we are able to predict the mental state or treatment response and a stimulus class by analyzing the neuroimaging features such as activity and connectivity of neural responses.

The SVM is one of the most popular machine learning tools, which attempts to find the optimal hyperplane that can separate the two group of subjects and maximize the margins (Vapnik, 1996). The SVM algorithm is effective and robust in general. Cox *et al.*, 2003 applied statistical learning algorithms such as SVM and LDA to separate brain activation based fMRI data and can accurately predict participants who viewed different images (butterflies, chairs, birds, cows, etc.).

ANN is another popular model, which simulates the biological neural networks to train a statistical learning model (Hastie *et al.*, 2009). Usually, an ANN consists of multiple layers including a input layer, a hidden layer, and a output layer along with an activation function such as sigmoid function. The parameters to estimate are the weights of nodes in each layer, and one plugs the weighted sum through all layers to the activation to determine the classification and prediction. ANN with an acyclic graph is called a feed forward ANN, otherwise is recurrent ANN. The complexity of ANN increase with added hidden layers, and an ANN without any hidden layers is called single layer perceptron which is equivalent to logistic regression.

Decision trees are also widely applied. For the tree, each interior node corresponds

to one of the input variables, and the edges represent children for each of the possible values of that input variable with cutoff values; each leaf represents a value of the outcome. The learning process is repeated on each derived subset in a recursive manner called recursive partitioning. This model is called Classification And Regression Tree (Breiman, 1983). The random forrest is an ensemble classifier that consists of many decision trees, which is based on the "bagging" idea to average low bias but high variance weak classifiers to achieve high robustness and accuracy (Breiman, 2001).

Methodology for statistical learning is a very important and active research area. The ultimate goal is to use neuroimaging data to guide clinical decision making and biomarker findings. Also, the machine learning tools provide a possible path way to integrate clinical outcomes, neuroimaging data, and underly genomics and proteomics data.

1.5 Motivation Examples

1.5.1 An fMRI Resting-state Study of Depression

Resting-state fMRI data were acquired from 40 treatment naive subjects currently diagnosed with major depression disease (MDD), aged between 18-65 years, and with no significant psychological comorbidities or neurological disorders. The subjects were randomized into two antidepressant treatment groups – a cognitive behavioral therapy group (14 subjects) and a pharmacologic treatment group (14 subjects); and both groups underwent twelve weeks of active treatment. The fMRI scans were acquired at baseline (prior to treatment) and at approximately two weeks following the initiation of treatment.

The data were collected on a 3T Siemens scanner with a Z-saga sequence to avoid

orbitofrontal signal ablation. At each scanning session, 150 fMRI volumes were scanned in 7.5 minutes during visual fixation with 30 slices, field of view covering = 220 mm, voxel resolution of 3.4375mm x 3.4375mm x 4mm, TR= 2.920ms, TE = 30ms, and FA=90°. The data preprocessing steps include: motion correction, slice timing correction, normalization and spatial smoothing using 6mm Gaussian FWHM by using SPM5 and the VBM5 toolbox. Also, the de-trending and demeaning steps were applied to remove the incoherent background shift and texture variation. We scaled each regional time series by its ℓ^2 norm to adjust for differences in variability across regions and subjects.

1.5.2 A PET Study of Alzheimer's Disease

The Alzheimer's Disease Neuroimaging Initiative (ADNI) (`http://www.loni.ucla.edu/ADNI/`) is a large national project with the goal to develop biomarkers of Alzheimer's Disease (AD) in elderly subjects, to define the rate of progress of mild cognitive impairment and Alzheimer's disease, and to provide a large database which will improve design of treatment trials. We analyzed PET data from 73 healthy controls and 73 Alzheimer's disease patients, obtained from the ADNI database. Our goal is to predict subject's follow-up (6 month) brain activity, based on the baseline activity.

1.6 Proposed Research

The longitudinal HND provides the developing trend, besides cross-sectional information, of the brain neurophysiology to aid with clinical decisions such as disease prediction and treatment selection. Correspondingly, we seek to develop statistical machine learning methods specifically for classification and prediction by using longitudinal HND. Our new method incorporates the temporal trend estimation when

optimizing support vector classifier separating function parameters. Thus, the new objective function includes two sets of parameters and is solved by iterative quadratic programming. The new method utilizes classification and prediction accuracy.

We also plan to develop a Bayesian hierarchical model for brain connectivity networks analysis. The hierarchical model consists of three levels: voxel level, region level, and group level (e.g. patients and healthy controls). The base level of this framework is a mixture model of voxel pair connectivity distribution with consideration of the "small worldness" property of the brain networks. The top level employs generalized linear mixed model to model connectivities between ROIs while adjusting the individual's covariate effects such as age, gender, and psychiatric conditions, as well as accounting for the correlation structure between region pairs. In practice, we can apply this model to functional connectivity analysis (fMRI data), structural connectivity analysis (DTI data), or the weighted joint of the two modalities.

The resting-state fMRI data analysis strategy is different from task-induced fMRI data analysis due to the absence of stimuli. Therefore, for resting-state fMRI data analysis we first measure brain activity by fractional Amplitude of Low-Frequency Fluctuation (fALFF) based on frequency properties of the fMRI signal; then we conjecture a Bayesian hierarchical framework for modeling the measured activity. The proposed Bayesian hierarchical framework estimates voxel and ROI level brain activity for each clinical group with considering spatial correlation structure by using the functional connectivity information. The inference is drawn from the joint posteriors of model parameters.

Chapter 2

Topic 1: A Novel Support Vector Classifier for Longitudinal High-dimensional Data and Its Application to Neuroimaging Data

2.1 Introduction

Current biomedical technology enables the collection of high-dimensional data (HDD) to gain insights regarding genomic, proteomic, and *in vivo* neural processing properties. Moreover, such HDD are more commonly being collected longitudinally, potentially revealing changes in biological properties that may provide clues to disease diagnosis, progression, or recovery. Machine learning tools have been widely applied for HDD classification and prediction (Mitchell *et al.*, 2004; LaConte *et al.*, 2005 Chen *et al.*, 2007). Support vector machine methods are among the most popular machine learning techniques due to their high prediction accuracy and robustness (Vapnik,

1998; Mourao *et al.*, 2005;Fu *et al.*, 2008; Craddock *et al.*, 2009). However, most current machine learning methods have been developed for cross-sectional rather than longitudinal high-dimensional data (LHDD) analysis. The "ideal" methodology for LHDD would take advantage of the additional data to determine temporal trends of features and use them as inputs within machine learning models. However, in practice the temporal trends are usually unknown, and currently no such model exists for simultaneously determining the temporal trends and building the classification model.

To address classification or prediction objectives in context of LHDD, one may opt to use data from only a single time point, e.g. baseline data. Another potential approach for handling LHDD is a naive procedure of simply combining the longitudinal data as independent sources of information. Using data from only a single time point or using longitudinal data as independent sources of information may lead to substantial information loss and may not fully capitalize on the available data. One may also consider fitting preliminary models, for example, using logistic regression for each feature, and then using the resulting estimates to preset the temporal trends for classification. Since this approach uses classification outcome of interest in the preliminary modeling stage, presetting temporal parameters for each feature using model based estimates would lead to the vast danger of overfitting and pose difficulty for the following feature selection procedure.

In this paper, we propose a novel support vector classifier (SVC) for LHDD that extracts key features of each cross-sectional component as well as temporal trends between these components for the purpose of classification and prediction. The objective function of our new method incorporates two groups of estimands: the decision hyperplane function parameters and the temporal trend parameters that determine an optimal way to combine the longitudinal data. The objective function is derived from

maximizing the margin width, with error-tolerated correct classification constraints. Within the framework of the Lagrange (Wolfe) dual of the objective function, we augment the dimension of the Hessian matrix by incorporating the temporal trend parameters. Then, we apply quadratic programming techniques to optimize the classification parameters and temporal trend parameters. With the kernels satisfying Mercer's conditions, the objective function is convex, leading to a finite dimensional representation of the decision function. The framework allows feature selection with unknown temporal trend parameters through recursive feature elimination (RFE) procedures.

Generally, our proposed framework is applicable to any type of high dimensional data that are measured longitudinally. For example, in the application to neuroimaging data, our method is applicable to longitudinal/multi-session studies collecting fMRI, PET, EEG, and MEG data. The longitudinal property refers to multiple scanning sessions (e.g. images or collections of images acquired on different days). Importantly, for some neuroimaging data (e.g. fMRI data), there may be a series of images measured at different time points within one session. Therefore, we usually use features reflecting various summaries from the original data at each session. For example, the features in our method may include functional connectivity, localized activity summary statistics (first level analysis results for fMRI data), or frequency domain summary statistics. Hence with appropriate summary statistics, our approach can handle a range of high-dimensional data modalities.

The rest of the paper is organized as follows. In Section 2, we present the new longitudinal SVC and provide an accompanying computational strategy. Furthermore, we discuss its extension to nonlinear kernels and RFE based feature selection algorithm. In Section 3, we examine the classification performance of the proposed method for a data example and in a stimulation study. Section 4 concludes the paper

with a summary and a discussion of the major strengths of our novel SVC for LHDD.

2.2 Methods

2.2.1 Classical Support Vector Classifier

SVC is a popular kernel machine learning algorithm that is derived to solve classification problems (Vapnik, 1996). For one subject indexed by s, the p dimensional feature space is denoted as $\boldsymbol{x}_s \in \mathbb{R}^p$, for $s = 1, 2..., N$ and group indicators $y_s \in \{-1, 1\}$ denote a binary state such as disease status (positive/negative) or treatment response (recovery or not). A classifier is defined by constructing a separating function (or hyperplane) $h(\boldsymbol{x}_s) = \boldsymbol{w} \cdot \boldsymbol{x}_s + b$ and then generating $\hat{y}_i = sign(h(\boldsymbol{x}_s))$, if the data are linearly separable. The SVC chooses the unique hyperplane that maximizes the margins, which are the distances between the hyperplane and the support vectors. For cases when data are not linearly separable, a 'soft margin' is introduced that allows some data points to be misclassified. Therefore, the SVC is subject to optimize the following objective function:

$$\min_{\boldsymbol{w}} \ \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + \mathcal{C} \sum_{s=1}^{N} \xi_s \ \ s = 1, 2, ...N, \tag{2.2.1}$$

subject to

$$y_s(\boldsymbol{w} \cdot \boldsymbol{x}_s + b) \geqslant 1 - \xi_s, \ \text{ and } \xi_s \geqslant 0.$$

where ξ_s is the distance of the subject s from its correct side of the margin and the constraint constant C is the tuning parameter regarding the tolerance level of misclassification.

Then, we obtain the Lagrange (Wolfe) dual by substituting $\boldsymbol{w} = \sum_{s=1}^{N} y_s \alpha_s \boldsymbol{x}_s$ to the

Lagrange primal function of formula 2.2.1.

$$\min_{\alpha_s} \frac{1}{2}\sum_{s,s'} \alpha_s \alpha_{s'} y_s y_{s'} \langle \Phi(\boldsymbol{x}_s), \Phi(\boldsymbol{x}_{s'}) \rangle - \sum_{s=1}^{N} \alpha_s \text{ for } s \text{ and } s' = 1, 2, ...N, \tag{2.2.2}$$

subject to

$$C \geqslant \alpha_s \geqslant 0, \text{and } \sum_s \alpha_s y_s = 0.$$

where $\mathbf{K}(\boldsymbol{x}_s, \boldsymbol{x}_{s'}) = \langle \Phi(\boldsymbol{x}_s), \Phi(\boldsymbol{x}_{s'}) \rangle$ means that we first map data into a higher dimension through the function $\Phi(.)$, then take the inner product of the mapped vectors. The formula 2.2.2 could be expressed as:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{G} \boldsymbol{\alpha} - \mathbf{1}' \boldsymbol{\alpha} \tag{2.2.3}$$

where , $\mathbf{G}$ is a Gram matrix $(N \times N)$ satisfying the Mercer's condition (requiring $\mathbf{G}$ is at least semi positive definite) multiplied by corresponding group labels, $\boldsymbol{\alpha}$ is a 1 by N vector of estimands. $G_{s,s'}$ is $\langle \Phi(\boldsymbol{x}_s), \Phi(\boldsymbol{x}_{s'}) \rangle y_s y_{s'}$. Then, the 'Wolfe' dual is well suited for quadratic programming (QP) optimization programs in most software. The objective function in formula 2.2.3 can be also considered as the sum of penalty and loss functions in terms of reproducing kernel Hilbert space with $h(\cdot) = \sum_{s=1}^{N} \alpha_s y_s \mathbf{K}(\boldsymbol{x}_s, \cdot) \in \mathscr{H}_K$ (Wahba, 1990; Hastie and Tibshirani, 1990). Once the separating hyperplane has been determined through quadratic programming optimization, the class label of a new observation $\boldsymbol{x}_{new}$ can be determined by the sign function of $h(\boldsymbol{x}_{new}) = \sum_{s=1}^{N} \alpha_s y_s \mathbf{K}(\boldsymbol{x}_s, \boldsymbol{x}_{new}) + b$.

2.2.2 Longitudinal Support Vector Classifier - LSVC

Consider longitudinal data collected from N subjects at T measurement occasions or scanning sessions, with p features quantified during each session. The expanded feature matrix is then TN by p. Let $\boldsymbol{x}_{s,t}$ be used to represent the features collected for one subject s at time t. Hence, our aim is to classify each individual $\widetilde{\boldsymbol{x}}_s = \{\boldsymbol{x}_{s,1}, \boldsymbol{x}_{s,2}, ..., \boldsymbol{x}_{s,T}\}'$ to a certain group $y_s \in \{-1, 1\}$. We characterize linear trends of change: $\boldsymbol{x}_s = \boldsymbol{x}_{s,1} + \beta_1 \boldsymbol{x}_{s,2} + \beta_2 \boldsymbol{x}_{s,3} ... + \beta_{T-1} \boldsymbol{x}_{s,T}$, with unknown parameter vector $\boldsymbol{\beta} = (1, \beta_1, \beta_2, ..., \beta_{T-1})'$. The trend information is input into the SVC. A key challenge that we address is how to jointly estimate the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. We propose a novel longitudinal support vector classifier (LSVC) that jointly estimates the separating hyperplane parameters and the temporal trend parameters using quadratic programming. We present our approach using a simple linear kernel, but the ideas naturally extend to other kernel functions.

Let $\widetilde{\mathbf{X}}_m = [\widetilde{\mathbf{X}}_{t=1}, \widetilde{\mathbf{X}}_{t=2}, ..., \widetilde{\mathbf{X}}_{t=T}]'$ be a p by TN matrix, with components $\widetilde{\mathbf{X}}_{t=k} = (y_1 \boldsymbol{x}_{1,t=k}, y_2 \boldsymbol{x}_{2,t=k}, ..., y_N \boldsymbol{x}_{N,t=k})$ representing data from N subjects each with p features. The corresponding $\boldsymbol{\beta}_m$ is a TN by N matrix.

$$\begin{aligned} \mathbf{G} &= (\widetilde{\mathbf{X}}_{t=1} + \boldsymbol{\beta}_1 \widetilde{\mathbf{X}}_{t=2} + ... + \boldsymbol{\beta}_{T-1} \widetilde{\mathbf{X}}_{t=T})^T (\widetilde{\mathbf{X}}_{t=1} + \boldsymbol{\beta}_1 \widetilde{\mathbf{X}}_{t=2} + ... + \boldsymbol{\beta}_{T-1} \widetilde{\mathbf{X}}_{t=T}) \\ &= (\widetilde{\mathbf{X}}_m \boldsymbol{\beta}_m)^T (\widetilde{\mathbf{X}}_m \boldsymbol{\beta}_m) \\ &= \boldsymbol{\beta}_m^T \mathbf{G}_m \boldsymbol{\beta}_m, \end{aligned} \tag{2.2.4}$$

with

$$\mathbf{G}_m = \begin{bmatrix} \widetilde{\mathbf{X}}_{t=1}^T \widetilde{\mathbf{X}}_{t=1} & \cdots & \widetilde{\mathbf{X}}_{t=1}^T \widetilde{\mathbf{X}}_{t=T} \\ \vdots & \ddots & \vdots \\ \widetilde{\mathbf{X}}_{t=T}^T \widetilde{\mathbf{X}}_{t=1} & \cdots & \widetilde{\mathbf{X}}_{t=T}^T \widetilde{\mathbf{X}}_{t=T} \end{bmatrix}$$

and $\boldsymbol{\beta}_m^T = [\mathbf{I}_{N \times N}, \beta_1 \mathbf{I}_{N \times N}, \beta_2 \mathbf{I}_{N \times N}, ..., \beta_{T-1} \mathbf{I}_{N \times N}]$

Then, we denote $\boldsymbol{w}_{nv}$ as the estimate of separating hyperplane parameter in the classical SVC with inputs in the form of $\boldsymbol{x}_s = \boldsymbol{x}_{s,1} + \beta_1 \boldsymbol{x}_{s,2} + \beta_2 \boldsymbol{x}_{s,3} ... + \beta_{T-1} \boldsymbol{x}_{s,T}$. The primal objective function becomes

$$\min_{\boldsymbol{w}_{nv}} \ \frac{1}{2} \parallel \boldsymbol{w}_{nv} \parallel^2 + \mathcal{C} \sum_{s=1}^{N} \xi_s \ \ s = 1, 2, ...N \tag{2.2.5}$$

Similarly, with substituting $\boldsymbol{w}_{nv} = \sum_{s=1}^{N} y_s \alpha_s (\widetilde{\boldsymbol{x}}_s \boldsymbol{\beta}_m)^T$, we can reparameterize the Langrange (Wolfe) dual function as:

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2} \boldsymbol{\alpha}_m^T \mathbf{G}_m \boldsymbol{\alpha}_m - \mathbf{1}' \boldsymbol{\alpha} \tag{2.2.6}$$

with subject to

$$C \geqslant \boldsymbol{\alpha}_m(s) \geqslant 0,$$

$$\sum_{t}^{T} \sum_{s}^{N} \boldsymbol{\alpha}_m(s + (t-1)N) y_s = 0,$$

$$\text{for } s = 1, 2, ...N \text{ and } t = 1, 2, ...T - 1.$$

In this way, the model augments the dimension of $\mathbf{G}_m$ to TN by TN and the augmented kernel is ensured to be semi-positive definite. After $\boldsymbol{\alpha}_m$ is determined, the separating hyperplane parameter becomes

$$\boldsymbol{w}_{nv} = \Big[\sum_{s=1}^{n} \boldsymbol{\alpha}_m(s) \boldsymbol{x}_{s,1} y_s + \sum_{s=1}^{n} \boldsymbol{\alpha}_m(s+N) \boldsymbol{x}_{s,2} y_s, ... + \sum_{s=1}^{n} \boldsymbol{\alpha}_m(s + N(T-1)) \boldsymbol{x}_{s,T} y_s \Big]. \tag{2.2.7}$$

Defining the $1 \times T$ vector $\boldsymbol{\alpha}_{m,s} = (\boldsymbol{\alpha}_m(s), \boldsymbol{\alpha}_m(s+N), ..., \boldsymbol{\alpha}_m(s + (T-1)N))$ we

then have $\boldsymbol{w}_{nv} = \sum_{s=1}^{n} y_s \boldsymbol{\alpha}_{m,s} \widetilde{\boldsymbol{x}}_s$. In either case, we can notice that

$$\boldsymbol{w}_{nv} = \sum_{s=1}^{n} y_s \boldsymbol{\alpha}_m(s)(\boldsymbol{x}_{s,1} + \beta_1 \boldsymbol{x}_{s,2} + \beta_2 \boldsymbol{x}_{s,3} ... + \beta_{T-1} \boldsymbol{x}_{s,T}).$$

After obtaining $\boldsymbol{w}_{nv}$, we have $b = \frac{1}{N} \sum_{s=1}^{N} (\boldsymbol{w}_{nv} \cdot (\widetilde{\boldsymbol{x}}_s \boldsymbol{\beta}_m)^T - y_s)$, in which $\boldsymbol{\beta}_m$ can be estimated based on $\boldsymbol{\alpha}_m$. Hence, the separating hyperplane is

$$h(\widetilde{\boldsymbol{x}}) = \boldsymbol{w}_{nv} \cdot (\widetilde{\boldsymbol{x}} \boldsymbol{\beta}_m)^T + b. \tag{2.2.8}$$

The subjects with all $\boldsymbol{\alpha}_m > 0$ are considered as support vectors. Therefore, this method is different from directly applying SVC after stacking up the features at different times as independent features. In fact, this naive expansion of the feature space is a special case of LSVC with all $\beta = 1$.

