Abstract

Neurological diseases constitute the leading disease burden worldwide. Existing symptom-based diagnostic methods are often insufficient to detect many of these diseases in their early stages. Recent advances in neuroimaging technologies have enabled non-invasive examination of the brain, which facilitates localization of disease-induced effects directly at the source. In particular, functional magnetic resonance imaging (fMRI) has become one of the dominant means for studying brain activity in healthy and diseased subjects. However, the low signal-to-noise ratio, the typical small sample size, and the large inter-subject variability present major challenges to fMRI analysis. Standard analysis approaches are largely univariate, which underutilize the available information in the data. In this thesis, we present novel strategies for activation detection, region of interest (ROI) characterization, functional connectivity analysis, and brain decoding that address many of the key challenges in fMRI research. Specifically, we propose: 1) new formulations for incorporating connectivity and group priors to better inform activation detection, 2) the use of invariant spatial features for capturing the often-neglected spatial information in ROI characterization, 3) an evolutionary group-wise approach for dealing with the high inter-subject variability in functional connectivity analysis, and 4) a generalized sparse regularization technique for handling ill-conditioned brain decoding problems. On both synthetic and real data, we showed that exploitation of prior information enables more sensitive activation detection, more refined ROI characterization, more robust functional connectivity analysis, and more accurate brain decoding over the current state-of-the-art. All of our results converged to the conclusion that integrating prior information is beneficial, and oftentimes, essential for tackling the challenges that fMRI research present.
Preface

The list of publications resulting from my thesis work is provided on pages iv to vi arranged according to chapters.

[P1–P5, P7–P9, P12–P16]: I was the primary author of these papers and the main contributor to the design, development, and testing of the presented methods under the supervision of Dr. Rafeef Abugharbieh. The papers were edited by the co-authors. Dr. Martin McKeown of the Pacific Parkinson’s Research Center (PPRC) provided data (Clinical Research Ethics Board, H04-70177) and input to [P7–P9]. Dr. Ghassan Hamarneh of Simon Fraser University (SFU) provided input to [P1, P3–P5]. Initial method testing of [P3] was partly conducted by Kelvin So of Berkeley University under my supervision.

[P6, P11]: I was the primary author of these papers and the main contributor to the design, development, and testing of the presented methods under the supervision of Dr. Rafeef Abugharbieh and Dr. Bertrand Thirion of the French National Institution for Research in Computer Science and Control (INRIA). The papers were edited by the co-authors. Dr. Jean Baptiste Poline of INRIA provided data and input to these works. Dr. Gael Varoquaux of INRIA also provided input.

[P10]: Dr. Samantha Palmer of PPRC was the primary author of this paper. The co-authors and I were responsible for editing. The experimental data were provided by Dr. Martin McKeown of PPRC (Clinical Research Ethics Board, H04-70177). The spatial analysis portion of this work was performed by me. The network analysis portion of this work was conducted by Dr. Samantha Palmer.

[P17–P20]: I was the primary author of these papers and the main contributor to the design, development, and testing of the presented methods under the supervision of Dr. Rafeef Abugharbieh. The papers were edited by the co-authors. The data were made publicly available online by Dr. Tom Mitchell of Carnegie Mellon University. Initial method testing of [P17] was partly conducted by Arash Vahdat of SFU under my supervision.
Chapter 2: Prior-informed Activation Detection


Chapter 3: Spatial Characterization of Regional Activation Patterns


Chapter 4: Group-wise Functional Network Analysis


Chapter 5: Generalized Sparse Classification


# Table of Contents

Abstract.................................................................................................................................... ii  
Preface..................................................................................................................................... iii  
Table of Contents .................................................................................................................. vii  
List of Tables ........................................................................................................................... x  
List of Figures ......................................................................................................................... xi  
List of Abbreviations ........................................................................................................... xiii  
Acknowledgements ............................................................................................................... xv  
Dedication ............................................................................................................................. xvi  

## 1 Introduction.................................................................................................................. 1  
1.1 Problem Statement and Motivation....................................................................................... 1  
1.2 Functional Magnetic Resonance Imaging............................................................................. 3  
1.3 Activation Detection .......................................................................................................... 6  
1.3.1 Intra-subject Analysis ....................................................................................................... 6  
1.3.2 Group Analysis ............................................................................................................... 9  
1.4 ROI Characterization ......................................................................................................... 11  
1.5 Functional Connectivity Analysis ...................................................................................... 12  
1.5.1 Intra-subject Analysis ..................................................................................................... 12  
1.5.2 Group Analysis ............................................................................................................. 14  
1.6 Brain Decoding ................................................................................................................... 15  
1.7 Thesis Contributions .......................................................................................................... 18  
1.7.1 Activation Detection ....................................................................................................... 18  
1.7.2 ROI Characterization ..................................................................................................... 19  
1.7.3 Functional Connectivity Analysis ................................................................................... 20  
1.7.4 Brain Decoding ............................................................................................................... 20  