Besides estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ vectors from $\boldsymbol{\alpha}_m$, we can alternatively employ an iterative procedure to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with respect to an objective function of 2.6. The algorithm will take T quadratic programming steps for each iteration. We rewrite the first part of the objective function in 2.6 as:

$$\boldsymbol{\alpha} \mathbf{G}_m^{0,0} \boldsymbol{\alpha} + \boldsymbol{\alpha} \boldsymbol{\beta}_m^T \mathbf{G}_m^{0,T} \boldsymbol{\alpha} + \boldsymbol{\alpha} \mathbf{G}_m^{0,T} \boldsymbol{\beta}_m \boldsymbol{\alpha} + \boldsymbol{\alpha} \mathbf{G}_m^{T,T} \boldsymbol{\beta}_m \boldsymbol{\alpha}, \tag{2.2.9}$$

where we denote

$$\mathbf{G}_m = \begin{bmatrix} \mathbf{G}_m^{0,0} & \mathbf{G}_m^{0,T} \\ \mathbf{G}_m^{T,0} & \mathbf{G}_m^{T,T} \end{bmatrix}. \tag{2.2.10}$$

For example, $\mathbf{G}_m^{0,0}$ $(N \times N)$ is the submatrix in the left top corner of the matrix $\mathbf{G}_m$ for the baseline data $(\widetilde{\mathbf{X}}_{t=1}^T \widetilde{\mathbf{X}}_{t=1})$.

Since the sum of convex functions is still convex, we only need to prove that the

objective function in 2.9 is convex with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We relegate the proof of convexity to the appendix. The convexity guarantees that the local minimum is also the global minimum and the solution for that minimum is unique. The algorithm is described as follows: (1) we start with initial values of $\boldsymbol{\beta}$ and use QP to optimize 2.9 to obtain $\boldsymbol{\alpha}$; (2) use the updated $\boldsymbol{\alpha}$ obtained in step 1 and apply QP again to estimate $\boldsymbol{\beta}$; (3) repeat the above two steps until convergence. The uniqueness of the solution leads to the convergence of the iterative algorithm.

2.2.3 Nonlinear Kernel Functions

Although the above derivations are considered in context of a linear kernel, it is natural to extend to nonlinear kernels. First, we can denote

$$\widetilde{\mathbf{K}}(\widetilde{\boldsymbol{x}}_s, \widetilde{\boldsymbol{x}}_{s'}) = \begin{bmatrix} \mathbf{K}(\widetilde{\boldsymbol{x}}_{s,1}, \widetilde{\boldsymbol{x}}_{s',1}) & \cdots & \mathbf{K}(\widetilde{\boldsymbol{x}}_{s,1}, \widetilde{\boldsymbol{x}}_{s',T}) \\ \vdots & \ddots & \vdots \\ \mathbf{K}(\widetilde{\boldsymbol{x}}_{s,T}, \widetilde{\boldsymbol{x}}_{s',1}) & \cdots & \mathbf{K}(\widetilde{\boldsymbol{x}}_{s,T}, \widetilde{\boldsymbol{x}}_{s',T}), \end{bmatrix} \tag{2.2.11}$$

and we have $< \beta\mathbf{K}(\cdot, \widetilde{\boldsymbol{x}}_{s,t}), \mathbf{K}(\cdot, \widetilde{\boldsymbol{x}}_{s',t}) >= \beta\mathbf{K}(\widetilde{\boldsymbol{x}}_{s,t}, \widetilde{\boldsymbol{x}}_{s',t})$, where $\mathbf{K}(\cdot, \widetilde{\boldsymbol{x}}_{s,t})$ indicates the reproducing kernel map of $\widetilde{\boldsymbol{x}}_{s,t}$. Therefore, the temporal trend is taken on the reproducing kernel mapped space which may be a set of nonlinear transformations of $\widetilde{\boldsymbol{x}}_{s,t}$, say $\mathbf{K}(\cdot, \widetilde{\boldsymbol{x}}_{s,t=0}) + \beta_1\mathbf{K}(\cdot, \widetilde{\boldsymbol{x}}_{s,t=1})..,+\beta_{T-1}\mathbf{K}(\cdot, \widetilde{\boldsymbol{x}}_{s,t=T-1})$.

Thus, the reproducing kernel function of separating hyperplane becomes

$$h(\widetilde{\boldsymbol{x}}) = \sum_{s=1}^{N} y_s \boldsymbol{\alpha}_m \widetilde{\mathbf{K}}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{x}}_s)\boldsymbol{\beta}_m^T + b, \tag{2.2.12}$$

where b is obtained by $b = \frac{1}{N}\sum_{s=1}^{N}\sum_{s'=1}^{N}(y_s y'_s \boldsymbol{\alpha}_m \widetilde{\mathbf{K}}(\widetilde{\boldsymbol{x}}_s, \widetilde{\boldsymbol{x}}_{s'})\boldsymbol{\beta}_m^T - y_s)$. In this way, the temporal trend parameter vector's length is increased in accordance with the dimen-

sion of features mapped. Hence, it also could be considered to estimate nonlinear temporal trends of the original features.

2.2.4 Feature Selection and Parameter Tuning

Feature selection is a critical step in supervised learning, as it can reduce the dimensionality of the feature space, leading to increased robustness, improved stability of the classifier, and reduced computational load. However, for longitudinal HDD feature selection based on 'filtering' may not be applicable because elements of $\boldsymbol{\beta}$ are unknown and thus no statistical test can be conducted for each feature. Nevertheless, 'wrapper' procedures such as SVC based recursive feature elimination (RFE) algorithm is valid under our new LSVC model. The SVC-RFE algorithm was first proposed by Guyon *et al.*, 2002, and it ranks all the features according to a classifier based weight function and eliminates one or more features with the lowest weights. This process is repeated until the minimal set of features achieve high classification accuracy. For a linear SVC, the weights are simply summarized from the $p \times 1$ vector $\boldsymbol{w}$. For non-linear kernel SVC, the rank of a feature is determined by the impact that its removal has on the variation of $\|\boldsymbol{w}\|^2$. In context of longitudinal HDD, the rank is determined by:

$$|\|\boldsymbol{w}_{nv}\|^2 - \|\boldsymbol{w}_{nv}^{-v}\|^2| = |\boldsymbol{\alpha}_m^T \mathbf{G}_m \boldsymbol{\alpha}_m - (\boldsymbol{\alpha}_m^{-v})^T \mathbf{G}_m^{-v} \boldsymbol{\alpha}_m^{-v}|, \tag{2.2.13}$$

where $\boldsymbol{w}_{nv}^{-v}$, $\boldsymbol{\alpha}_m^{-v}$, and $\mathbf{G}_m^{-v}$ are the estimates and inputs without feature or features v.

In addition, the tuning parameters such as cost C are also important, as they can affect the estimate of separating hyperplane parameters as well as temporal trend parameters. If we consider C as the level of shrinkage, and large C corresponds

to light regularization and small C stands for heavy regularization. Therefore, we can use the SVC path algorithm by starting with large C (low regularization) and increase it gradually, and observe the path of shrinkage in terms of $\boldsymbol{\alpha}_m(C)$ (Hastie and Tibshirani, 2004). This process will provide insight concerning the bias and variance trade off. For LSVC, we can utilize this shrinkage path algorithm to better estimate the temporal trend parameters.

2.3 Results

We investigate the performance of the proposed method by using simulation data and by using data from a longitudinal neuroimaging study.

2.3.1 Simulation study

To evaluate the performance of our proposed LSVC, we generate longitudinal data for 200 subjects and evenly divide them into two groups. Data for each subject includes $p = 100$ features at two time points ($T = 2$). We generate a group label $y_s \in \{-1, 1\}$ and features $\mathbf{x}_s$ for each subject. We also use a binary variable z to determine the baseline feature expression level, e.g. if $z_s = 1$ then $\mathbf{x}_s = \mathbf{1}$ otherwise $\mathbf{x}_s = \mathbf{0}$. Within each group, half of of the subjects have $z_s = 1$ at the lowest level of separability of the baseline data. If $z_s = y_s$, the baseline data are 100% separable. We then set up the temporal change variable Δ that depends on the group label y by letting $\Delta_s = \mathbf{1}$, if $y_s = 1$, otherwise $\Delta_s = \mathbf{0}$. Thus, different groups have different temporal trends. Therefore, the simulation is generated as follows:

$$\mathbf{x}_{s,t=1}|z = 0 \sim N(\mathbf{0}, \mathbf{I} \cdot \sigma^2),$$

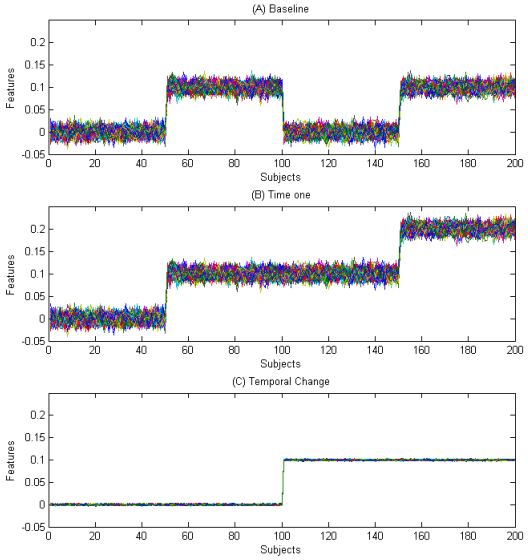
$$\mathbf{x}_{s,t=1}|z = 1 \sim N(\mathbf{1}, \mathbf{I} \cdot \sigma^2),$$

$$\Delta_s|y = -1 \sim N(\mathbf{0}, \mathbf{I} \cdot \tau^2),$$

$$\Delta_s|y = 1 \sim N(\mathbf{1}, \mathbf{I} \cdot \tau^2),$$

and $\mathbf{x}_{s,t=2} = \mathbf{x}_{s,t=1} + a \cdot \Delta$, where a is scalar to denote the magnitude and direction of the change. In this simulation, we use $\sigma^2 = 0.01$, $\tau^2 = 0.001$, and $a = 1$ to generate the data. The generated data is depicted in Figure 1, with the x-axis indicating the subject number (the first 100 subjects are in group one, the rest are in group two), and the y-axis indicating the feature expression level. The three subplots describe baseline, time one, and the temporal trend.

Figure 2.1: Simulated Data Set: (A) Baseline data, (B) Time one data, and (C) Temporal change



We test the performance of the model using different parameter and separability conditions. The variance has little influence on the model if the data are not separable, but separability definitions do impact the results. Therefore, we consider four methods: SVC based on baseline data, SVC based on both baseline and time one data

stacked and treated as independent (i.e. no temporal trends), our proposed LSVC, and SVC with a known trend. We test these methods using separability levels of 50%, 60% and 70%. The separability level between groups could also be considered as a function of the variation between subjects within each group, where a lower level of separability between groups results from higher between-subjects (in one group) variation. Also, we run an additional simulation by introducing different random "subject" effects upon the features both at baseline and time 1. The random effects are assumed to have zero mean and variance σ^2, $2\sigma^2$, and $5\sigma^2$ (totally blurring), and the higher level of noise leads to lower level of SNR ratio. We then evaluate the performances of the classifiers under different levels of SNR ratios (see table 2). For all cases, we only consider linear kernels for equitable comparisons. In addition, for the tuning parameter C there is no closed form estimator though cross validation can be applied to assist in determining the best-performing value of C from a pre-specified list of values (Hastie and Tibshirani, 2004). We feel that generating prediction results based on different levels of C provides a better evaluation of the SVC's performance when comparing different models.

We present the accuracy results (and standard deviation) for each method and for each simulation setting in Table 1. The results indicate that our LSVC has excellent performance, which is comparable to the 'oracle' model with perfect accuracy in our simulation example. Here, SVC with 'oracle' represents the SVC as if the temporal trend is known and maximal information is obtained for LHDD. The traditional SVC performs very poorly across all simulation settings.

Similarly, Table 2 shows the results for the simulated data with 50% separability and different levels of noise. When the noise level is low, LSVC performs better than traditional methods and approximates the 'oracle' model. When we double the original noise level considered, i.e. the noise is taken to be $2\sigma^2$, the LSVC still performs

quite well and shows marked improvements over the traditional SVC. We also consider a case where the noise level saturates the signal, specifically $5\sigma^2$. Although this case is not likely to arise in practice, we wanted to evaluate the performance of our method under extreme conditions. Naturally, the performance of our model declines in this setting, but it still outperforms the conventional SVC approaches.

Table 2.1: Simulation Classification Results with Different Separability

Cost (C)	SVC baseline	SVC stack	LSVC	SVC 'oracle'
50% separable at baseline				
0.1	.49 (0.06)	.51 (0.04)	.99 (0.01)	1(0.0)
1	.52 (0.03)	.50 (0.03)	1 (0.00)	1(0.0)
100	.50 (0.02)	.53 (0.04)	1 (0.01)	1(0.0)
10000	.53 (0.03)	.48 (0.06)	.99 (0.03)	1(0.0)
60% separable at baseline				
0.1	.57 (0.16)	.71 (0.31)	1 (0.0)	1(0.0)
1	.52 (0.23)	.75 (0.11)	1 (0.0)	1(0.0)
100	.58 (0.12)	.73 (0.14)	1 (0.01)	1(0.0)
10000	.63 (0.08)	.72 (0.26)	1 (0.02)	1(0.0)
70% separable at baseline				
0.1	.72 (0.13)	.83 (0.04)	1 (0.01)	1(0.0)
1	.78 (0.07)	.87 (0.03)	1 (0.02)	1(0.0)
100	.73 (0.12)	.82 (0.04)	1 (0.02)	1(0.0)
10000	.71 (0.21)	.81 (0.06)	1 (0.06)	1(0.0)

2.3.2 Data Example

We analyze data from the ADNI database (www.loni.ucla.edu/ADNI), which includes longitudinal PET scans acquired at baseline, 6 months, and 12 months. We used data from 80 subjects, 40 Alzheimer's disease (AD) patients and 40 healthy controls, ages 62 to 84. We used SPM5 for data preprocessing. We illustrate our longitudinal SVC procedure using PET scans from baseline and 12 months.

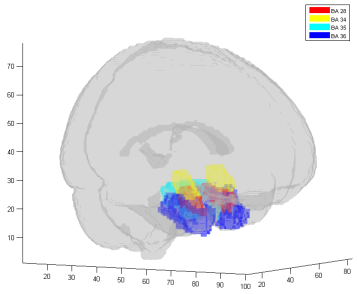
We use 1877 voxels within AD relevant regions of interest (ROI) as features, for

Table 2.2: Simulation Classification Results with Different Noise Level

Cost (C)	SVC baseline	SVC stack	LSVC	SVC "oracle"
		Noise: σ^2		
0.1	.47 (0.13)	.54 (0.07)	.98 (0.03)	1(0.0)
1	.56 (0.09)	.50 (0.12)	.99 (0.01)	1(0.0)
100	.51 (0.06)	.61 (0.14)	.99 (0.01)	1(0.0)
10000	.52 (0.10)	.55 (0.08)	.99 (0.01)	1(0.0)
		Noise:$2\sigma^2$		
0.1	.57 (0.16)	.55 (0.21)	.96 (0.03)	1(0.0)
1	.52 (0.23)	.55 (0.14)	.96 (0.02)	1(0.0)
100	.48 (0.12)	.52 (0.18)	.98 (0.01)	1(0.0)
10000	.55 (0.08)	.49 (0.21)	.97 (0.02)	1(0.0)
		Noise: $5\sigma^2$		
0.1	.48 (0.33)	.47 (0.28)	.56 (0.22)	.73 (0.11)
1	.54 (0.17)	.52 (0.23)	.61 (0.18)	.68 (0.20)
100	.53 (0.22)	.51 (0.26)	.54 (0.15)	.70 (0.17)
10000	.51 (0.24)	.49 (0.16)	.58 (0.06)	.62 (0.13)

example the hippocampus and entorhinal cortex (see Figure 2). Based on voxels within selected ROIs, we applied our novel longitudinal SVC to discriminate healthy and AD groups. Our goal here is not to chase perfect classification of accuracy through tuning parameters and feature selection, rather we demonstrate the usage of the proposed method and compare with the alternatives. For the validation procedure, we choose leave one out cross-validations. We tested on the data by using three

Figure 2.2: Voxels in these ROIs are used for analysis



classification methods SVC with baseline session only, SVC with two sessions stacked independently (N by $2p$); and the proposed LSVC. Also, different kernels are used. The results show that the accuracies for the two alternative methods across all costs are around 50% when polynomial (degree 2, 3, 5, 10) and Gaussian kernels (with various values of σ) are used. We tune the cost parameter across C = (0.1, 1, 100, 10000). The accuracies are listed in Table 3 based on a leave-one-out cross validation across all costs. In general the accuracy of LSVC method is 10 to 15 percent higher than the other two alternative methods.

Table 2.3: ADNI PET Data Classification Results

Cost (C)	SVC baseline	SVC stack	LSVC
0.1	.65	.66	.78
1	.66	.67	.76
100	.65	.67	.75
10000	.66	.66	.75

Overall, based on the simulation study and neuroimaging data analysis, our proposed method outperforms the traditional methods.

2.4 Discussion

In this article, we present a novel support vector classifier for LHDD. Our proposed method estimates decision function parameters and longitudinal parameters simultaneously using quadratic programming. The classifier can be extended to any kernel that satisfies Mercer's condition, and then the temporal trend is based on the nonlinear transformations of the original feature space. The SVC-RFE feature selection procedure can also be conducted in our LSVC, with ranking weight based on the width of the separating margins.

We apply the proposed method to longitudinal neuroimaging data which is a type of LHDD with temporal and spatial correlation structure. A growing literature has addressed the issue of temporal and spatial correlation when modeling neuroimaging data as dependent variables (Bowman *et al.*, 2008, Derado *et al.*, 2010). However, in our model the LHDD represent independent variables, and the group label for each subject is the dependent variable, and usually we do not explicitly account for correlations of the predictors. Note that we model the temporal trend for the LHDD to account for the temporal correlations introduced by the longitudinal experimental design. For fMRI data, since we use the first level analysis results as features, the scan to scan temporal correlation is considered in the first level analysis using conventional approaches such as prewhitening or precoloring.

In our data example, we use biological information to effectively reduce the number of features from around 300,000 to 1,877, rather than performing variable selection empirically. When such biological information is not present, some supervised methods are applicable. Based on the results from our simulation study and our data example, the LSVC leverages the additional information from longitudinal measurements to achieve higher prediction accuracy. The computational load of our LSVC technique is generally quite manageable, and on average training a LSVC model of 200 subjects with 100 features takes roughly 14 minutes on a PC with Intel Core2 Duo 2.83G CPU and 4G memory.

2.5 Appendix: Proof of Convexity of Objective function w.r.t. α and β

Proposition. $f = \boldsymbol{\alpha}_m^T \mathbf{G}_m \boldsymbol{\alpha}_m$ is a convex function regarding $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\alpha}_m = (\boldsymbol{\alpha}, \beta_1\boldsymbol{\alpha}, ..., \beta_{T-1}\boldsymbol{\alpha})$.

Proof. The second order condition of convexity requires the Hessian matrix $\nabla^2 f$ to be positive semidefinite (p.s.d.). and

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial \boldsymbol{\alpha}^2}, \frac{\partial^2 f}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}} \\ \frac{\partial^2 f}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}}, \frac{\partial^2 f}{\partial \boldsymbol{\beta}^2} \end{pmatrix}.$$

Here we first present the case of two time points and extend it to T time points. Therefore,

$$f = \begin{pmatrix} \boldsymbol{\alpha} \\ \beta\boldsymbol{\alpha} \end{pmatrix}^T \begin{bmatrix} \mathbf{G}_m^{0,0} & \mathbf{G}_m^{0,1} \\ \mathbf{G}_m^{1,0} & \mathbf{G}_m^{1,1} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha} \\ \beta\boldsymbol{\alpha} \end{pmatrix}$$

where $\mathbf{G}_m^{0,0} = \widetilde{\mathbf{X}}_{t=1}^T \widetilde{\mathbf{X}}_{t=1}$, $\mathbf{G}_m^{0,1} = \widetilde{\mathbf{X}}_{t=1}^T \widetilde{\mathbf{X}}_{t=2}$, $\mathbf{G}_m^{1,0} = \widetilde{\mathbf{X}}_{t=2}^T \widetilde{\mathbf{X}}_{t=1}$and $\mathbf{G}_m^{1,1} = \widetilde{\mathbf{X}}_{t=2}^T \widetilde{\mathbf{X}}_{t=2}$.

Then, the four derivatives are:

$$\begin{aligned} \frac{\partial^2 f}{\partial \boldsymbol{\alpha}^2}_{(N\times N)} &= \mathbf{G}_m^{0,0} + \beta\mathbf{G}_m^{0,1} + \beta\mathbf{G}_m^{1,0} + \beta^2\mathbf{G}_m^{1,1} \\ \frac{\partial^2 f}{\partial \boldsymbol{\alpha} \partial \beta}_{(N\times 1)} &= (\mathbf{G}_m^{0,1} + \beta\mathbf{G}_m^{1,1})\boldsymbol{\alpha} \\ \frac{\partial^2 f}{\partial \boldsymbol{\beta} \partial \beta}_{(1\times N)} &= \boldsymbol{\alpha}^T(\mathbf{G}_m^{0,1} + \beta\mathbf{G}_m^{1,1}) \\ \frac{\partial^2 f}{\partial \beta^2}_{(1\times 1)} &= \boldsymbol{\alpha}^T\mathbf{G}_m^{1,1}\boldsymbol{\alpha} \end{aligned}$$

Next, we need to prove $\nabla^2 f$ is p.s.d. For any nonzero vector $\mathbf{v}$ of length N and

scalar u,

$$\begin{pmatrix} \mathbf{v} \\ u \end{pmatrix}^T \begin{pmatrix} \frac{\partial^2 f}{\partial \boldsymbol{\alpha}^2}, & \frac{\partial^2 f}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}} \\ \frac{\partial^2 f}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}}, & \frac{\partial^2 f}{\partial \boldsymbol{\beta}^2} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ u \end{pmatrix}$$

$$= \mathbf{v}^T \mathbf{G}_m^{0,0} \mathbf{v} + \mathbf{v}^T \mathbf{G}_m^{0,1} \boldsymbol{\alpha} u + \beta \mathbf{v}^T \mathbf{G}_m^{0,1} \mathbf{v} + u \boldsymbol{\alpha}^T \mathbf{G}_m^{0,1} \mathbf{v} + \mathbf{v}^T \mathbf{G}_m^{1,0} \mathbf{v} \beta$$

$$+ u \boldsymbol{\alpha}^T \mathbf{G}_m^{1,1} \mathbf{v} \beta + \beta \mathbf{v}^T \mathbf{G}_m^{1,1} \boldsymbol{\alpha} u + \beta \mathbf{v}^T \mathbf{G}_m^{1,1} \mathbf{v} \beta + u \boldsymbol{\alpha}^T \mathbf{G}_m^{1,1} \boldsymbol{\alpha} u$$

$$= [\widetilde{\mathbf{X}}_{t=2}(\beta \mathbf{v} + u\boldsymbol{\alpha}) + \widetilde{\mathbf{X}}_{t=1}\mathbf{v}]^T [\widetilde{\mathbf{X}}_{t=2}(\beta \mathbf{v} + u\boldsymbol{\alpha}) + \widetilde{\mathbf{X}}_{t=1}\mathbf{v}] + \beta \mathbf{v}^T \mathbf{G}_m^{1,1} \mathbf{v} \beta + u \boldsymbol{\alpha}^T \mathbf{G}_m^{1,1} \boldsymbol{\alpha} u \geq 0$$

because $\mathbf{G}_m^{1,1}$ is p.s.d.

Similarly for T time points data set, the Hessian matrix

$$\begin{aligned} \nabla^2 f = & [\widetilde{\mathbf{X}}_{t=1}\mathbf{v} + \sum_{k=1}^{T-1} \widetilde{\mathbf{X}}_{t=k+1}(\beta_k \mathbf{v} + u_k \boldsymbol{\alpha})]^T [\widetilde{\mathbf{X}}_{t=1}\mathbf{v} + \sum_{k=1}^{T-1} \widetilde{\mathbf{X}}_{t=k+1}(\beta_k \mathbf{v} + u_k \boldsymbol{\alpha})] \\ & + \sum_{k=1}^{T-1} u_k \boldsymbol{\alpha}^T \mathbf{G}_m^{k,k} \boldsymbol{\alpha} u_k + \sum_{k=1}^{T-1} \beta_k \mathbf{v}^T \mathbf{G}_m^{1,1} \mathbf{v} \beta_k \end{aligned}$$

is also p.s.d.

Moreover, the objective functions with nonlinear kernels are also convex if each $\widetilde{\mathbf{K}}(\widetilde{\mathbf{X}}_{t=k}, \widetilde{\mathbf{X}}_{t=k})$ follows Mercer's condition and is p.s.d. for $k = 1, 2, ..., T$.

Chapter 3

Topic 2: Bayesian Hierarchical Model for Comprehensive Brain Connectivity Analysis

3.1 Introduction

Functional connectivity in the human brain refers to the inter-links between neuronal processing units. Specific patterns of functional connectivity are linked to corresponding actions, emotions, and cognitions, and disruptions of these functional connectivity patterns are associated with pyschiatric and neurological disorders. Modern neuroimaging technologies including functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) provides a pathway to detect brain connectivity (Jirsa *et al.*, 2007; Sporns *et al.*, 2010, Lindquist, 2008; Caffo *et al.*, 2010). For example, the fMRI technique enables the assessment of functional connectivity by measuring temporal coherence between distinct neural processing units within the human brain. Furthermore, the brain connectivity can be used at a population level to explore its

association with age, gender, disease status (e.g. normal vs. Alzheimer disease or schizophrenia), and different treatment responses. For functional connectivity, there are several connectivity metrics available to measure the coherence of the temporal profiles such as Pearson correlation, spectral coherence, and dynamic causal modeling (Sun *et al.*, 2004; Friston *et al.*, 2003; Ombao *et al.*, 2008; Freyermuth *et al.*, 2010).

There are several intrinsic challenges to evaluating functional connectivity in the brain. Typical neuroimaging data in standard 2mm MNI space usually includes roughly 300 thousand voxels, therefore more than 40 billion voxel pairs for whole brain connectivity analysis. In addition, the voxel pairs are not independent and their covariance matrix would include 2.3×10^{26} parameters. If voxel pair level connectivity analysis is conducted for the whole brain, such ultra-high dimensionality will pose an infeasible computational load for parameter estimation. Therefore, the current methods either (i) choose a small set of seed voxels among all voxels and associate the rest of the voxels with the seed voxels or (ii) quantify region pair level connectivity that first summarizes all voxels within each region by taking averages (or similar dimension reduction), and then measuring the connectivity by using the regional representatives (Greicius*et al.*, 2003).

The seed voxel approach is simple for computation, but it is inherently limited in scope as it ignores connectivities other than the ones associated with the seeds and does not account for the spatial correlation structure between connectivity unit pairs. The region based method neglects the variations within regions that may lead to substantial bias and information loss. Recently, there has been a focus on the application of graph theory metrics to depict complex brain networks mostly by using regions' representatives (Bullmore, 2009). Graph theory metrics reflect important topological properties of brain networks, for example the "small-worldness" property, which indicates that most neural units are only highly likely to be connected with

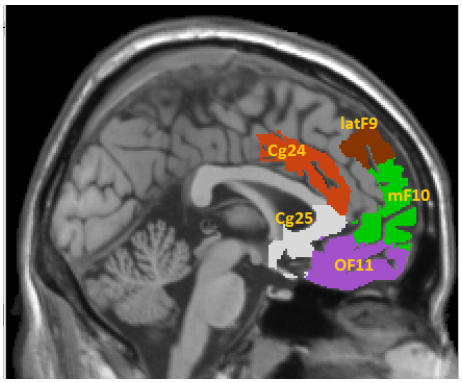
their neighbors, but they may achieve the distant connnections, typically through only a few hub neural units (Achard *et al.*, 2005). The "small-worldness" property is common in highly organized networks such as the internet, vehicle traffic, and social networks. Nevertheless, those metrics usually only function as summary statistics rather than as motifs for establishing brain connectivity.

We propose a novel Bayesian hierarchical model for brain connectivity analysis. The base level of this framework assumes that the connectivities between voxel pairs between two brain regions follow a mixture distribution with two modes reflecting connected and non-connected voxel pairs. The assumption follows naturally from the "small-worldness" property of the brain networks, since two distant ROIs could be well connected through only a small proportion of efficiently connected neural units. The connectivity information from all cross-region voxel pairs is utilized. Then, the mixture model parameters capturing the proportion are passed to the next level as connectivity strength. Additionally, we adjust for a individual's covariate effects such as age, gender, and psychiatric conditions in the GLM.

The parameters across different levels of this model can be estimated using Markov Chain Monte Carlo (MCMC) methods, and inferences are drawn based on the joint posterior probability distribution for all of the model parameters. Although the focus of our data application in this paper targets functional connectivity, the proposed method is also well suited for structural connectivity analysis by using DTI data along with an associated metric for structural connectivity strength, for instance probabilistic diffusion tensor tractography (Behrens *et al.*, 2007; Hagmann *et al.*, 2003; Parker *et al.*, 2003).

The rest of the paper is organized as follows. In section 2, we introduce the motivating data set. In Section 3, we present the new Bayesian hierarchical model for brain connectivity and parameter estimation strategies by using MCMC. In Section

Figure 3.1: Sagittal View of Brain with ROIs of Our Main Research Interest

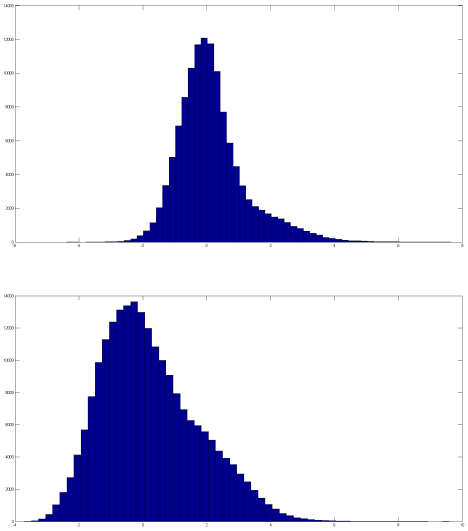


4, we discuss the results from our analysis and the simulation study. Section 5 concludes the paper with a summary and a discussion of the major strengths and future improvement of our proposed method.

3.2 Motivating Example

There is a major focus on resting-state functional connectivity in the field of neuroimaging, which examines connectivity properties during a so-called default mode of neural activity when subjects are left to think for themselves. We consider resting-state fMRI data from 32 treatment naive subjects currently diagnosed with major depression disease, ages 24 to 57 years. The sample includes 12 males and 20 females who have no significant psychological comorbidities or neurological disorders. The fMRI scans were acquired at baseline and clinical covariates includes age, gender, and baseline Hamilton Depression Rating Scale 17 items (HAM-D17). HAM-D17 ranges from 0 to 54 and is used to indicate the depression severity. We consider 0 to 7 as normal, 8 to 17 as mild/moderate depression, and 18 or above as moderate/severe (Maruish, 1999; Schutte and John, 1995). Therefore, the patients could be categorized to two groups: 11 mildly depressed subjects and 21 severely depressed subjects. The data were collected on a 3T Siemens scanner with a Z-saga sequence to

Figure 3.2: Histograms of voxel pair functional connections for two example subjects. The asymmetry and highly connected component at the right side are demonstrated.



avoid orbitofrontal signal ablation. At each scanning session, 150 fMRI volumes were scanned in 7.5 minutes during visual fixation with 30 slices, field of view covering = 220 mm, voxel resolution of 3.4375mm x 3.4375mm x 4mm, TR= 2.92ms, TE = 30ms, and FA=90°. The data preprocessing steps include: motion correction, slice timing correction, normalization and spatial smoothing using 6mm Gaussian FWHM by using AFNI and FSL. Also, the de-trending and demeaning steps were applied to remove the incoherent background shift and texture variation. We want to assess the functional connectivities between ROI, then make inference at the group level that how would depression level affect the connectivities. Particularly, we are interested in connections between prefrontal cortex (PFC) including medial frontal cortex (mF10),orbital frontal cortex (OF11), and lateral prefrontal cortex (latF9) and subgenual cingulate cortex (Cg25) and anterior cingulate (Cg24) which may be involved in emotion experience and processing. (Seminowicz *et al.*, 2004; Ressler and Mayberg, 2007). The regions of interest are demonstrated in Figure 3.1.

When we explore the distribution of voxel-level functional connections (using cor-

relation metric for example) between two ROIs, we notice that it follows a mixture distribution with two components: relatively weakly connected voxel pairs and more closely connected voxel pairs (see Figure 3.2). Moreover, each subject has different proportion of connected component. According to the "small-worldness" property, the size of connected component can represent the connectivity breadth. Therefore, we would develop a method to study such connectivity breadth and identify what factors may influence the connectivity breadth.