## 2 Prior-informed Activation Detection ............................................................................... 21  
2.1 Univariate Activation Detection with GLM ........................................................................... 22  
2.2 Univariate Activation Detection with GMM ......................................................................... 23
2.3 Modeling Local Spatial Correlations with MRF ................................................................. 25
2.4 Modeling Functional Connectivity with RW ................................................................. 26
2.5 Modeling Group Information with GMRF ......................................................................... 28
  2.5.1 GMRF Formulation ........................................................................................................ 29
  2.5.2 Implementation Details ................................................................................................... 30
2.6 Synthetic Task Data Experiments ....................................................................................... 31
  2.6.1 Data Generation .............................................................................................................. 31
  2.6.2 Results and Discussion ................................................................................................... 32
2.7 Real Task Data Analysis ..................................................................................................... 36
  2.7.1 Materials ......................................................................................................................... 36
  2.7.2 Results and Discussion ................................................................................................... 37
2.8 Resting State-informed Activation Detection ................................................................. 40
  2.8.1 RS-informed Activation Model ...................................................................................... 41
  2.8.2 Posterior Activation Effects Estimation ........................................................................ 42
  2.8.3 Functional Connectivity Estimation ............................................................................... 42
  2.8.4 Hyper-parameters Estimation ......................................................................................... 43
2.9 Synthetic RS/Task Paired Data Experiments ...................................................................... 45
  2.9.1 Data Generation .............................................................................................................. 45
  2.9.2 Results and Discussion ................................................................................................... 45
2.10 Real RS/Task Paired Data Analysis .................................................................................... 47
  2.10.1 Materials ......................................................................................................................... 47
  2.10.2 Results and Discussion ................................................................................................... 47
2.11 Summary ............................................................................................................................. 49

3 Spatial Characterization of ROI Activation ........................................................................ 50
  3.1 3D Moment-based Invariant Spatial Features ................................................................. 51
3.2 Effects of Spatial Normalization ............................................................................................ 53
  3.2.1 Assessment Methods ....................................................................................................... 54
  3.2.2 Results and Discussion ................................................................................................... 56
  3.2.2.1 Effects of Spatial Normalization on Smaller ROIs ................................................ 57
  3.2.2.2 Effects of Spatial Normalization on Larger ROIs .................................................. 61
3.3 Effects of L-dopa on Parkinson’s Disease ............................................................................ 64
  3.3.1 Statistical Analysis ......................................................................................................... 64
  3.3.2 Results and Discussion ................................................................................................... 65
3.4 Large Scale Real Data Experiment ..................................................................................... 67
List of Tables

Table 2.1  Summary of activation detection methods............................................................ 49
Table 4.1  Summary of functional connectivity analysis methods. ......................................... 89
Table 5.1  Summary of feature selection approaches for fMRI classification......................... 105
List of Figures

Figure 1.1 Typical fMRI block design experiment................................................................. 4
Figure 1.2 Sample raw intensity time course of an activated voxel. ................................. 4
Figure 2.1 Pictorial depiction of GLM. ............................................................................. 22
Figure 2.2 Graphical depiction of GMM. ........................................................................... 24
Figure 2.3 Histogram of correlation values between random Gaussian noise time courses.. 27
Figure 2.4 Proposed GMRF graph structure................................................................. 29
Figure 2.5 Synthetic data results. ................................................................................... 34
Figure 2.6 TPR, FPR, and DSC at two typical SNR levels. ............................................... 35
Figure 2.7 Homunculus................................................................................................. 37
Figure 2.8 Real data results for control subjects............................................................... 38
Figure 2.9 Real data results for PD subjects................................................................. 39
Figure 2.10 Synthetic data results. ................................................................................. 46
Figure 2.11 Real data results......................................................................................... 48
Figure 3.1 Spatiotemporal ROI analysis pipeline......................................................... 52
Figure 3.2 Extraction of ROI activation statistics.............................................................. 55
Figure 3.3 Discriminability comparisons between the contrasted procedures for extracting
ROI activation statistics........................................................................................................ 57
Figure 3.4 ROI activation statistics in the right thalamus under the slow condition (Section
2.7.1). .......................................................................................................................... 58
Figure 3.5 ROI activation statistics in the right thalamus under the fast condition (Section
2.7.1). .......................................................................................................................... 59
Figure 3.6 ROI activation statistics in the right cerebellum under the slow condition (Section
2.7.1). .......................................................................................................................... 62
Figure 3.7 ROI activation statistics in the right cerebellum under the fast condition (Section
2.7.1). .......................................................................................................................... 63
Figure 3.8 Spatial focusing effect of L-dopa seen in a typical PD subject ...................... 66
Figure 3.9 Normalization of the spatial extent of ROI activation statistics by L-dopa. .... 66
Figure 3.10 Percentage of ROIs detected with significant activation differences using $J_1$. 69
Figure 3.11 Activation detection comparisons between $J_1$ and mean intensity ........... 69
Figure 4.1  Synthetic data results. ........................................................................................... 80
Figure 4.2  Convergence of GRD. .......................................................................................... 82
Figure 4.3  Outlier analysis ................................................................................................... 83
Figure 4.4  Effects of varying $\lambda$. .................................................................................. 84
Figure 4.5  Real fMRI data results. ....................................................................................... 87
Figure 5.1  Predictor matrix. .................................................................................................. 98
Figure 5.2  Spatiotemporal neighborhood. ........................................................................... 99
Figure 5.3  Prediction accuracy comparisons. ....................................................................... 101
Figure 5.4  Classifier weight patterns. .................................................................................. 102
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCc</td>
<td>Caudal anterior cingulate cortex</td>
</tr>
<tr>
<td>ACCr</td>
<td>Rostral anterior cingulate cortex</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer’s disease</td>
</tr>
<tr>
<td>AMY</td>
<td>Amygdale</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>ANSPCA</td>
<td>Adaptive negative sparse principal component analysis</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregressive</td>
</tr>
<tr>
<td>ASLDA</td>
<td>Anatomically-informed sparse linear discriminant analysis</td>
</tr>
<tr>
<td>ASSLDA</td>
<td>Anatomically-informed spatial smooth sparse linear discriminant analysis</td>
</tr>
<tr>
<td>ASTSLDA</td>
<td>Anatomically-informed spatiotemporally smooth sparse linear discriminant analysis</td>
</tr>
<tr>
<td>BOLD</td>
<td>Blood oxygen level dependent</td>
</tr>
<tr>
<td>CAU</td>
<td>Caudate</td>
</tr>
<tr>
<td>CER</td>
<td>Cerebellum</td>
</tr>
<tr>
<td>CM</td>
<td>Connectivity-informed model</td>
</tr>
<tr>
<td>CTC</td>
<td>Cerebello-thalamo-cortical</td>
</tr>
<tr>
<td>DSC</td>
<td>Dice similarity coefficient</td>
</tr>
<tr>
<td>ED</td>
<td>Euclidean distance</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalography</td>
</tr>
<tr>
<td>ENLDA</td>
<td>Elastic net linear discriminant analysis</td>
</tr>
<tr>
<td>EPI</td>
<td>Echo planar imaging</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional magnetic resonance imaging</td>
</tr>
<tr>
<td>FRP</td>
<td>False positive rate</td>
</tr>
<tr>
<td>GE</td>
<td>Graph embedding</td>
</tr>
<tr>
<td>GL</td>
<td>Graphical least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>GL-CM</td>
<td>Connectivity-informed model with graphical LASSO</td>
</tr>
<tr>
<td>GLM</td>
<td>General linear model</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian mixture model</td>
</tr>
<tr>
<td>GMRF</td>
<td>Group Markov random field</td>
</tr>
<tr>
<td>GRD</td>
<td>Group replicator dynamics</td>
</tr>
<tr>
<td>GRF</td>
<td>Gaussian random field</td>
</tr>
<tr>
<td>GRW</td>
<td>Group random walker</td>
</tr>
<tr>
<td>GSR</td>
<td>Generalized sparse regularization</td>
</tr>
<tr>
<td>HAMMER</td>
<td>Hierarchical attribute matching mechanism for elastic registration</td>
</tr>
<tr>
<td>HDP</td>
<td>Hierarchical Dirichlet process</td>
</tr>
<tr>
<td>HIP</td>
<td>Hippocampus</td>
</tr>
<tr>
<td>HRF</td>
<td>Hemodynamic response function</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent component analysis</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>Independently and identically distributed</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear discriminant analysis</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>L-Dopa</td>
<td>Levodopa</td>
</tr>
<tr>
<td>MCICA</td>
<td>Motion-corrected independent component analysis</td>
</tr>
<tr>
<td>MEG</td>
<td>Magnetoencephalography</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>MM</td>
<td>Mixture model</td>
</tr>
<tr>
<td>MNI</td>
<td>Montreal Neurological Institutes</td>
</tr>
<tr>
<td>MRF</td>
<td>Markov random field</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
</tr>
<tr>
<td>MS</td>
<td>Multiple Sclerosis</td>
</tr>
<tr>
<td>M1</td>
<td>Primary motor cortex</td>
</tr>
<tr>
<td>OAS</td>
<td>Oracle approximating shrinkage</td>
</tr>
<tr>
<td>OAS-CM</td>
<td>Connectivity-informed model with oracle approximating shrinkage</td>
</tr>
<tr>
<td>PAL</td>
<td>Pallidum</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PCC</td>
<td>Posterior cingulate cortex</td>
</tr>
<tr>
<td>p.d.</td>
<td>Positive definite</td>
</tr>
<tr>
<td>PD</td>
<td>Parkinson’s disease</td>
</tr>
<tr>
<td>PDar</td>
<td>Akinetic Parkinson’s disease</td>
</tr>
<tr>
<td>PDt</td>
<td>Tremor dominant Parkinson’s disease</td>
</tr>
<tr>
<td>PET</td>
<td>Positron emission tomography</td>
</tr>
<tr>
<td>PM</td>
<td>Premotor area</td>
</tr>
<tr>
<td>PMA</td>
<td>Primary motor area</td>
</tr>
<tr>
<td>PUT</td>
<td>Putamen</td>
</tr>
<tr>
<td>RD</td>
<td>Replicator dynamics</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of interest</td>
</tr>
<tr>
<td>RS</td>
<td>Resting state</td>
</tr>
<tr>
<td>R-UM</td>
<td>Univariate model with ridge penalty</td>
</tr>
<tr>
<td>RW</td>
<td>Random walker</td>
</tr>
<tr>
<td>SLDA</td>
<td>Sparse linear discriminant analysis</td>
</tr>
<tr>
<td>SMA</td>
<td>Supplementary motor area</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-noise ratio</td>
</tr>
<tr>
<td>SPG</td>
<td>Spectral projected gradient</td>
</tr>
<tr>
<td>SPECT</td>
<td>Single photon emission computed tomography</td>
</tr>
<tr>
<td>SSA</td>
<td>Somatosensory area</td>
</tr>
<tr>
<td>SSLDA</td>
<td>Spatially smooth sparse linear discriminant analysis</td>
</tr>
<tr>
<td>STC</td>
<td>Striato-thalamo-cortical</td>
</tr>
<tr>
<td>STSLDA</td>
<td>Spatiotemporally smooth sparse linear discriminant analysis</td>
</tr>
<tr>
<td>S-UM</td>
<td>Standard univariate model</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>SVM-RFE</td>
<td>Support vector machine-recursive feature extraction</td>
</tr>
<tr>
<td>THA</td>
<td>Thalamus</td>
</tr>
<tr>
<td>TRP</td>
<td>True positive rate</td>
</tr>
<tr>
<td>UM</td>
<td>Univariate model</td>
</tr>
<tr>
<td>UPDRS</td>
<td>Unified Parkinson’s disease rating scale</td>
</tr>
<tr>
<td>WHO</td>
<td>World health organization</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to thank my advisor, Dr. Rafeef Abugharbieh, for supervising me throughout my graduate studies and providing me the freedom to explore different research directions. I would also like to thank Dr. Martin McKeown for data sharing and orienting me to the clinical side of neuroscience research. Moreover, I would like to thank Dr. Ghassan Hamarneh for his input on my research from a computer scientist’s perspective. In addition, I would like to thank Dr. Bertrand Thirion and his colleagues, Dr. Gael Varoquaux and Dr. Jean-Baptiste Poline, for hosting me at INRIA and sharing data. Furthermore, I would like to thank my friends inside and outside of BiSICL for their support. I would also like to thank NSERC, MSFHR, Walter C. Sumner Association, SPIE, Government of France, and UBC for their generous funding. Last but not least, I would like to thank my family for their encouragement.
Dedication