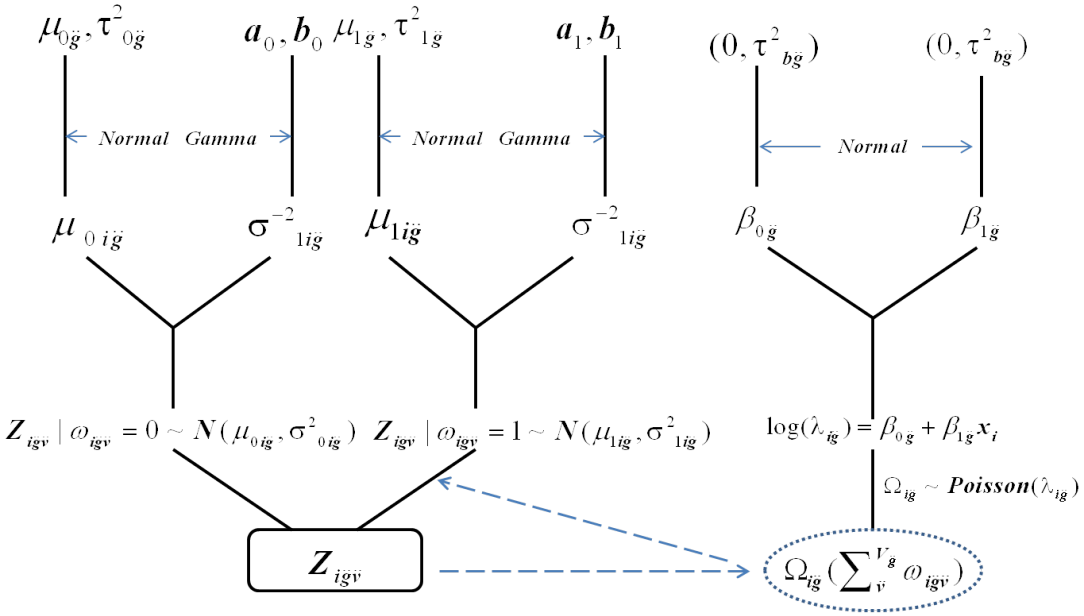
3.3 Methods

We formulate a model that builds on the calculated voxel-level functional connections within ROIs and between ROIs. The ROIs are defined by existing anatomical parcellation of the brain, for example AAL or Brodmann brain atlas (**?**; Garey, 1994). ROIs are labeled as $g = 1, \ldots, G$, where we may set G as high as 116 for AAL map and 48 for Brodmann map. Therefore, we denote $z_{i\ddot{g}\ddot{v}}$ connectivity (suitably transformed, e.g. using Fisher transformation) between voxels v and v' labeled by $\ddot{v}$, between regions g ($v \in g$)and g' ($v' \in g'$) labeled by $\ddot{g}$, for subject i ($i = 1, \ldots, n$). For a region pair $\ddot{g}$, the total number of voxel pairs is $V_{\ddot{g}}$. If region g and g' include V_g and $V_{g'}$ voxels respectively, then we have $V_{\ddot{g}} = V_g \times V_{g'}$ voxel pairs and we assume all subjects share the same region masks.

3.3.1 Bayesian Hierarchical Model

The Bayesian hierarchical framework is developed to jointly model the voxel pair connectivities, ROI pair connectivities, and clinical covariate effects while bookkeeping the variability at each level (see Figure 3.3).

Figure 3.3: chart illustration of the Bayesian hierarchical model. From the bottom to top are voxel pair data (observed and augmented), region level parameters, and hyperpriors.



Distribution for connectivity

In the proposed Bayesian hierarchical model, the first level specifies the links between ROIs per subject by using distribution of voxel pair connections within the region pair. The heuristic behind this is that according to the "small-world" property of brain networks, two regions can be efficiently connected through only a small proportion of connected neural units, for here voxel pair connections. Therefore, at the first level, we assume that $z_{i\ddot{g}\ddot{v}}$ follows a mixture with two modes representing two populations: connected or non-connected voxel pairs. The two components vary around means $\mu_{0i\ddot{g}}$ (non-connected population) and $\mu_{1i\ddot{g}}$ (connected population), with the variances $\sigma^2_{0i\ddot{g}}$ and $\sigma^2_{1i\ddot{g}}$, and we set $\mu_{1i\ddot{g}} > \mu_{0i\ddot{g}}$ to aid identifiability.

$$
\begin{aligned}
&z_{i\ddot{g}\ddot{v}}|\omega_{i\ddot{g}\ddot{v}}, \mu_{0i\ddot{g}}, \sigma^2_{0i\ddot{g}}\mu_{1i\ddot{g}}, \sigma^2_{1i\ddot{g}} \sim \omega_{i\ddot{g}\ddot{v}}\mathrm{N}(\mu_{1i\ddot{g}}, \sigma^2_{1i\ddot{g}}) + (1-\omega_{i\ddot{g}\ddot{v}})\mathrm{N}(\mu_{0i\ddot{g}}, \sigma^2_{0i\ddot{g}}),\\
&\omega_{i\ddot{g}\ddot{v}}|\pi_{i\ddot{g}} \sim \text{Bernoulli}(\pi_{i\ddot{g}}),\\
&\Omega_{i\ddot{g}}|\lambda_{i\ddot{g}} \sim \text{Poisson}(\lambda_{i\ddot{g}}), \text{ where } \Omega_{i\ddot{g}} = \sum_{\ddot{v}}^{V_{\ddot{g}}} \omega_{i\ddot{g}\ddot{v}} \text{ and } \lambda_{i\ddot{g}} = V_{\ddot{g}}\pi_{i\ddot{g}}
\end{aligned}
\tag{3.3.1}
$$

The latent indicator variable $\omega_{i\ddot{g}\ddot{v}}$ assigns a voxel pair connectivity to either connected or non-connected group. In Bayesian data analysis, the latent variables are usually treated as missing data and data augmentation techniques can be used for purpose of estimation. Our interest lies in detecting the between ROIs connectivity breadth represented by the number of connected pairs $\lambda_{i\ddot{g}} \doteq V_{\ddot{g}}\pi_{i\ddot{g}}$ which is the sum of latent indicator variables. Since $V_{\ddot{g}}$ is very large number at the order of hundreds of thousands, $\Omega_{i\ddot{g}} = \sum_{\ddot{v}}^{V_{\ddot{g}}} \omega_{i\ddot{g}\ddot{v}}$ can be considered to follow a Poisson distribution with parameter $\lambda_{i\ddot{g}}$.

Prior specifications

As shown in formula 2, at the second level of the hierarchical model expresses a prior belief that each subject's mixture component means $\mu_{0i\ddot{g}}$ and $\mu_{1i\ddot{g}}$ arise from normal distributions with across subject means $\mu_{0\ddot{g}}$ and $\mu_{1\ddot{g}}$ and variances $\tau^2_{0\ddot{g}}$ and $\tau^2_{1\ddot{g}}$. The functional connections of voxel pairs within a ROI across all subjects shed light on the prior assumption for $\mu_{1\ddot{g}}$, because numerous amount of voxel pairs are connected within a ROI. On the other hand, the connectivities of voxel pairs between ROIs across all subjects yield information about $\mu_{0\ddot{g}}$. Thus, we empirically specify $\mu_{1\ddot{g}}$ by calculating the mode of voxel pair functional connections within a ROI across all subjects, because it seems to represent a reasonable starting point to assume that voxel pairs within anatomically-defined regions exhibit high association. Note that

we exclude the voxel pairs within a ROI of geometric distance greater than 30 mm to avoid the incoherent voxel pairs. Similarly, $\mu_{0\ddot{g}}$ could be specified by using the mode of voxel pair functional connections between ROIs across all subjects The variances $\tau^2_{0\ddot{g}}$ and $\tau^2_{1\ddot{g}}$ reflect the confidence of the observed information from data. Usually, we set $\tau^2_{1\ddot{g}}$ as a relatively small value (say $\tau_{1\ddot{g}} = 0.01$) to ensure all subjects use similar criteria to determine that a voxel pair is connected or non-connected. In addition, for the first level variance parameters $\sigma^2_{0i\ddot{g}}$ and $\sigma^2_{1i\ddot{g}}$ we set Gamma distributions with hyperprior parameters $a_0 = 0.1, b_0 = 0.005, a_1 = 0.1, b_1 = 0.005$ as vague priors.

$$
\begin{aligned}
&\mu_{0i\ddot{g}}|\mu_{0\ddot{g}}, \tau^2_{0\ddot{g}} \sim N(\mu_{0\ddot{g}}, \tau^2_{0\ddot{g}}),\\
&\mu_{1i\ddot{g}}|\mu_{1\ddot{g}}, \tau^2_{1\ddot{g}} \sim N(\mu_{1\ddot{g}}, \tau^2_{1\ddot{g}}),\\
&\sigma^{-2}_{0i\ddot{g}} \sim \text{Gamma}(a_0, b_0),\\
&\sigma^{-2}_{1i\ddot{g}} \sim \text{Gamma}(a_1, b_1),\\
&log(\lambda_{i\ddot{g}}) = \beta_{\ddot{g}0} + \beta_{\ddot{g}} x_i, \text{ where } \lambda_{i\ddot{g}} = E(\Omega_{i\ddot{g}}|\beta_{\ddot{g}0}, \beta_{\ddot{g}}),
\end{aligned}
\tag{3.3.2}
$$

More importantly at the second level, we model the region pair connectivity breadth $\Omega_{i\ddot{g}}$ through a generalized linear model (GLM) with *log* link to accommodate the Poisson distribution. The main effects $\boldsymbol{\beta}_{\ddot{g}}$ of the GLM include the clinical covariates such as disease status, age, and treatment group. The priors for $\boldsymbol{\beta}_{\ddot{g}}$ are set vague enough to let the large amount of observed data determine posterior distribution by using normal distributions with zero mean and large variance, for example 20.

$$
\beta_{\ddot{g}0} \sim \text{N}(0, \tau^2_{0b}), \beta_{\ddot{g}} \sim \text{N}(0, \tau^2_b),
\tag{3.3.3}
$$

A link to local *fdr*

We esteem that there is a natural link between the mixture model in the proposed method and the local *fdr* method by Efron, 2005. Both models target identification the true positive elements from a mixture model consisting of a dominating null distribution and a small hump rather than using hard cut-off values. But, the local *fdr* method tends to focus on $fdr(z)$ at specific z values, while we are interested in the marginal effect of $\Omega_{i\ddot{g}}$:

$$\begin{aligned} E(\Omega_{i\ddot{g}}) &= \int \sum_{\ddot{v}}^{V_{\ddot{g}}} P(z_{i\ddot{g}\ddot{v}}) f(z_{i\ddot{g}\ddot{v}}) \mathrm{d}z_{i\ddot{g}\ddot{v}} \\ &= \sum_{\ddot{v}}^{V_{\ddot{g}}} \int \left(\frac{\pi_{i\ddot{g}} f_1(z_{i\ddot{g}\ddot{v}})}{\pi_{i\ddot{g}} f_1(z_{i\ddot{g}\ddot{v}}) + (1-\pi_{i\ddot{g}}) f_0(z_{i\ddot{g}\ddot{v}})} \right) f(z_{i\ddot{g}\ddot{v}}) \mathrm{d}z_{i\ddot{g}\ddot{v}} \\ &= V_{\ddot{g}} \pi_{i\ddot{g}} \end{aligned} \tag{3.3.4}$$

where $f(z_{i\ddot{g}\ddot{v}}) = \pi_{i\ddot{g}} f_1(z_{i\ddot{g}\ddot{v}}) + (1-\pi_{i\ddot{g}}) f_0(z_{i\ddot{g}\ddot{v}})$ and $P(z_{i\ddot{g}\ddot{v}})$, the probability of "true positive" is identical to $1 - fdr$. Clearly, the marginal variable is subject to less variance, and $\Omega_{i\ddot{g}}$ is invariant to the choice of f_0 and f_1 as long as they can capture the mixture distribution curves. Usually, the local *fdr* methods apply parametric or nonparametric regression to estimate curves of the two components in order to obtain accurate estimate of $\widehat{fdr}$ at each z. However, in our case, we are interested in the marginal estimate and include multiple subjects with each subject having its own mixture distribution and parameters; and these parameters follow hyperprior distributions with hyperparameters in the hierarchical structure. Therefore, we feel that the normal mixture model is effective to yield robust estimates of the marginal variables while it does introduce too much complexity to the multilevel model.

3.3.2 Estimation and posterior inference

MCMC algorithms are employed to estimate the massive numbers of parameters in our hierarchical probability model. Since the GLM is introduced, no conjugate prior could be specified for the main effects β. Therefore, we use a Gibbs sampler with Metropolis updates where needed for model fitting. The full conditionals are given by the following:

$$\omega_{i\ddot{g}\ddot{v}} \sim \text{Bernoulli}\left(\frac{\pi_{i\ddot{g}}\text{N}(z_{i\ddot{g}\ddot{v}}|\mu_{1i\ddot{g}},\sigma^2_{1i\ddot{g}})}{\pi_{i\ddot{g}}\text{N}(z_{i\ddot{g}\ddot{v}}|\mu_{1i\ddot{g}},\sigma^2_{1i\ddot{g}})+(1-\pi_{i\ddot{g}})\text{N}(z_{i\ddot{g}\ddot{v}}|\mu_{0i\ddot{g}},\sigma^2_{0i\ddot{g}})}\right)$$

$$\mu_{0i\ddot{g}} \sim \text{Normal}\Big(\Psi_{0i\ddot{g}}\{\tau^{-2}_{0i\ddot{g}}\mu_{0\ddot{g}}+(V_{\ddot{g}}-\Omega_{i\ddot{g}})\sigma^{-2}_{0i\ddot{g}}\overline{z}_{0i\ddot{g}}\},\Psi_{0i\ddot{g}}\Big)$$

$$\mu_{1i\ddot{g}} \sim \text{Normal}\Big(\Psi_{1i\ddot{g}}\{\tau^{-2}_{1i\ddot{g}}\mu_{1\ddot{g}}+\Omega_{i\ddot{g}}\sigma^{-2}_{1i\ddot{g}}\overline{z}_{1i\ddot{g}}\},\Psi_{1i\ddot{g}}\Big)$$

$$\sigma^{-2}_{0i\ddot{g}} \sim \text{Gamma}\left\{a_0+(V_{\ddot{g}}-\Omega_{i\ddot{g}})/2,\Big(\frac{1}{b_0}+\frac{1}{2}\sum_{\forall\omega_{i\ddot{g}\ddot{v}}=0}(z_{i\ddot{g}\ddot{v}}-\mu_{0i\ddot{g}})^2\Big)^{-1}\right\} \tag{3.3.5}$$

$$\sigma^{-2}_{1i\ddot{g}} \sim \text{Gamma}\left\{a_1+(\Omega_{i\ddot{g}})/2,\Big(\frac{1}{b_1}+\frac{1}{2}\sum_{\forall\omega_{i\ddot{g}\ddot{v}}=1}(z_{i\ddot{g}\ddot{v}}-\mu_{1i\ddot{g}})^2\Big)^{-1}\right\}$$

$$\beta_{\ddot{g}0} \propto \prod_i^n e^{(\beta_{\ddot{g}0}+\beta_{\ddot{g}}x_i)\Omega_{i\ddot{g}}}exp(-e^{(\beta_{\ddot{g}0}+\beta_{\ddot{g}}x_i)})\text{N}(\beta_{\ddot{g}0}|0,\tau^2_{0b})$$

$$\beta_{\ddot{g}} \propto \prod_i^n e^{(\beta_{\ddot{g}0}+\beta_{\ddot{g}}x_i)\Omega_{i\ddot{g}}}exp(-e^{(\beta_{\ddot{g}0}+\beta_{\ddot{g}}x_i)})\text{N}(\beta_{\ddot{g}}|0,\tau^2_{b})$$

where $\Psi_{0i\ddot{g}}=\Big((V_{\ddot{g}}-\Omega_{i\ddot{g}})\sigma^{-2}_{0i\ddot{g}}+\tau^{-2}_{0i\ddot{g}}\Big)$, $\Psi_{1i\ddot{g}}=\Big(\Omega_{i\ddot{g}}\sigma^{-2}_{1i\ddot{g}}+\tau^{-2}_{1i\ddot{g}}\Big)$, $\overline{z}_{1i\ddot{g}}$ is the mean for all $\ddot{v}$ of $\omega_{i\ddot{g}\ddot{v}}=1$, $\overline{z}_{0i\ddot{g}}$ is the mean for all $\ddot{v}$ of $\omega_{i\ddot{g}\ddot{v}}=0$, and $\pi_{i\ddot{g}}=e^{\beta_{\ddot{g}0}+\beta_{\ddot{g}}x_i}/V_{\ddot{g}}$.

In the first stage of MCMC, we first sample the latent indicator variable, then sample the four parameters of the Gaussian mixture model. For one region pair, the latent indicators are summarized by the number of connected voxel pairs $\Omega_{i\ddot{g}}$ out of total number $V_{\ddot{g}}$ of pairs. At the second stage, the GLM parameters are

updated through Metropolis steps with input of $\Omega_{i\ddot{g}}$ following a Poisson rather than a Binomial distribution (log link rather than logistic link). The reason is that $\Omega_{i\ddot{g}}$ and $V_{\ddot{g}}$ are usually on the order of thousands and millions respectively and therefore the ratio function of Metropolis step may be easily overflowed and generate zeros in the denominator if the logistic link is used but not for the log link. Also, from the perspective of neurophysiology, the connection breadth between two regions would be represented by number of connected voxels because the ratio could heavily depend on the sizes of the regions. The ratios of βs in the Metropolis step are:

$$r_\beta = \exp\Bigg\{\sum_i^n [\Omega_{i\ddot{g}}(\beta'_{i\ddot{g}} - \beta_{i\ddot{g}}) + exp(-\eta'_{i\ddot{g}}) - exp(-\eta_{i\ddot{g}})] + log(p_0(\beta'_{i\ddot{g}}|0,\tau_b^2)) - log(p_0(\beta_{i\ddot{g}}|0,\tau_b^2))\Bigg\}$$

where β' are proposed values, and accordingly plug in $\eta_{i\ddot{g}} = \beta_{\ddot{g}0} + \beta_{\ddot{g}} x_i$ to obtain $\eta_{i\ddot{g}}$. Let p_0 indicate the prior distribution. The proposed value will be accepted with probability of $min(1, r_\beta)$. All the other parameters can be sampled from known distributions through Gibbs sampler as shown in formula 5.

The proposed Bayesian model formulation and implementation would take samples from the joint posterior distribution for all of the model parameters for statistical inferences (e.g. Bayes factor and credible intervals) regarding connectivity and brain network properties at different levels. First, the posteriors include the voxel pair connectivity (the indicator function) information for each subject, thus we can make inference about 1: The probability of a voxel pair being connected denoted by $PP_{i\ddot{g}}(\ddot{v})$ by calculating the rate of $\omega_{i\ddot{g}}(\ddot{v}) = 1$; 2: Identifying which voxels are hub voxels that connect with numerous other voxels; and 3: which voxels are highly connected to each other(modularity). Second, at the region pair level, we can draw inference about the connectivity breadth of region pair by calculating the posterior of $e^{\beta_{\ddot{g}0}+\beta_{\ddot{g}}x_i}$; furthermore at the whole brain level, the graph theoretical property metrics such as

"small-worldness", assortativity, and centrality could be measured with their posterior distributions based on the voxel pair level inferences. Last and most important, at the population level, we can investigate impacts of the clinical covariate effects β such as age, disease groups, and treatment responses on connectivity breadth between ROIs.

3.4 Results

We apply our Bayesian hierarchical model to both the resting-fMRI for depression study and a simulation study.