To my family
1 Introduction

Functional magnetic resonance imaging (fMRI) is currently one of the most widely-used imaging modality for studying human brain function. However, the low signal-to-noise ratio (SNR) coupled with the small sample size presents major challenges to fMRI analysis. Standard analysis techniques are largely univariate in nature, which neglect important information encoded by the interactions between voxels. Also, the high inter-subject variability both anatomically and functionally renders establishment of subject correspondence difficult, which complicates group analysis. In this thesis, we propose new prior-informed approaches for dealing with the key challenges in fMRI analysis. Our proposed methods enable an elaborate exploration of the functional organization of the human brain from detecting localized activation effects to deciphering mental representations encoded in whole-brain activity patterns. In this chapter, we first discuss the impact of neurological diseases, which serves as the main motivation for our work. We then provide an overview of fMRI and the key research questions and challenges associated with this imaging modality. The limitations of the existing analysis methods are then discussed, concluding with a summary of the thesis contributions.

1.1 Problem Statement and Motivation

Based on the Global Burden of Disease study conducted by the World Health Organization (WHO) in collaboration with the World Bank and the Harvard School of Public Health, neurological diseases rank as the number one disease burden in the world [1]. In fact, over 1,000 neurological diseases, including Alzheimer’s disease (AD), Parkinson’s disease (PD), Multiple Sclerosis (MS), Huntington’s disease, Stroke, Migraine, and Epilepsy among others, are currently affecting the world population [2]. In Canada, neurological diseases attribute to an estimated annual cost of $8.8 billion with $2.3 billion devoted to hospital care, physician care, and drug expenditures [3]. These costs represent a major financial burden on the country. Also, functional impairments associated with these neurodiseases greatly hinder the
daily activities of those affected and place a major toll on their family, both physically and emotionally. For instance, AD is known to induce memory loss, confusion, aggressiveness, and language breakdown among other symptoms. These debilitating effects may result in unintended inappropriate behaviors in social situations or one getting lost even in his or her own home. AD can thus cause great inconvenience, and at times, danger for the patients and those around them. Other neurodisease-induced symptoms, such as hand tremors and rigidity arising from PD, render tasks as simple as eating difficult for the patients. Worse yet, many of the neurological diseases are incurable and tend to worsen over time. Thus, with the life expectancy of the general population increasing, neurological diseases are predicted to become an even greater burden to healthcare systems worldwide.

Among the common neurological diseases, AD and PD are by far the most prevalent [4, 5]. Diagnosis of these neurodegenerative diseases is typically based on clinical symptoms. For instance, AD is often diagnosed using the Alzheimer’s Criteria set by the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s Disease and Related Disorders Association, where clinical tests are performed to assess impairments of cognitive functions, such as memory and attention [6]. Similarly, PD is typically diagnosed using the Unified Parkinson’s Disease Rating Scale (UPDRS), where mentation, behaviour, mood, activities of daily living, and motor changes are examined [7]. However, clinical symptoms are easily confused with effects of normal ageing and do not appear until years after disease onset [8]. Also, current therapies help reduce symptom severity, but typically do not modify disease progression. Thus, objective biomarkers are needed to facilitate earlier disease detection, where more effective treatments can be designed to slow disease progression or even prevent disease onset [8, 9].

In recent years, neuroimaging has greatly contributed to the characterization and understanding of neurological diseases. The application of MRI, for instance, has unveiled abnormal anatomical changes in the brain, such as hippocampal atrophy in AD patients [10] and brain volume reduction in MS patients [11]. Other disease-induced anatomical changes, such as degeneration of white matter fiber tracks in AD patients [12], have also been observed using diffusion MRI. In addition to anatomical changes, many neurological
diseases are associated with abnormal functional changes. For example, brain studies with electroencephalography (EEG) and magnetoencephalography (MEG) have shown impaired neural synchrony between brain regions in AD and PD patients [13]. Positron emission tomography (PET) and single photon emission computed tomography (SPECT) have also been used to quantify the extent of striatal dopaminergic degeneration in PD [14] and the reduction of cerebral metabolic rate for glucose in AD [15]. Furthermore, the use of fMRI has detected decreased functional synchrony in the hippocampus of AD patients [16], compensatory changes in brain activity in MS patients [17], and functional reorganization in stroke patients during recovery [18], among other clinical findings.

In this thesis, we focused on studying the functional aspects of the human brain with the ultimate goal of identifying reliable biomarkers for various neurological diseases. The most commonly-used imaging modalities for such purposes include EEG, MEG, PET, SPECT, and fMRI. Since fMRI provides much higher spatial resolution than EEG and MEG without the need for injecting radioactive tracers as in PET and SPECT, we have chosen to concentrate on fMRI throughout this work.

1.2 Functional Magnetic Resonance Imaging

Functional MRI provides a non-invasive means for studying brain activity based on changes in blood flow, blood volume, and blood oxygenation level. Among the techniques for creating image contrast, blood oxygen level dependent (BOLD) contrast is by far the most widely used [19]. The assumption is that neural activity directly affects the local blood flow near the active brain areas through a process called, “hemodynamic response.” Specifically, neuronal firing consumes oxygen, which increases local magnetic inhomogeneity, and hence decreases the BOLD signals near the brain areas involved. When an inflow of oxygenated blood rushes into these brain areas in response to the oxygen demand, the relative decrease in de-oxygenated blood reduces the local magnetic inhomogeneity, which increases the BOLD signals. In a typical fMRI experiment, a subject lying in the MR scanner performs a certain task as brain volumes are acquired at regular time intervals, as shown in Figure 1.1.
Figure 1.1 Typical fMRI block design experiment. A subject placed in the MR scanner performs a certain task, e.g. finger tapping, as BOLD volumes are acquired at regular time intervals. Alternating between rest and task generates the image contrast required for inferring brain activity. Pre-processing, such as motion correction to account for the subject’s head movements and high-pass filtering to remove low frequency scanner drifts, is then performed. The pre-processed voxel time courses are compared with an expected response to generate activation statistics maps, which are thresholded to localize the activated brain voxels.

Figure 1.2 Sample raw intensity time course of an activated voxel. Blue = voxel time course. Red = expected response.
The contrast in BOLD signals between the baseline and stimulus conditions is then used to infer brain activity. However, the inherently low SNR [20] (typically between 0.2 to 0.5 [21, 22]) and the complex noise structure of BOLD signals [23] greatly complicate analysis. Shown in Figure 1.2 is a sample intensity time course of an active brain voxel, plotted against the expected response (a boxcar function convolved with the canonical hemodynamic response function (HRF) [24]). As evident from Figure 1.2, the raw BOLD signals are highly contaminated by noise, and must be preprocessed prior to analysis [23].

Standard preprocessing steps include slice timing correction [25], head motion correction [26], and temporal autocorrelation correction [27, 28]. Spatial smoothing [29] and spatial normalization [26] may be optionally performed depending on the type of analysis (i.e. intra-subject or group, Section 1.3). Slice timing correction accounts for how brain slices are not acquired simultaneously, and is corrected by shifting the phase spectrums of the voxel time courses, while retaining their original magnitude spectrums [25]. Head motions are corrected by rigidly registering all brain volumes to the first volume assuming the shape of the brain stays fixed throughout the fMRI experiment [26]. Temporal autocorrelations mainly arise from scanner drifts, cardiac noise, respiratory noise, and head motions [28]. Since the noise component of the voxel time courses is often assumed to be identical independently distributed (i.i.d.), pre-whitening is typically performed to ensure correct statistical inference [28]. Specifically, temporal autocorrelations can be decomposed into a short-range and a long-range component [28]. The long-range temporal correlations are removed by high-pass filtering, while the short-range autocorrelations are accounted for using autoregressive (AR) models [28]. Since standard pre-processing techniques are widely available [30], the focus of this thesis is instead on the analysis of the pre-processed fMRI data.

The key research questions addressed by fMRI revolve around the two fundamental organization principles to which the brain adheres, namely functional specialization and functional integration [30]. Functional specialization refers to how specific brain regions are specialized for certain functions. This brain property is typically studied using fMRI by examining which areas of the brain are activated during various tasks [31] as well as by characterizing the activity patterns within brain regions of interest (ROIs) [32, 33]. The other
organizational principle, functional integration, refers to how different cognitive processes are mediated through the interactions between brain regions [30]. These interactions are usually inferred from fMRI data based on signal correlations [34], commonly referred to as “functional connectivity.” Also, recent advances in pattern classification techniques have ignited major research interest in examining how different cognitive states are encoded in whole-brain activity patterns [35, 36]. In this thesis, we present our contributions within these four aspects of fMRI analysis: activation detection, ROI characterization, functional connectivity analysis, and brain decoding. These aspects play complimentary roles in providing a more complete picture of the functional organization of the brain. In particular, activation detection and ROI characterization offer information concerning the individual brain units (i.e. voxels) and modules (i.e. ROIs), while connectivity analysis and brain decoding provide a more system level perspective of how cognitive functions are encoded through the interactions between the various brain units and modules. State-of-the-art for each of these four aspects of fMRI analysis are summarized in Sections 1.3, 1.4, 1.5, and 1.6, respectively. For the interested readers, details regarding other active areas of fMRI research, such as characterizing functional integration in terms of “effective connectivity” and HRF estimation, can be found in [37] and [38].