3.4.1 MDD Study Using Resting-state fMRI Data

In the MDD study, our goal is to explore the neural circuits associated with regions of prefrontal cortex (e.g. latF9, mF10, OF11) and regions that may be involved in emotion experience and processing including Cg25 and Cg24. Among all the region pairs, the functional connections between Cg25 and mF10, as well as Cg25 and OF11 are our most interest according to previous studies and neurophysiological knowledge (Ressler and Mayberg, 2007). For resting-state fMRI data, the voxel pair functional connections are determined by the temporal coherence between time series of distant voxels. Hence, it is natural to use correlation or spectral coherence as metrics, and here we use correlation for the purpose of demonstration. In addition, the Fisher transformation and linear transformation are used towards the correlations as voxel pair connectivity metrics. As demonstrated in Figure 3.2, the functional connections follow a mixture distribution. The average of the connected voxel pairs is centered around correlation of 1.65 (which maps to Pearson correlation of 0.4), which is calculated by the mode of the distribution of within region voxel pair functional connections. Then, we apply our Bayesian hierarchical model for connectivity analysis

based on the transformed correlation metrics. For the purpose of demonstration, we use the region pair mF10 and Cg25 for model comparison and results illustration.

Table 3.1: Patient Group Comparisons for Breadth of Connectivity between Specified Regions (Entries represent number of connected voxel pairs.)

Parameters	Median (Percent)	std	25 pct	75 pct	Mode	Difference(CI)	Percent
Cg25 & mF10 (122,591 voxel pairs)							
(Mildly depressed)	6967 (5.7)	17	6961	6988	6974	-880	-13 %
(Severely depressed)	6087 (4.9)	15	6069	6100	6088	(-892,-846)	
Cg25 & latF9 (119,426 voxel pairs)							
(Mildly depressed)	954 (0.8)	14	928	993	961	-240	-25 %
(Severely depressed)	714(0.6)	11	686	740	715	(-226,-250)	
Cg25 & OF11 (214,376 voxel pairs)							
(Mildly depressed)	29202 (13.6)	409	28624	29349	29173	-8593	-30 %
(Severely depressed)	20609 (9.6)	487	20486	21171	20578	(-8611,-8568)	
Cg25 & Cg24 (32,705 voxel pairs)							
(Mildly depressed)	4064 (12.4)	49	4048	4105	4072	135	3 %
(Severely depressed)	4201 (12.8)	46	4155	4222	4201	(110,147)	

We first draw inferences on the voxel pair level functional connections. Within each region pair, we summarize the probability of each cross-region voxel pair is connected, then identify which voxels are connected to an amount of voxels in the other region as a hub at the population level. In Figure 3.4, each row in the heatmap represents a voxel and each column represents a subject, and the intensity indicates its number of connected voxels. Based on the results, using region mF10 as a example, we can notice there is a clear pattern that there are an amount of voxels are highly connected to region Cg25 across all 32 subjects and these highly connected voxels serves has neural processing hubs since they reveal coherent activity with a lot of other voxels. This will enhance our understanding of the neurophysiology of the human brain. In addition, we can also obtain the information of which voxel pairs are connected with high probability (Figure 3.5), and we can infer which voxel pairs are most commonly connected functionally for one region pair. Based on the 3D figure, we notice that the voxels highly connected to the other region are evenly distributed.

Then, we investigate the results at the voxel pair levels, as well as the impact

Figure 3.4: Heatmap showing the number of functional connections for each voxel in mF10 across 32 Subjects, with the connections extending to voxels in Cg25. The voxels are ordered such that the voxel with most connections are listed first. Note that some voxels in mF10 consistently show a large number of connections to voxels in Cg25 across the 32 subjects. These highly connected voxels serves has neural processing hubs since they reveal coherent activity with a lot of other voxels

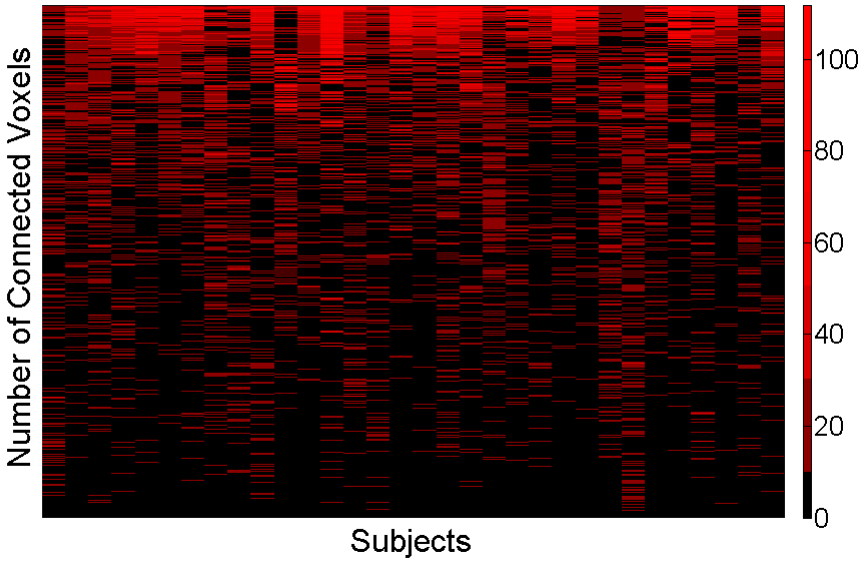


Figure 3.5: 3D plot showing voxel pairs connected with high probabilities(cut-off 0.9 for across subject average) falling in regions Mf10 and Cg25.

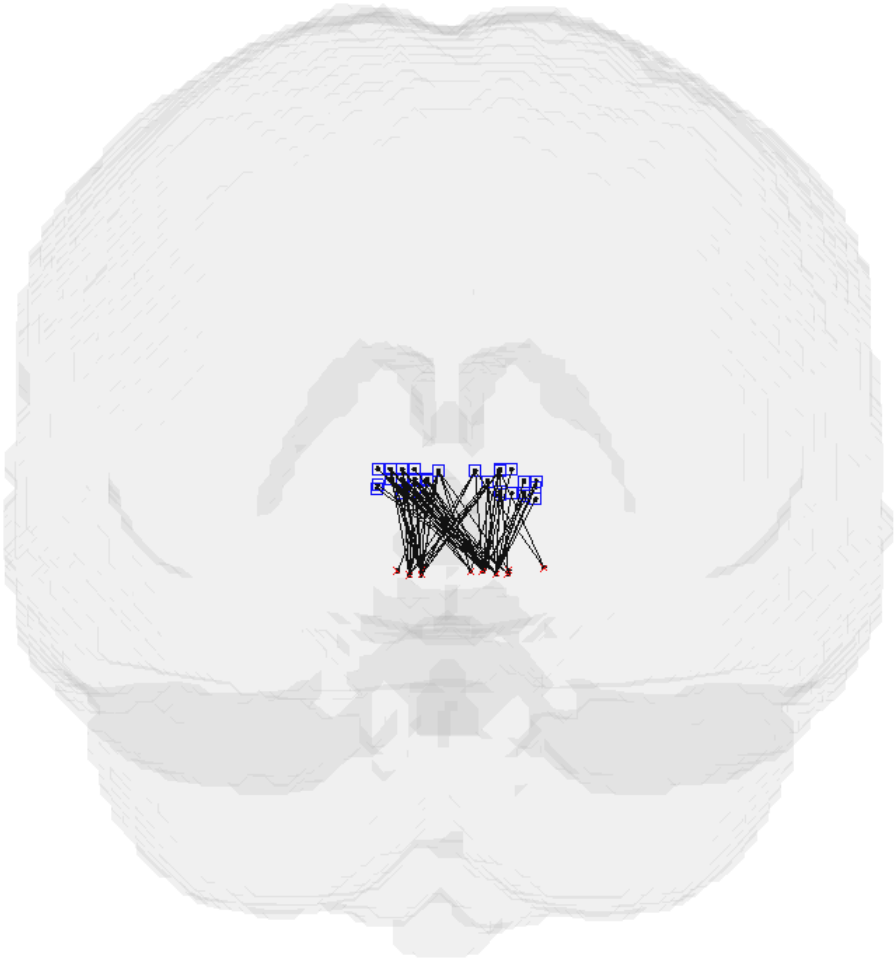


Figure 3.6: Trace plots for main effects of connectivity between Cg25 and mF10. The plots indicate the chains converge after 1000 iterations with small variations.

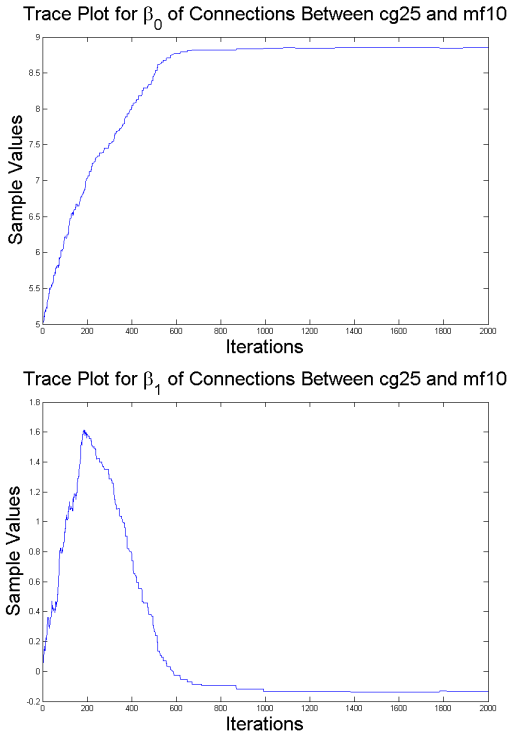


Figure 3.7: Histograms for connectivities between Cg25 and mF10 by using region representatives for all subjects in two groups. The region representatives connectivities indicate that the two region in two groups are equally connected, but the mildly depressed group is lower than the severely depressed group.

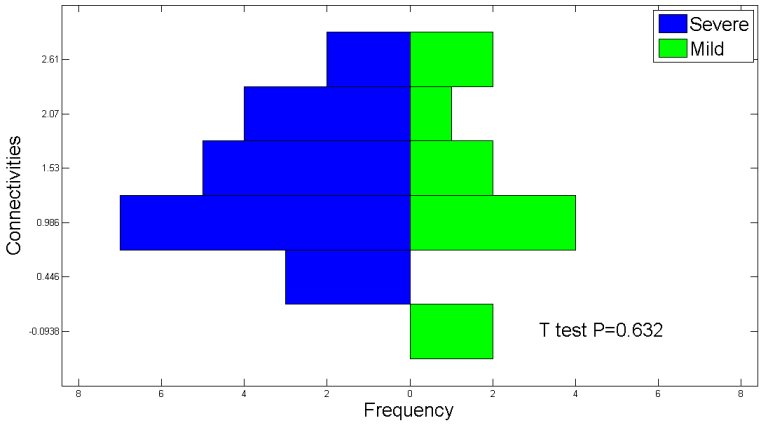
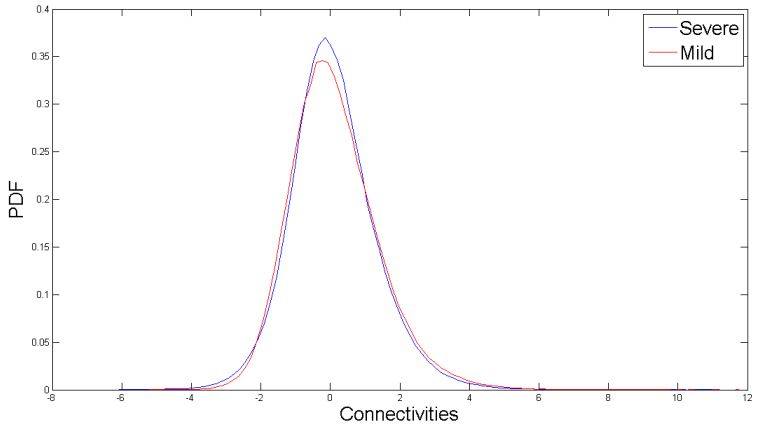


Figure 3.8: Density plots for all voxel pairs of Cg25 and mF10 for all subjects in two groups. The mildly depressed group is slightly more connected than the severely depressed group.



of the clinical disease status on connections between regions. Although functional connections of all voxel pairs in all ROIs could be used for analysis and the computational load is manageable, we only demonstrate the results from our most interested ROI pairs and those reflect significant differences between different depression levels, since most region pairs show similar connectivity extent. We summarize the breadth of connected voxel pairs by different HAM-D score severity groups. In Table 3.1, we can notice the region pair of mF10 and Cg25, OF11 and Cg25, Cg24 an Cg25 are relatively highly connected. The connectivity breadth between mF10 and Cg25, OF11 and Cg25 differ between the two patient groups: the patients with more severe depression have 15% and 30% less functional connections , which may reveal the fact that the neural units for emotion processing interact less with the prefrontal cortex neural units. But, there is almost no difference in functional connections between groups for Cg24 and Cg25 (3% difference), which indicates that the two group have the similar functional connections between the regions for emotion processing. The difference in the table indicates the difference of numbers of connected voxel pairs in two groups with reference of mildly depressed group. Also, we investigate the age

and gender effect and find the parameters are very close to zero.

As a comparison, we further apply the region representative method by using correlations of the temporal profiles of the first vectors of SVD decomposition of all temporal profiles in mF10 and Cg25 for all subjects. Figure 3.7 depicts the results: based on the functional connections between mF10 and Cg25 representatives there is no difference between the two groups (with t test p value at 0.63), and the mean number of functional connections in mildly depressed group is lower than in the severely depressed group which contradicts the findings by our model. Furthermore, we check the density plot (Figure 3.8) for voxel pairs functional connections in two groups and found the fact that the mildly depressed group has higher numbers of functional connections than the severely depressed group. Therefore, we conclude that our model is able to reveal the subtle difference caused only by the connected voxel pairs, and accordingly detect the group effects.

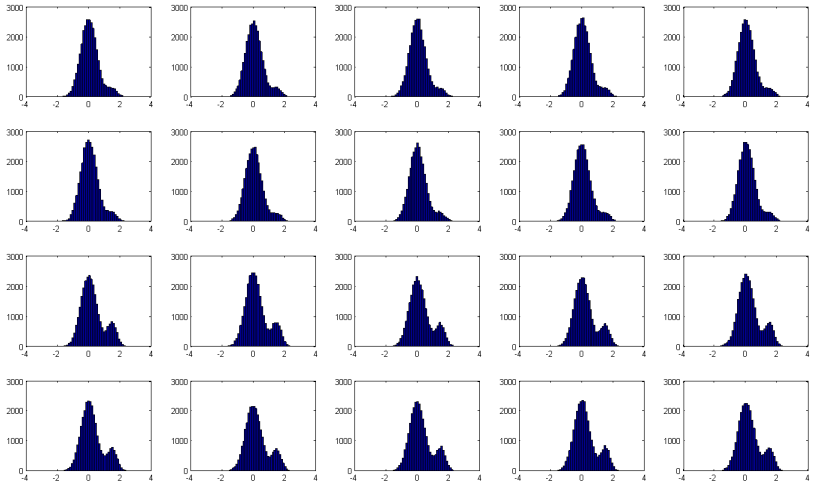
In addition, we investigated the voxel pairs are highly connected

The MCMC runs were checked to ensure the stability of the results and the convergence of the chains. However, the massive number of voxel pairs of the data in question prevents formal use of MCMC convergence tools for voxel level parameters. We monitor the trace plots at the region level. The resulting posterior samples converge after around 1000 burn-in period (Figure 3.6 for example).

3.4.2 Simulation Study

In the simulation, we assume that there are 20 subjects with 10 controls and 10 cases, and within each subject there are 30,000 ($V_{\vec{g}}$) voxel pairs for each region pair. We first generate the random numbers of connected cross-region voxel pairs by using Poisson

Figure 3.9: Histograms of voxel pair connectivities for simulated 20 simulated subjects. The big bump in each histogram represents unconnected voxel pairs, and the small bump represents the connected voxel pairs. The size of the small bumps decide the breadth of the connectivities.



distributions with mean parameters:

$$\Omega_{i\ddot{g}} = exp(\beta_{\ddot{g}0} + \beta_{\ddot{g}} x_i).$$

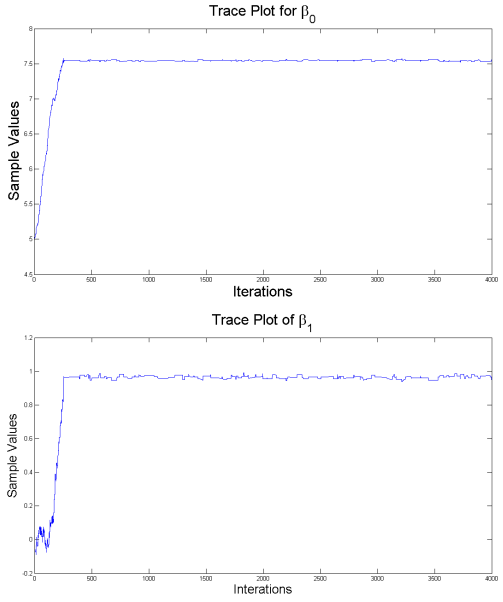
Here, we use $\beta_{\ddot{g}0} = 7.5$ and $\beta_{\ddot{g}1} = 1$ to indicate the controls have more breadth in connectivity than cases do. Then, we simulate $\Omega_{i\ddot{g}}$ voxel pairs following normal distribution of mean 1.5 and variance 0.5, and $V_{\ddot{g}} - \Omega_{i\ddot{g}}$ voxel pairs following normal distribution of mean 0 and variance 0.5. The simulated data are demonstrated in the histograms in Figure 3.9, where the controls have a larger component of connected voxel pairs. Our main goal is to detect main group effect as well to check the number of voxel pair are sampled as connected for each subject. We apply our proposed hierarchical Bayesian framework to the simulated data set.

We check the trace plots of Figure 3.10 to ensure the convergence of the chains of all parameters. Based on the results, we notice the chains converge after around

Table 3.2: Summaries of Distributions of Posteriors in the Simulation Study

Parameters	Median	std	25 pct	75 pct	Mode
$\overline{\Omega}_0$	1891.5	9.44	1884.8	1898	1892
$\overline{\Omega}_1$	4965.1	11.28	4957.9	4972.7	4966
β_0 (7.5)	7.54	0.0083	7.54	7.55	7.54
β_1 (1)	0.97	0.0088	0.96	0.97	0.97

Figure 3.10: Trace plots for main effects of the simulation data set. Both parameters converge to the true after burn-in iterations.