1.3 Activation Detection

The goal of brain activation detection is to map brain regions to functions by exerting different stimuli and examining which areas of the brain respond to which stimuli. Such analysis may be performed either at the intra-subject level or at the group level. Intra-subject analysis is more often used in clinical applications, such as neurosurgical planning [39], whereas group analysis is typically for identifying functional characteristics that are generally common within a population.

1.3.1 Intra-subject Analysis

The standard approach for inferring intra-subject brain activation from fMRI data involves comparing each voxel’s intensity time course independently against an expected response
using a general linear model (GLM) [31] to estimate what is referred to as “activation effect” (Section 2.1). Statistics, such as t-values, pertaining to the probabilities (p-values) of observing the estimated effects given that no activation occurred are then computed and assembled into an activation statistics map, referred to as a t-map. Voxels with t-values above a user-defined threshold (conventionally set to a t-value corresponding to an experimental-wise p-value of 0.05) are declared as activated. To control the number of voxels falsely declared as active, one must account for multiple comparisons during threshold selection [29]. Since the voxels are certainly correlated due to factors, such as physiological noise, interpolation during motion correction, and the functional integration property of the brain in general [30], correcting for multiple comparisons using the traditional Bonferroni correction technique, which assumes independence between voxels, will result in an overly stringent threshold [40]. To account for correlations between spatially neighboring voxels, Worsley et al. proposed a thresholding technique based on Gaussian random field (GRF) theory [29]. Deriving this GRF-based threshold requires estimating the smoothness of the activation statistics maps, which is non-trivial. The standard way of dealing with this issue is to spatially smooth the BOLD volumes using a fixed Gaussian kernel and assume the smoothness of the activation statistics maps is equal to the kernel width [30]. Spatial smoothing also helps increase the SNR based on the matched filter theorem [30], provided that the Gaussian kernel matches the shape of the active spatial clusters. However, the shape of the active clusters is likely to vary across the brain. To remedy this limitation, scale-space approaches have been proposed [41, 42], where kernels of varying widths are applied with the threshold adjusted for the additional scale dimension [41].

The main drawback of GRF thresholding is the need for spatial smoothing, which blurs the details of the activation statistics maps [41]. The use of permutation tests [40] has been proposed to mitigate this limitation at the expense of longer computational time. Alternatively, one may model the activation statistics maps using mixture models (MM) [43-45] (Section 2.2), which assumes the activation statistics of the active and non-active voxels are drawn from two distinct probability distributions. Under this setting, activated voxels are delineated by comparing the posterior probabilities of a voxel being active or non-active given data, which mitigates the need for multiple comparisons correction [44]. However,
mixture models alone do not account for spatial correlations. To mitigate this limitation, Woolrich et al. proposed a spatial mixture model based on continuous Markov random field (MRF) [46, 47]. Discrete MRF approaches have also been proposed [48-50], which convert activation detection into a class labeling problem, i.e. active or non-active, with discrepancies in labels between neighboring voxels penalized (Section 2.3). Spatial continuity is thereby encouraged, which helps suppress false declarations of isolated voxels as being activated.

The methods described thus far model local voxel interactions at the inference stage after the activation statistics maps are computed. The alternative approach would be to model these local interactions during the activation effect estimation stage. For instance, Ruttimann et al. proposed fitting spatial wavelets to BOLD volumes to account for local spatial correlations [51]. In their approach, spatial wavelet transform is separately applied to each BOLD volume in generating wavelet coefficient time courses. GLM is then performed on these time courses to compute activation statistics. The statistically significant wavelet coefficients are then identified by thresholding the activation statistics maps with e.g. Bonferroni or false discovery rate (FDR) correction [52]. The advantage of performing statistical inference in the wavelet domain is that the decorrelation property of wavelet transform renders the independence assumption in Bonferroni and FDR correction more appropriate [53]. However, the optimal way for transforming the results back to the spatial domain is not obvious. If the inverse wavelet transform is naively applied to the activation statistics of the significant wavelet coefficients to reconstruct an activation statistics map in the spatial domain, most voxels will be assigned non-zero activation statistics [53], thus thresholding will again be required. Therefore, Van De Ville et al. proposed using wavelet transform as a denoising step and performing statistical testing in the spatial domain but with the effect of the wavelet processing taken into account during threshold selection [54].

In addition to the wavelet approach, substantial research efforts have been placed on using Bayesian approaches [46, 47, 55-59] to model local voxel interactions through a spatial prior. For instance, Penny et al. proposed using a spatial Laplacian prior on the activation effects to encourage spatial smoothness [55]. Extending Penny et al.’s work, Harrison et al. employed a diffusion-based prior that accounts for how the spatial correlation structure is non-stationary
across the brain [56]. Other approaches include using spatiotemporal autoregressive models to account for correlations in the noise [58]. Most of these models require deploying approximate inference or computationally expensive sampling methods in estimating the activation effects. Also, only local spatial correlations are modeled with longer range voxel interactions completely ignored. Such long-range interactions may be implicative of activation, and hence beneficial in improving activation detection.

### 1.3.2 Group Analysis

In standard fMRI group analysis [31], brain images of all subjects are first warped onto a common brain template [60] to create voxel correspondence across subjects. This procedure is referred to as “spatial normalization.” Activation effects at each voxel location are then statistically compared across subjects, e.g. using a t-test, to generate a group activation statistics map. The resulting group map is subsequently thresholded to isolate the commonly activated voxels. The implicit assumption is that a one-to-one correspondence exists between the active voxels of the subjects, and that all voxels are perfectly aligned in the template space after spatial normalization. However, the large inter-subject variability renders whole-brain spatial normalization prone to mis-registration errors [61]. High dimensional registration techniques, such as Hierarchical Attribute Matching Mechanism for Elastic Registration (HAMMER) [62], have been shown to improve alignment over commonly-used low dimensional techniques [63, 64] with more precise, localized activation foci detected [65]. However, even if perfect anatomical alignment can be achieved, whether a one-to-one functional correspondence exists between voxels across subject is debatable [65, 66].

To account for inter-subject variability in the location of the activated voxels, the conventional way is to spatially smooth the BOLD volumes to increase the functional overlaps across subjects, which also serves the purpose of deriving a GRF-based threshold [29]. However, spatial smoothing tends to smear the activations statistics maps. To mitigate this drawback, a couple of methods have been proposed to deal with this challenging problem of inter-subject functional misalignment. Thirion et al. proposed using a “structural” strategy to draw correspondence at the cluster level [67, 68]. Specifically, the activation statistics map of each subject is first thresholded to segment out the active voxel clusters.
Associated with each cluster is a local maximum, whose location is used to draw probabilistic correspondences between subjects, as estimated with belief propagation [69]. Replicator dynamics (RD) [70] is then applied to the “belief matrix” to segregate clusters that are commonly present in the subjects. Another approach, proposed by Keller et al., is to introduce a “spatial jitter” variable into GLM and marginalize out that variable to account for the various ways that the activation maps may be misaligned across subjects [71].

In addition to the standard approach of drawing group inference by thresholding a group activation statistics map, a number of hybrid methods, where group information is integrated into intra-subject activation detection, have been explored. Such “group-wise” approach facilitates modeling of subject differences while exploiting group information to better disambiguate signal from noise. For instance, Calhoun et al. [72] proposed “group independent component analysis” (ICA), where voxel time courses are concatenated across subjects and decomposed into independent spatial components and associating time courses of the respective spatial components. A “back reconstruction” procedure is then performed to extract subject-specific spatial component maps. The optimal way to delineate active voxels from these intra-subject spatial component maps, however, is not trivial. One approach is to apply a threshold to z-scores estimated from the spatial component maps [73]. However, these z-scores do not have a clear statistical interpretation, in contrast to activation statistics generated from GLM [73]. Therefore, the spatial component maps are more often used for generating group maps instead [72]. Recently, Bathula et al. [74] proposed a leave-one-out strategy, where all subjects except one are analyzed to compute a group map, which is then used as a prior to segment the active voxels of the left-out subject. However, mis-registration errors arising from whole-brain spatial normalization, as required for generating group maps, may limit the accuracy of this strategy. To relax the one-to-one voxel correspondence criterion, Kim et al. [75] proposed using a hierarchical Dirichlet process (HDP) model with “random effects” to extract Gaussian-shaped functional clusters, as opposed to detecting active voxels. The authors showed that adding “random effects” to the HDP model permits slight inter-subject variations in the location of the detected clusters without losing cluster correspondence between subjects. However, considering the possible shapes that a functional cluster may take, the assumption that all functional clusters are Gaussian-shaped can be
limiting. Removal of this assumption, along with the one-to-one voxel correspondence criterion, would thus be desirable.

### 1.4 ROI Characterization

ROI analysis is typically employed for testing specific hypotheses. For instance, one might be interested in studying how Parkinson’s disease affects ROIs within the motor system in addition to causing cell death in the substantia nigra [76]. This analysis approach involves specifying ROIs for each subject and comparing some “properties” of the regional activation statistics across subjects to draw group inference. The main advantage of this ROI analysis approach is that an exact voxel correspondence across subjects is not required, hence no need for spatial normalization. The original information in the activation statistics maps, which is certainly altered during spatial normalization, is thus retained. The challenge is that defining the ROIs and specifying a metric to rigorously characterize and compare the regional activation statistics across subjects without merely amplifying inter-subject variability is non-trivial. ROIs are typically defined anatomically [32, 33, 77], but such ROI definition may incorporate non-activated voxels into the analysis, which could result in reduced sensitivity in detecting task-related effects, depending on the metric used.

By far, mean activation statistics, percentage of activated voxels, and variants of these metrics are most widely used in fMRI ROI analysis [32, 33]. For instance, Buck et al. used the mean activation statistics of different proportions of voxels within an ROI to examine the reproducibility of results for various cognitive tasks [33]. Allen and Courchesne used the percentage of activated voxels and mean percentage signal changes to study the activation differences between subjects with and without autism [78]. Similarly, Aizenstin et al. used the percentage of voxels with negative activation statistics values to compare the activation distributions of young and elderly subjects [79]. All these features are derived based on the amplitude of the activation statistics with the underlying assumption that the spatial distribution of the regional activation statistics is immaterial.
Characterizing ROI activation statistics using amplitude-based features provides a simple means for subject comparisons, since these features are insensitive to variability in brain shape, brain size, subjects’ orientation, and location of the activated voxels. However, Haxby et al. in their seminal work [80] have shown that the spatial distribution of the BOLD signals also encode important cognitive information. Thus, instead of only looking at mean amplitude changes in activation statistics averaged over an ROI, exploring spatial distribution changes may also be beneficial [81-83].