300 iterations to vary around the true values with relatively small variances. We summarize some of the results in Table 3.2 and we feel the estimates are accurate and the model can reveal the group effect effectively and accurately. Therefore, this simulation supports the validity of the proposed method and parameter estimation algorithm. The small bias of βs may come from the false negative classification of the mixture model because of gravity of the huge null component.

All the programming including MCMC and imaging processing was implemented in MATLAB. With sample size of one group of 20 subjects, the MCMC will take around 6 minutes to run for one region pair on a Intel i7 3.4 GHz CPU and 8G

memory PC.

3.5 Discussion

Our proposed method is the first approach that is able to use all available voxel pair connectivity information to gain region pair connectivity information, and the first model to incorporate the brain network graph theoretical properties to determine the brain connectivity. Based on the posterior distributions of the parameters at all levels, we can make inferences on numerous quantities about brain connectivities. For most clinical studies, we are interested in how the main effects affect the connectivities. By our model setup, the main effects detect the change of the "small proportion" of connected neural units and therefore the subtle difference. It is also robust to the bias due to non-perfect image registration, because inferences about the main interests do not depend on the specific location s of voxel pairs and marginal region pair connectivity can be effectively and efficiently estimated. In addition, each voxel pair has a posterior probability of being connected. The inter-region "hub" voxel could be detected if it is connected to many voxels with high probabilities.

The mixture model in the first level determines the number of connected voxel pairs in a fashion of *local* fdr rather than choosing hard cutoff points, which improves the accuracy of sampled numbers of connected voxel pairs. In the model, we assume the independence of all voxel pairs between a region pair, and this assumption does not affect the marginal quantity of total number of voxel pairs and hence the group parameters. Furthermore, based on the 3D plot showing the voxel pair connections (Figure 3.5), it does not imply any explicit correlation of spatially adjacent voxels. In the future, we will further investigate the impact of the assumption and alternative strategies for modeling. Although numerous voxel pairs per subject and

many subjects may be included, the computational load is manageable with around 10 minutes per region pair by average on a regular PC and better performance on clusters. In the future, we will run multiple chains for the purpose of model checking by parallel computing on a cluster. The proposed model is also applicable to other functional connectivity metrics and other brain connectivity measures such as structural connectivity from DTI data.

Chapter 4

Topic 3: An Unified Bayesian Framework for Resting-state fMRI Data Analysis: Jointly Modeling Frequency Activity and Connectivity

4.1 Introduction

Functional magnetic resonance imaging (fMRI) techniques provides a pathway to investigate *in vivo* neural activity in a detailed 3D space. fMRI has been widely applied in the fields of human cognition, emotion, and behavior as well as mental illness studies. During an fMRI experiment, the brain signal is recorded as the neural responses based on external stimuli (task-induced fMRI) or brain activity at resting status (resting-state fMRI). The fMRI data analysis generally include brain activation

analysis and functional connectivity analysis, depending on the research objectives. The typical brain activation analysis is based on task induced fMRI data, and the goal is to detect the intensity of the neural responses towards the experimental tasks and the differences of those responses in patterns of brain activity between various experimental conditions, between different subgroups of subjects, e.g. normal subjects and patients with major depression. For a task-induced fMRI study, a two stage data analysis strategy is often employed (Worsley et al., 2002). The first stage involves the general linear model towards single voxel time series with task event based design matrix, but for resting-state fMRI data such analysis is not feasible because no event is included. Therefore, for resting-state fMRI analysis we first measure the brain activity by using fractional amplitude of low frequency fluctuation (fALFF), which is the ratio of power spectrum of low-frequency (0.01 to 0.08 Hz) to that of the entire frequency range and reflects the intensity of regional spontaneous brain activity (Zou et.al, 2008). Also, resting-state fRMI data enable us to conduct functional connectivity analysis to cluster brain areas to different networks based on the similarity of their temporal/frequency domain properties (Bowman 2004, 2005).

The Bayesian hierarchical framework has been shown as an efficient and effective method for detecting task-related changes in brain activity and covariance between distinct brain locations simultaneously, based on the first level analysis statistics (Bowman et al 2008). However, such a model is designed for the task data analysis and the dimension of spatial covariance matrix is restricted by no more than the number of subjects in the smallest study group. In this article, we propose a novel Bayesian hierarchical model particularly for resting-state fMRI data analysis to jointly model the localized fALFF "brain activity" and functional connectivity based on frequency coherence. The functional connectivity is incorporated as a part of the hierarchical framework based on the Bayesian infinite mixture model, also known as

Dirichlet process mixture model (Ferguson, 1973, Neal, 2000, Rasmussen, 2000). It does not require pre-specifying the number of the mixture components. The clustering procedure is implemented by using the Chinese restaurant process by treating the number of networks and network distribution as latent components (Aldous, 1985; Pitman, 1996). This entire multilevel framework is estimated using Markov Chain Monte Carlo (MCMC) techniques via Gibbs sampling. The Bayesian framework provides inference based on the posterior distribution of the model parameters with regard to the voxel level and region level frequency band "brain activity", and functional connectivity network maps based on the frequency coherence. We monitor the likelihood as the behavior of the CRP process.

The rest of the paper is organized as follows. In Section 2, we present the Bayesian framework and provide an accompanying computational strategy. In Section 3, we examine the classification performance of the proposed method for a data example. Section 4 concludes the paper with a summary and a discussion of the major strengths of our model.

4.2 Motivating Example

We consider resting-state fMRI data from 36 subjects currently diagnosed with major depression disease. The sample includes 16 males and 20 females who have no significant psychological comorbidities or neurological disorders. The fMRI scans were acquired at baseline and clinical covariates includes age, gender, and baseline Hamilton Depression Rating Scale 17 items (HAM-D17). HAM-D17 ranges from 0 to 54 and is used to indicate the depression severity. We consider 0 to 7 as normal, 8 to 17 as mild/moderate depression, and 18 or above as moderate/severe (Maruish, 1999; Schutte and John, 1995). Therefore, the patients could be categorized to two

groups: 18 mildly depressed subjects and 18 severely depressed subjects. The frequency descriptor is an important feature for resting-state fMRI data, and we intend to differentiate the difference between mildly depressed patients with severely depressed patients as well as to investigate the region level connectivity network based on the frequency coherence.

The data were collected on a 3T Siemens scanner with a Z-saga sequence to avoid orbitofrontal signal ablation. At each scanning session, 150 fMRI volumes were scanned in 7.5 minutes during visual fixation with 30 slices, field of view covering = 220 mm, voxel resolution of 3.4375mm x 3.4375mm x 4mm, TR= 2.92ms, TE = 30ms, and FA=90°. The data preprocessing steps include: motion correction, slice timing correction, normalization and spatial smoothing using 6mm Gaussian FWHM by using AFNI and FSL. Also, the de-trending and demeaning steps were applied to remove the incoherent background shift and texture variation.

4.3 Method

In this section, we describe the unified Bayesian framework for frequency response and brain network analysis in details.

The region of interest (ROI) is often defined based on existing brain anatomical parcellation maps, for example, the AAL map includes up to 116 ROIs and the Brodmann map includes 48 ROIs with region index $g = 1, \ldots, G$(Tzourio- Mazoyer et al., 2002; Garey, 1994). We use V_g to denote the number of voxels in a particular region indexed by g. The individualized fALFF frequency band proportions are denoted by $\boldsymbol{\beta}_{ig} = \{\beta_{ig}(1), \ldots, \beta_{ig}(V_g)\}'$ (localized effects from all voxels in region g are collected into a single vector), and x_{iq} is the qth covariate of subject i. Suppose we have N subjects, and all regions could be allocated to a cluster k, with unknown number of

total clusters (infinite mixture model) with the cluster size n_k and the initial number of cluster K_0.

Then, our hierarchical Bayesian framework has the following structure:

$$
\begin{aligned}
&\boldsymbol{\beta}_{ig}|\boldsymbol{\mu}_g, \alpha_{ig}, \sigma_g^2 \sim \text{MVN}(\boldsymbol{\mu}_g + \mathbf{1}\alpha_{ig} + \sum_{q=1}^{Q} \boldsymbol{\gamma}_{gq} x_{iq}, \sigma_g^2 \mathbf{I}) \\
&\boldsymbol{\mu}_g|\lambda_g^2 \sim \text{MVN}(\mathbf{1}\mu_{0g}, \lambda_g^2 \mathbf{I}) \\
&\boldsymbol{\gamma}_{gq}|\tau_g^2 \sim \text{MVN}(\mathbf{0}, \tau_{gq}^2 \mathbf{I}) \\
&\sigma_g^{-2} \sim \text{Gamma}(a_0, b_0) \\
&\lambda_g^{-2} \sim \text{Gamma}(c_0, d_0) \\
&\tau_{gq}^{-2} \sim \text{Gamma}(e_{0q}, f_{0q}) \\
&\alpha_{ig}|\theta_{\mathbf{ig}} \sim \text{N}(\theta_{\mathbf{ig}}) \text{ and } \theta_{\mathbf{ig}} = \{\phi_{ig}, \epsilon_{ig}^2\} \\
&\theta_{\mathbf{ig}} \sim G, G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_{\mathbf{ik}}}, \text{ and } \theta_{\mathbf{ik}} = \{\phi_{ik}, \epsilon_{ik}^2\} \\
&G \sim DP(h_0, H_0) \\
&\phi_{ik} \sim \text{N}(0, \xi_k^2) \\
&\epsilon_{ik}^2 \sim \text{Gamma}(a_0, b_0) \\
&\xi_k^2 \sim \text{Gamma}(a_0, b_0)
\end{aligned}
\tag{4.3.1}
$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iV_g})'$, $\boldsymbol{\mu}_g = (\mu_g(1), \ldots, \mu_g(V_g))'$, and $\boldsymbol{\gamma}_{gq} = (\gamma_{gq}(1), \ldots, \gamma_{gq}(V_g))'$. μ_{0g} is the global mean across all subjects and intra-regional voxels. $\theta_{\mathbf{ig}}$ is a vector of the mean and variance parameters for the normal distribution for region g, and given the cluster allocation (say k), then they are the the mean and variance parameters for the kth mixture component represented by $\theta_{\mathbf{ik}} = \{\phi_{ik}, \epsilon_{ik}^2\}$.

The top level of the model assumes that the brain activity $\boldsymbol{\beta}_{ig}$ vector in region g

follows a multivariate normal distribution. Each element of the vector, $\boldsymbol{\beta}_{ig}(v)$ varies around the population-level mean parameter $\mu_g(v)$, q clinical effects $\sum_{q=1}^{Q} \boldsymbol{\gamma}_{gq} x_{iq}$ (e.g. age and gender), and an subject-wise region level random effect component α_{ig}.

At the second level of the model, it indicates that each voxel within the ROI arises from a normal distribution with a mean given by the overall region mean μ_{0g} and variance λ_g^2. We obtain the mean prior parameter μ_{0g} by taking the region mean, as it is reasonable to assume that the voxels within the same anatomically defined regions often tend to exhibit more coherent frequency property due to neurophysiology as well as spatial smoothing. In addition, we assume clinical covariates regression coefficients γ_{iq} follow a normal distribution with zero mean and variance τ_{gq}^2.

Gaussian Infinite Mixture Model (GIMM) and functional connectivity: Our model also allows to investigate the functional connectivity by constructing α_{ig} with GIMM. We assume that α_{ig} of all ROIs from one subject arises from the GIMM, in that each ROI can be allocated to a brain network cluster (one mixture component). However, the allocation and number of mixture components are both unknown, and we use θ_{ig} to present the unknown component mean and variance parameter for region g. Given a cluster, say k all ROIs wihtin it follow a normal distribution with the component mean ϕ_{ik} and variance ϵ_{ik}^2. Then, we denote G as a probability measure for the mixture component and specifying distributional assumption for G by using Dirichlet process $DP(h_0, H_0)$ to determine g going to which cluster. H_0 is the base measure of G and h_0 is the concentration parameter.

To complete our Bayesian hierarchical model, we specify diffuse or weakly informative hyperpriors for the variance parameters with inverse Gamma distribution by setting $a_0 = 0.1, b_0 = 0.005, c_0 = 0.1, d_0 = 0.01, e_0 = 0.1, f_0 = 0.01$. The rationale for using these vague priors is to ensure that the results based on the posteriors are primarily governed by the information in the data, especially when no reliable prior

information is available.

Full conditionals: For estimation, MCMC methods with Gibbs sampling was implemented. Applying MCMC is complicated by the massive amount of the data, large number of spatial locations, and the large number of parameters, but the Gibbs-friendly model reduces the computing time and the memory usage. Full conditional distributions are derived for each parameter. For simplification, we denote by

$$\overline{\alpha}_{ig} = \frac{1}{N}\sum_{g\in k}\alpha_{ig}, \quad \overline{\boldsymbol{\beta}}_g = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\beta}_{ig}, \quad \overline{\beta}_{ig} = \frac{1}{Vg}\sum_{v=1}^{V_g}\boldsymbol{\beta}_{ig}(v),$$

$$\overline{\boldsymbol{\mu}} = (\bar{\mu}_1, \ldots, \bar{\mu}_G), \quad \overline{\boldsymbol{\gamma}}_q = (\bar{\gamma}_{1q}, \ldots, \bar{\gamma}_{Gq}), \quad \bar{\mu}_g = \frac{1}{V_g}\sum_{v=1}^{V_g}\mu_g(v),$$

$$\text{and} \quad u_{ig}^2 = (\boldsymbol{\beta}_{ig} - \boldsymbol{\mu}_g - \mathbf{1}\alpha_{ig} - \sum_{q=1}^{Q}\boldsymbol{\gamma}_{gq}x_{iq})'(\boldsymbol{\beta}_{ig} - \boldsymbol{\mu}_g - \mathbf{1}\alpha_{ig} - \sum_{q=1}^{Q}\boldsymbol{\gamma}_{gq}x_{iq})$$

Then the full conditionals are given by:

$$
\begin{aligned}
\boldsymbol{\mu}_g &\sim \text{MVN}\Big(\boldsymbol{\Omega}_g\{\lambda_g^{-2}\mu_{0g}\mathbf{1} + N\sigma_g^{-2}(\overline{\boldsymbol{\beta}}_g - \mathbf{1}\bar{\alpha}_g - \frac{1}{N}\sum_{i=1}^{N}\sum_{q=1}^{Q}\boldsymbol{\gamma}_{gq}x_{iq})\}, \boldsymbol{\Omega}_g\Big) \\
\boldsymbol{\gamma}_{gq} &\sim \text{MVN}\Big(\boldsymbol{\Phi}_{gq}\{\sigma_g^{-2}\sum_{i=1}^{N}x_{iq}(\boldsymbol{\beta}_{gi} - \boldsymbol{\mu}_g - \mathbf{1}\alpha_{ig} - \sum_{q'\neq q}^{Q}\boldsymbol{\gamma}_{gq'}x_{iq'})\}, \boldsymbol{\Phi}_{gq}\Big) \\
\sigma_g^{-2} &\sim \text{Gamma}\Big\{a_0 + NV_g/2, \Big(\frac{1}{b_0} + \frac{1}{2}\sum_{i=1}^{N}\mu_{ig}^2\Big)^{-1}\Big\} \\
\tau_{gq}^{-2} &\sim \text{Gamma}\Big\{e_{0q} + N/2, \Big(\frac{1}{f_{0q}} + \frac{1}{2}\sum_{i=1}^{N}\boldsymbol{\gamma}_{gq}\boldsymbol{\gamma}_{gq}'\Big)^{-1}\Big\} \\
\lambda_g^{-2} &\sim \text{Gamma}\Big\{c_0 + V_g/2, \Big(\frac{1}{d_0} + \frac{(\boldsymbol{\mu}_g - \mathbf{1}\mu_{0g})'(\boldsymbol{\mu}_g - \mathbf{1}\mu_{0g})}{2}\Big)^{-1}\Big\} \\
\alpha_{ig}(g\in k) &\sim N\Big(\Psi\{{\sigma_g}^{-2}V_g(\overline{\beta}_{ig} - \overline{\mu}_g - \sum_{q=1}^{Q}\overline{\boldsymbol{\gamma}}_{jq}x_{iq}) + \epsilon_{ik}^{-2}\phi_{ik}\}, \Psi\Big) \\
\phi_{ik} &\sim N\Big(\Gamma\{n_k\epsilon_{ik}^{-2}\overline{\alpha}_{ik}\}, \Gamma\Big)
\end{aligned}
\tag{4.3.2}
$$

$$
\begin{aligned}
\epsilon_{ik}^{-2} &\sim \text{Gamma}\Big\{a_0 + n_k/2, \Big(\frac{1}{b_0} + \frac{1}{2}\sum_{g\in k}(\alpha_{ig} - \phi_{ik})^2\Big)^{-1}\Big\} \\
\xi_k^{-2} &\sim \text{Gamma}\Big\{a_0 + N/2, \Big(\frac{1}{b_0} + \frac{1}{2}\sum_{i=1}^{N}\phi_{ik}^2\Big)^{-1}\Big\}
\end{aligned}
$$

where we denoted $\boldsymbol{\Omega}_g = (\lambda_g^{-2} + N\sigma_g^{-2})^{-1}\mathbf{I}$, $\boldsymbol{\Phi}_{gq} = (\tau_{gq}^{-2} + {\sigma_g}^{-2}\sum_{i=1}^{N}{x_{iq}}^2)^{-1}\mathbf{I}$, $\Psi = (V_g{\sigma_g}^{-2} + \epsilon_{ik}^{-2})^{-1}$, and $\Gamma = (n_k\epsilon_{ik}^{-2} + \xi_k^{-2})^{-1}$ (For simplicity, we omitted notation for conditional variables).

To determine that region g is allocated to which cluster ($E(g)$), we implement the Chinese restaurant process to sample $E(g)$ from multinomial distribution where the proportion for region g going to cluster k is modeled with the posterior probabilities:

$$P(E(g) = j \mid E(1), \cdots, E(g-1)) \propto \begin{cases} n_k P(\alpha_{ig} \mid E(g) = j) & j = k \\ h_0 P(\alpha_{ig} \mid E(g) = 0) & j = 0 \end{cases}$$

where $j = 0$ indicates to allocate g to a new cluster, the likelihood $P(X \mid E) = \prod_{i=1}^{N} \prod_{k=1}^{K} \prod_{g \in k} f(X|E, \theta_{ik}, \epsilon_{ik}^2)$, where f is a normal probability density.

Furthermore, we illustrate the detailed CRP algorithm for GIMM as:

- Initialization: randomly assign brain regions into an arbitrary number of K_0 clusters $1 <= K_0 <= G$ before all MCMC steps.
- Start MCMC, at each iteration update the other parameters than the mixture component parameters $E(g)$, ϕ_{ik}, ϵ_{ik}^2 and ξ_k^2.
- Update the the assignment of each brain region g to the mixture component, perform the following reassignment:
  - Remove brain region g from its current cluster, given the current assignment of all other brain regions, sample ϕ_{ik} and ϵ_{ik}^2, calculate the probability of this brain region joining each of the existing cluster as well as starting a new cluster.
  - Assign brain region g to the $K + 1$ possible clusters according to probabilities. Update indicator variable $E(g)$ based on the assignment.
- Update ϕ_{ik} and ϵ_{ik}^2 after all regions are assigned, then base current cluster parameters and assignment update all α_{ig}.
- Repeat the above steps in each MCMC iteration.

During the process of MCMC, we not only store the parameters of interest but also the likelihood in order to monitor the clustering effect of the mixture model. The

posteriors of parameters and functions of parameters yield inference via point estimates (mode or median of the samples) or credible intervals. In addition, the results from the Chinese Restaurant Process provides clustering assignment of the regions which could be considered as the frequency property based connectivity.

4.4 Results

We first calculate the frequency domain summary quantity fALFF for all fMRI data. Then we test the normality of all voxels by using Shapiro-Wilks test, and the test results indicate that 1938 (8%) among total 22127 voxels (in the selected ROIs) reject the null hypothesis by setting the $\alpha = 0.05$. Therefore, we feel that the normality assumption is appropriate (Figure 4.1 randomly selects several voxel to show the histograms). Next, we apply our our Bayesian hierarchical model to the fALFF quantity for the resting-state fMRI data. We also conducted chain diagnosis based on the results, the evidence indicate the chain converge after burn-in iterations (Figure 4.2), due to massive number of parameters we choose the chip parameter as 5. In the data analysis, we use 90 AAL regions for this study, the number of voxels within each ROI ranges from 90 to 650.

4.4.1 Voxel-level results

The analysis results provide voxel-specific inference, and we focus on the frequency band ratio (fALFF) difference between mildly depressed and severely depressed patients. We fit the model by using the two groups separately and contrast the μ for each voxels (Bowman 2008) Figure 4.3 and 4.4 display axial slices showing voxels exhibiting difference between mildly depressed patients and severely depressed patients based on the metric of fALFF. The highlighted voxels have posterior probabilities

Figure 4.1: Histogrms of Selected Voxels' fALFF across subjects

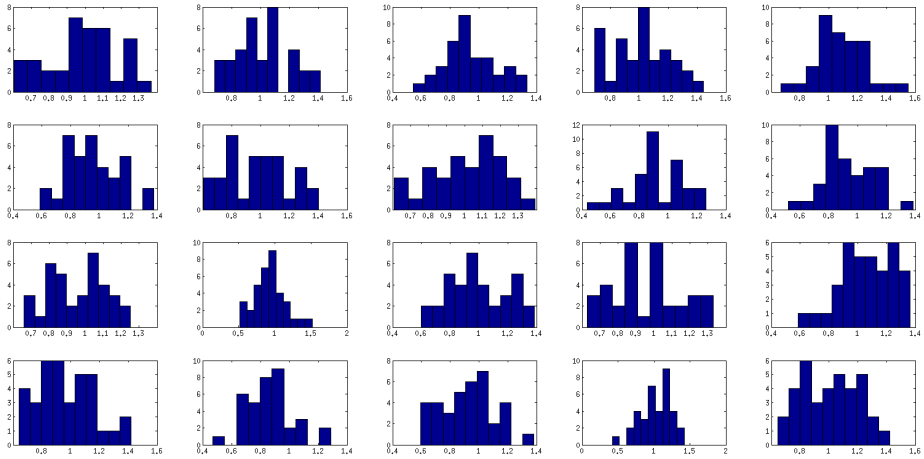
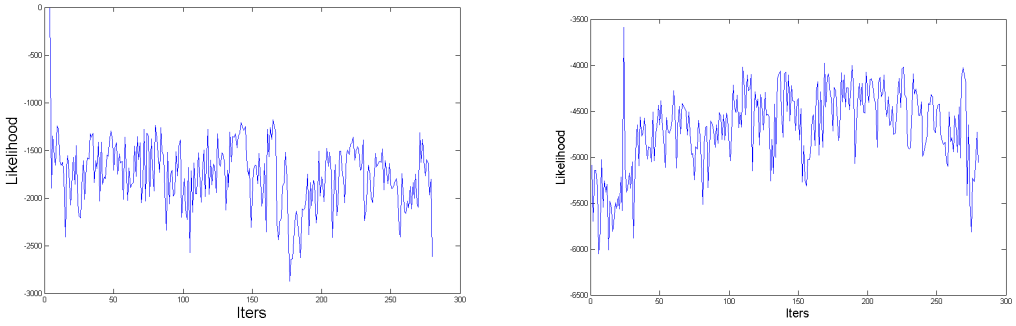


Figure 4.2: Likelihood trace plots for two groups (recorded once for every 5 iterations)



exceeding 0.80 of increased or decreased fALFF difference between two groups. The figures show that the severely depressed group have higher fALFF at cingulate gyrus especially at right anterior and posterior cingulate gyrus, somasensory gyrus, left medial frontal gyrus; while mildly depressed group have higher fALFF right prefrontal gyrus and precentral gyrus.

4.4.2 Regional results

Our model also enables the investigation of fALFF levels at a regional level. Comparing to task-induced fMRI analyses, the resting-state fMRI regional-level analyses may

Figure 4.3: Brain map showing the voxels with voxel-specific posterior probabilities exceeding 0.8 that severely depressed patients have higher fALFF than mildly depressed patients

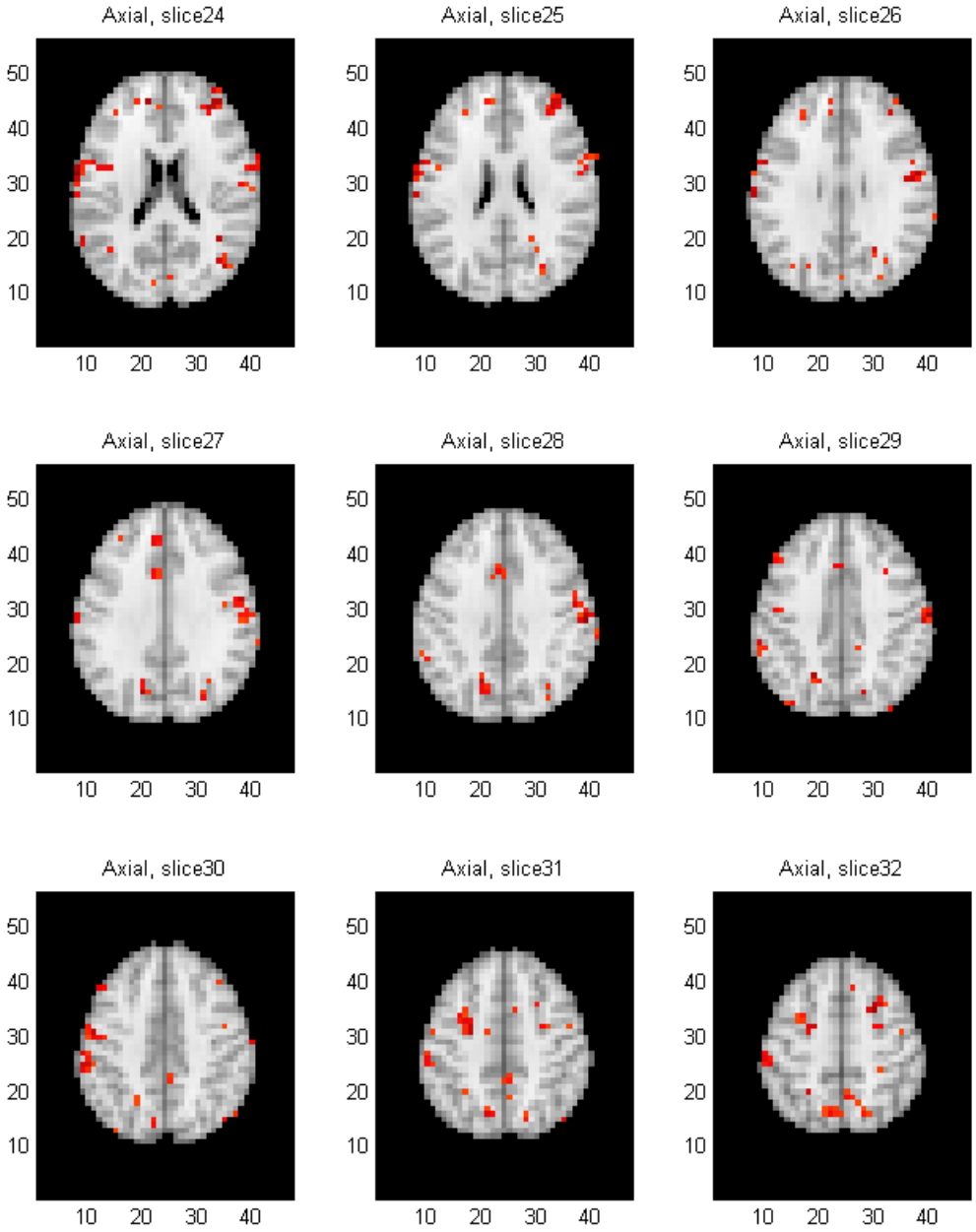


Figure 4.4: Brain map showing the voxels with voxel-specific posterior probabilities exceeding 0.8 that mildly depressed patients have higher fALFF than severely depressed patients

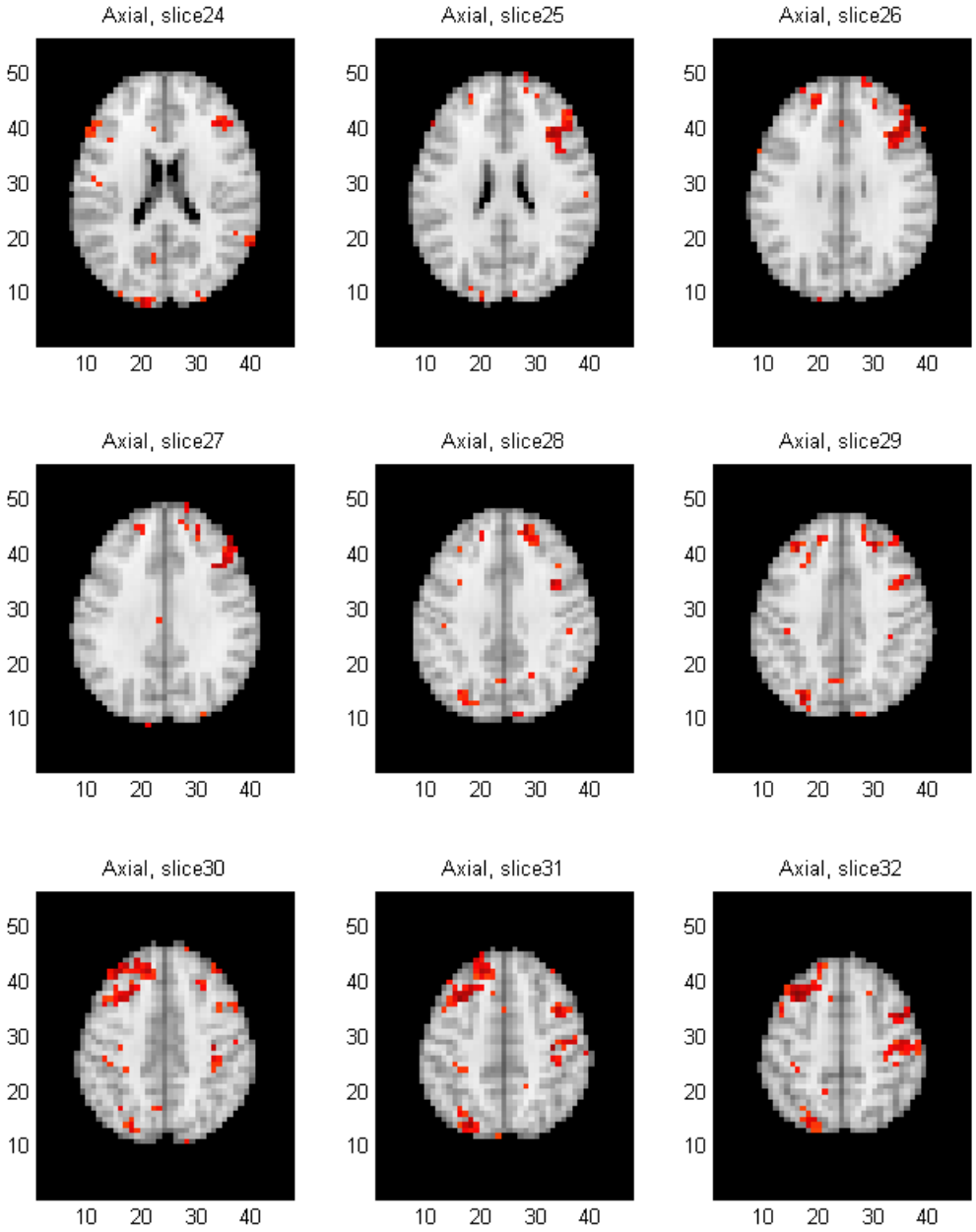
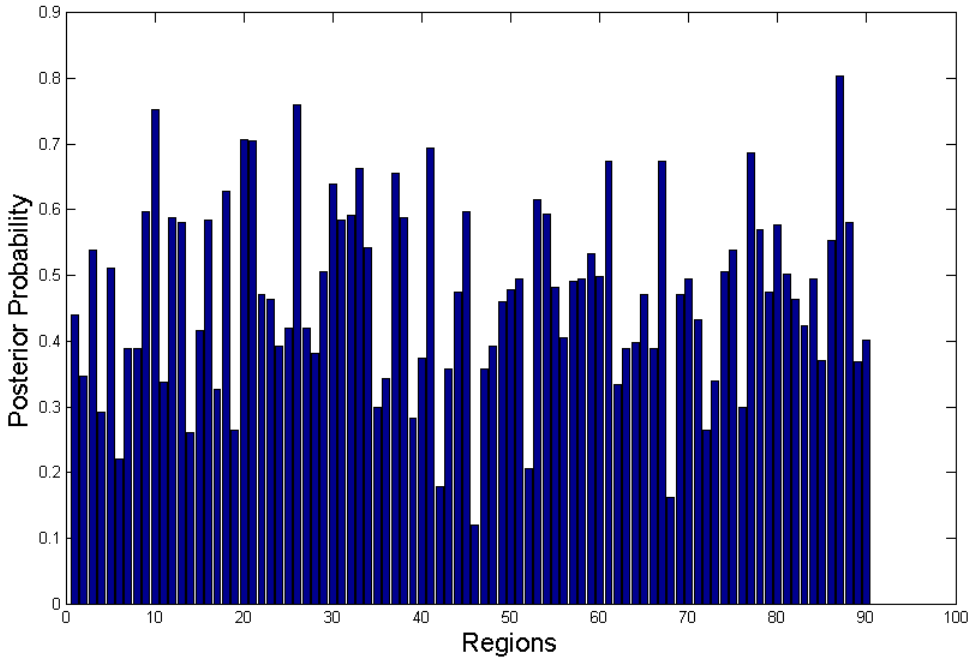


Figure 4.5: Brain map showing the voxels with voxel-specific posterior probabilities that severely depressed patients have higher fALFF than mildly depressed patients

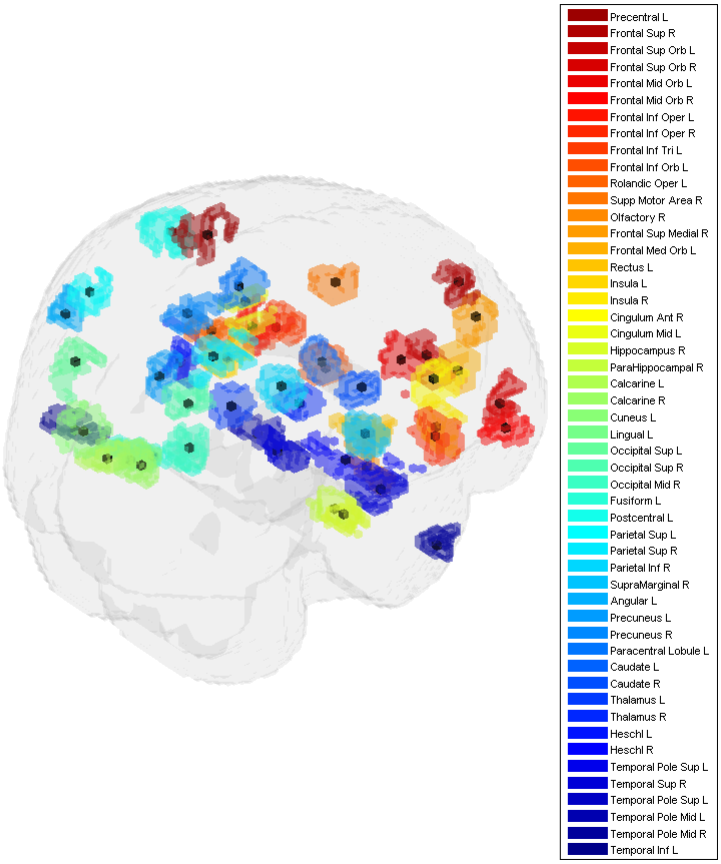


include more variance between voxels within same region without task orientation. Therefore, the difference of region-level fALFF between the two groups tend to be attenuated by such variance. We summarize the posterior probabilities in Figure 4.5, and it indicates that only a handful of regions exceed the thresholds with above 0.80 or below 0.20. The mildly depressed group have higher fALFF at Amygdala Right, Lingual Left, at Thalamus Left than severely depressed group.

4.4.3 Regional connectivity networks

In addition to the fALFF intensity analysis at a voxel and a region level, the estimation of the infinite mixture model yield connectivity network clustering results. We obtain different clustering results for different groups and since CRP does not require the preselection of the numbers of clusters and the model determines them adaptively, the

Figure 4.6: Connectivity network clusters for mildly depressed patients: the major cluster



two groups have different number clusters(6 for mildly depressed and 8 for severely depressed patients). Each clinical group contain a major cluster which includes most of default mode network regions such as medial prefrontal cortex (MPFC), medial temporal lobes (MTLs), posterior cingulate cortex (PCC)/retropslenial cortex, and medial occipital cortex as well as the regions spatially adjacent to them. In addition, each clinical group has smaller networks that are spatially adjacent or functionally correlated. The brain clusters are demonstrated in the following figures.

Figure 4.7: Connectivity network clusters for severely depressed patients: the major cluster

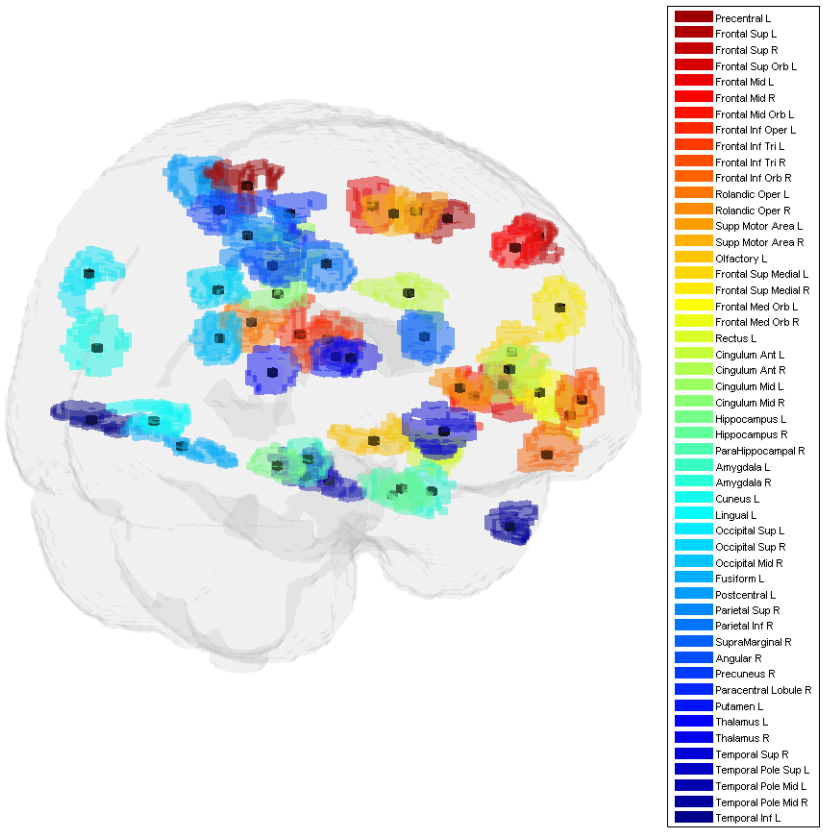


Figure 4.8: Connectivity network clusters for mildly depressed patients: the smaller cluster

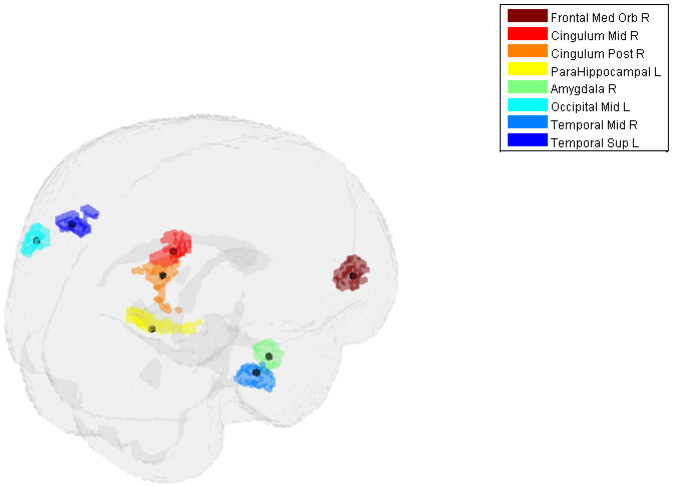
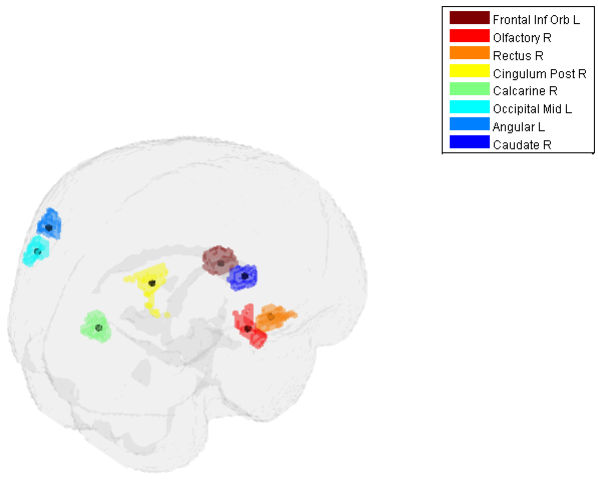


Figure 4.9: Connectivity network clusters for severely depressed patients: the smaller cluster



4.5 Discussion

We propose a unified Bayesian hierarchical model for analyzing resting-state fMRI neuroimaging data for simultaneous frequency descriptor (fALFF) analysis as well as connectivity analysis. This model provides a pathway to investigate neuroactivity at certain frequency band at both a voxel level and at a regional level. Therefore, we can estimate voxel-specific fALFF to detect partially activated regions as well as detect regional summaries while accounting for the correlations between the parameters. Our model accounts for prominent spatial correlations or functional connections in the brain by using the subject-wise random effect and infinite mixture model. Comparing our approach to the use Wishart distribution as a conjugate prior for the covariance matrix between regions, the CRP and infinite mixture model not only account for the spatial correlation but also provides the network clustering results, and is not subject to the restriction that region number must be less than the minimum group subject number. The discovered networks show functional coherence based on previous studies or/and spacial adjacency. Regarding to the computational load, the MCMC estimation procedures produce samples from the joint posterior distribution of all model parameters, which facilitates the point and interval estimate of those parameters and functions of the parameters. In our example, the voxel and region level fALFF analysis can detect difference between clinical groups and the clustering networks reveal the frequency coherence. Despite the complex model structure and rather rich formulation of our Bayesian hierarchical model as well as massive number of the data, computations for estimation are manageable. For instance, our example (36 subjects) took less than 40 minutes each group based on a PC i7 CPU (3.1Ghz) and 8GB RAM.

Chapter 5

Summary and Future Work

The nature of neuroimaging data: high dimensionality, complex spatial and temporal structure, and substantial noises at each stage of data acquisition raises challenges for data analysis and information retrieval. Our goal is to decipher or mine the meaningful neurophysiological and neuropathological information from the massive and complex data sets. To achieve this aim, therefore we usually do not restrict ourselves by only using one or a few specific statistical tools for, rather we often are problem-solving oriented. In this dissertation, we develop three novel models to tackle several challenges in current neuroimaging data analysis by using statistical techniques such as machine learning, Bayesian modeling, graph theory, and signal processing.

First, we develop a general classification approach for repeatedly measured high-dimensional data based on support vector classifier. We use neuroimaging data as predictors and disease status or treatment response as dependent variable. Rather than treating the features at each time as independent variable, we also seek the optimal temporal trend while calculating the separating hyperplane parameters. In the future, we mainly consider two extensions based on this method: i) to develop

a scheme for multiclass classifier of longitudinal high dimensional data; ii) to develop a 'wrapper' feature selection method based LSVC. Both of these two extensions are subject to expensive computational cost, efficient algorithms are desired for the implementation.

Second, we propose a Bayesian hierarchical framework first to summarize voxel level connectivity for the region level connectivity analysis and further account the clinical covariates. Based this work, we plan to jointly model structural connectivity and functional connectivity by letting structural connectivity as predictors for functional connectivity at the region level and the model will determine the fibre tracks along which region pairs could affect the FC for the chosen region pair. In addition, we could also extend all above frameworks for the longitudinal setting by accounting for the temporal correlation and predict the later images.

Lastly, we build a unified hierarchical Bayesian framework to jointly quantify localized resting-state brain 'frequency band' activity as well as the functional connectivity brain network by allocating each region to a latent network cluster. A natural extension could be rather using a single scalar fALFF descriptor, we use a series of power spectral band descriptors for the analysis by using more information. In the future, we will conduct more model validation and sensitivity analysis such as how subject-wise clustering varies from group clustering and MCMC performance on larger data sets. The extended framework could also be applied to task-induced fMRI data by summarize each HRF to scalar at each event, then use the vector of summary statistics for all events as input of the model, then conduct temporal (frequency for resting-state) and spatial localized activity analysis, while accounting for the spatial correlation by clustering coherently behaved regions as one component.

Bibliography

Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E. (2005). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs.*J Neurosci* 26, 63–72.

Behrens T EJ, Woolrich M W, Jenkinson M, Johansen-Berg H, Nunes R G, Clare S, Matthews P M, Brady J M, and Smith S M. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn Reson Med.* 2003; 50(5):1077-88.

Behrens T.E.J., Johansen-Berg H., Jbabdi S., Rushworth M.F.S., and Woolrich M.W. (2007). Probabilistic diffusion tractography with multiple fibre orientations. What can we gain? *NeuroImage*, 23, 144–155.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone., C. J. (1983). *Classification and Regression Trees.* Wadsworth.

Breiman, Leo. "Random Forests". *Machine Learning.* 2001; 45 (1): 5–32.

Bowman, F. Spatio-temporal modeling of localized brain activity. *Biostatistics* 2005; 6, 558–575.

Bowman, F., Caffo, B., Bassett, S., and Kilts, C. A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage* 2008; 39, 146–156.

Bowman, F. and Kilts, C. Modeling intra-subject correlation among repeated scans in positron emission tomography (PET) neuroimaging data. *Human Brain Mapping* 2003, 20, 59–70.

Bowman, F. and Patel, R. (2004). Identifying spatial relationships in neural processing using a multiple classification approach. *NeuroImage* 2004; 23, 260–268.

Bowman, F., Patel, R., and Lu, C. Methods for detecting functional classi- cations in neuroimaging data. *Human Brain Mapping* 2004; 23, 109–119.

Bullmore, E., and Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience,* 2009; 10, 186–198.

Caffo B., Crainiceanu C., Verduzco G., Mostofsky S., Spear-Bassett S., and Pekar J. (2010). Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimerdisease risk. *Neuroimage* 51, 1140–1149.

Calhoun, V. and Pekar, J. When and where are components independent? On the applicability of spatial- and temporal- ICA to functional MRI data, volume 11. Sixth International Conference on Functional Mapping of the Human Brain., Neuroimage, Academic Press, 2000.

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. Spatial and temporal independent component analysis of functional MRI data containing a pair of task- related waveforms. *Human Brain Mapping.* 2001; 13, 43–53.

Chen S., Hong D., and Shyr Y., Wavelets-based Procedures for Proteomic Mass Spectrometry Data Processing, *Computational Statistics and Data Analysis*. 2007; 52: 211–220.

Craddock R.C., Holtzheimer P.E., Hu X.P., and Mayberg, H.C., Disease State Prediction From Resting State Functional Connectivity. *M agnetic Resonance in Medicine* 2009 62 1619–28.

Common, P. Independent component analysis, a new concept *Signal Processing* 1994; 36, 287–314.

Cox, D. and Savoy, R. Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage.* 2003 19, 261–270.

Derado, G., Bowman, F., and Kilts, C. Modeling the spatial and temporal dependence in fMRI data. *Biometrics* 2010; 66, 949–957.

Efron, B., Tibshirani, R., Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* 2002; 23, 70–86.

Efron B. (2005). Local false discovery rates. Department of Statistics Technical Report 234, Stanford University.

Evans, A., Collins, D., Mills, S., Brown, E., Kelly, R., and Peters, T. (1993). 3d statistical neuroanatomical models from 305 mri volumes. *In Proc. IEEE Nucl. Sci. Symp. Med. Imaging Conf.* 1993; pages 1813–1817.

Evans, K., Dougherty, D., Pollack, M., and Rauch, S. Using neuroimaging to predict treatment response in mood and anxiety disorders. *Ann Clin Psychiatry* 2006; 18, 33–42.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*; 2001. 1348–1360.

Fadili, M., Ruan, S., Bloyet, D., and Mazoyer, B. (2001). On the number of clusters and the fuzziness index for unsupervised FCA application to BOLD fMRI time series. *Medical Image Analysis* 2001; 5, 55–67.

Freyermuth, J-M., Ombao, H. and von Sachs, R. (2010). Spectral Estimation from Replicated Time Series: An Approach Using the Tree-Structured Wavelets Mixed Effects Model. *Journal of the American Statistical Association* 105, 634–646.

Friston, K., Glaser, D. E., Henson, R. N., Kiebel, S., Phillips, C., and Ashburner, J. Classical and Bayesian inference in neuroimaging: applications. *Neuroimage* 2002; 16, 484–512.

Friston, KJ, Harrison, L, Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302.

Fu CHY, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SCR, Brammer MJ. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol Psychiatry* 2008; 63:656-662.

Garey, L. (1994). Brodman's localisation in the cerebral cortex: the principles of comparative localisation based on cytoarchitectonics. Springer, London.

Greicius MD, Krasnow B, Reiss AL, Menon V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc Natl Acad Sci USA*. 2003; 100:253–258.

Greicius MD, Supekar K, Menon V, Dougherty RF. Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cereb Cortex.* 2009; 19: 72–78.

Guo, Y. and Pagnoni, G. A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImag.* 2008; 42, 1078–1093.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning.* 2002; 46(1-3):389–422.

Guyon I. and Elisseeff A., Introduction to Variable and Feature Selection. *Journal of Machine Learning Research.* 2003; 3, pp. 1157–1182.

Hagmann P., Thiran JP., Jonasson L., Vandergheynsta P., ClarkecS., Maederb P., Meulib R. (2003). DTI mapping of human brain connectivity: statistical fibre tracking and virtual dissection *Neuroimage* 19(3), 545–554.

Hastie, T. and Tibshirani, R. 1990, *Generalized Additive Models*, Chapman and Hall.

Hastie T. , Tibshirani R., and Friedman J. *The Elements of Statistical Learning* 2nd ed. Springer series in statistics. Springer, New York, 2009.

Hastie T., Rosset S., Tibshirani R., and Zhu J. The Entire Regularization Path for the Support Vector Machine. *Journal of Machine Learning Research*, 5, 2004; 1391–1415.

Higgins, J. Brain imaging in psychiatry, chapter Normal brain anatomy imaged by CT and MRI, pages 63–107. Oxford: Blackwell Science, 1996.

C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann Predicting human resting-state functional connectivity from structural connectivity *Proc Natl Acad Sci* 2009; 106(6): 2035–2040.

Hyvarinen, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks.* 1999; 10, 626–634.

Jirsa VK., McIntosh AR. (2007). Handbook of Brain Connectivity. Springer.

LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. Support vector machines for temporal classification of block design fMRI data. *Neuroimage* 2005; 26:317-329.

Lindquist M. (2008). The Statistical Analysis of fMRI Data. *Statistical Science*, 23(4), 439-464.

Maruish MR. (1999)The Use of Psychological Testing for Treatment Planning and Outcomes Assessment. Lawrence Erlbaum Associates.

McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., and Sejnowski, T. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping.* 1998; 6, 160–188.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. Learning to decode cognitive states from brain images. *Machine Learning*, 2004; 57:145 – 175.

Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *Neuroimage* 2005; 28: 980-995.

Ogawa, S., Lee, T. M., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciencies, Biophysics* 1990; 87, 9868–9872.

Ombao, H, Shao, X., Rykhlevskaia, E, Fabiani, M and Gratton, G. (2008). Spatio-Spectral Analysis of Brain Signals, *Statistica Sinica*, 18, 1465-1482.

Parker G., Haroon H., Wheeler-Kingshott C. (2003). A framework for a streamline-based probabilistic index of connectivity (PICo) using a structural interpretation of MRI diffusion measurements *Journal of Magnetic Resonance Imaging* 18 (2),242–254.

Rencher, A. *Methods of Multivariate Analysis*, 2nd ed. John Wiley & Sons, Inc., New York., 2nd edition, 2002.

Ressler KJ. and Mayberg HS. (2007). Targeting abnormal neural circuits in mood and anxiety disorders: from the laboratory to the clinic. *Nature Neuroscience* 10(9): 1116–1124.

Schutte, NS. and John MM. (1995). Sourcebook of Adult Assessment Strategies. New York: Plenum Press.

Seminowicz DA, Mayberg HS, McIntosh AR, Goldapple K, Kennedy S, Segal Z, Rafi-Tari S. Limbic-frontal circuitry in major depression: a path modeling metanalysis. *Neuroimage* 22(1):409–18.

Skudlarski P, et al. Measuring brain connectivity: Diffusion tensor imaging validates resting state temporal correlations. *NeuroImage.* 2008; 43:554 561.

Sporns O (2010). Networks of the Brain. MIT Press.

Sun F., Miller, L., D'Esposito, M., (2004). Measuring interregional functional connectivity using coherence and partial coherence analysesof fmri data. *Neuroimage* 21, 647–658.

Talairach, J. and Tournoux, P. Co-planar stereotaxic atlas of the human brain. Thieme Medical Publishers, Inc., New York, 1988.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. Automated anatomical labelling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single subject brain. *NeuroImage* 2002; 15, 273–289.

Vapnik V. *The nature of statistical learning theory.* New York: Springer; 1996; 188 p.

Vapnik V. *Statistical Learning Theory*. Wiley, 1998.

Wahba, G., *Spline Models for Observational Data*. 1990; SIAM, Philadelphia, PA.

Woods, R., Grafton, S., Holmes, C., Cherry, S., and Mazziotta, J. Automated image registration: I. general methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography* 1998; 22, 139–152.

Woods, R., Grafton, S., Watson, J., N.L., S., and J.C., M. Automated image registration: II. Intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography* 1998; 22, 153–165.

Worsley, K. and Friston, K. Analysis of fMRI time-series revisited-again. *Neuroimage* 1995; 2, 173–181.

Zou H. and Hastie T. Regularization and variable selection via the elastic net, *Journal Of The Royal Statistical Society Series B,* 2005; 67(2), 301–320.