1.5 Functional Connectivity Analysis

Cognitive functions are known to be manifested through functional integration in addition to function specialization [84, 85]. In fact, recent studies suggest that voxel interactions are present even at rest without any external stimulus [20, 86-92]. These interactions are typically characterized in terms of functional connectivity [34] and effective connectivity [37]. Functional connectivity refers to the statistical dependencies between brain regions, whereas effective connectivity refers to the causal interactions among regions. We focus on the exploratory detection of brain networks based on functional connectivity inferred from fMRI data, and refer readers to [93-95] for descriptions of methods, such as dynamic causal modeling and structural equation modeling, for analyzing effective connectivity.

1.5.1 Intra-subject Analysis

The objective of functional connectivity analysis is to infer brain networks based on BOLD signal correlations. Viewed from a graph-theoretic perspective, voxels correspond to graph nodes with edges connecting nodes that form networks. One may thus infer brain networks by either finding cliques of connected nodes or by finding presence/absence of edges.

Methods that search for connected nodes can be broadly divided into seed-based and exploratory multivariate approaches [96]. The seed-based approach infers functional connectivity by examining the correlations between a seed voxel or region and all other voxels in the brain using GLM. This approach is equivalent to the standard univariate
analysis for activation detection except the expected response is replaced with the intensity 
time course of the seed. Though this approach has shown promising results [86, 97], the 
detected networks are inherently restricted by the choice of seed [98]. For instance, if a seed 
is selected from a motor region, only motor networks will be detected with the rest of the 
brain left unexplored. Also, joint interactions between multiple voxels are neglected [99]. To 
alleviate these limitations, various multivariate techniques, such as cluster analysis and 
decomposition-based methods, have been explored [96]. Cluster analysis [100, 101] 
partitions voxels based on similarity between the intensity time courses, but results are often 
sensitive to initialization and the number of clusters chosen [96]. Decomposition techniques, 
like principal component analysis (PCA) [34] and ICA [73, 99], factorize the voxel time 
courses into spatial components and associating time courses of the respective spatial 
components based on statistical criteria, such as maximum variance and independence. 
Functional networks are then inferred based on the weights assigned to the voxels in the 
spatial component maps. Specifically, voxels jointly assigned “large” weights within the 
same spatial component are considered to belong to the same network. However, grouping 
voxels into networks using this strategy is complicated by how non-zero weights are assigned 
to majority of the voxels due to the diffused weighting property of PCA and ICA [102, 103]. 
Also, determining which spatial components pertain to true brain activity can be non-trivial 
for complex experiments [104]. To extract stimulus-invoked spatial components, Todd et al. 
proposed using constrained PCA, where GLM is first performed to isolate the task-related 
variance from the voxel time courses prior to applying PCA [105]. The limitation of this 
approach is that any misspecifications in the expected stimulus responses will result in 
removal of signals from which brain networks are inferred. Some brain networks might thus 
be mistakenly neglected.

The alternative approach for discovering brain networks is to look for the presence/absence 
of graph edges. Pairwise correlations between voxel time courses are typically used as 
estimates of the degree of voxel interactions. By computing the correlations between all 
possible pairs of voxels, one can generate an inverse covariance matrix, i.e. precision matrix, 
where elements with non-zero values imply the corresponding pairs of voxels are connected. 
However, since there are typically more voxels than time points, estimation of the precision
matrix by inverting the empirical covariance matrix will be unstable. Also, due to noise, most (if not all) elements of the estimated precision matrix will be assigned non-zero weights. A recently proposed strategy to deal with these problems is to enforce sparsity during precision estimation [106-108]. The main advantage is that it shrinks elements of the precision matrix that correspond to non-connected voxels to exactly zero, hence provides an unambiguous answer as to which voxels are disconnected. However, existing methods for sparse inverse covariance estimation are very computationally intensive and can only handle up to thousands of variables. Thus, these methods are typically used for region level analysis [108].

1.5.2 Group Analysis

The seed-based approach naturally extends to group analysis since the same machinery for standard group activation detection can be applied. However, the limitations of standard group activation analysis, such as voxel misalignment, are also propagated. Moreover, the inherent difficulties of the seed-based approach, such as seed selection, limit the possible group networks that may be discovered.

Cluster analysis learns the seeds (i.e. cluster centers) in a data-driven, but the optimal way for statistically inferring commonly recruited clusters across subjects is not obvious. One approach that avoids spatial normalization is to pool the voxel time courses across all subjects and apply clustering based on time course similarity. However, inter-subject variability in HRF (if analyzing task-based data) renders this approach prone to errors. In fact, in the case of resting state (RS) data where no external stimulus is present to guide the timing of the BOLD signals, this approach would be completely infeasible. To ease the HRF variability issue in multi-stimuli task-based experiments, Lashkari et al. proposed applying GLM to generate “activation profiles,” i.e. vectors of activation effect estimates associated with different stimuli, normalizing and pooling all subjects’ activation profiles into a single set, and clustering these profiles using a mixture model [109]. However, it is unclear whether and how this approach can be generalized to RS data.

The decomposition methods, such as PCA and ICA, also do not required seed selection. However, drawing a correspondence between the intra-subject spatial component maps for
group analysis can be challenging [72]. One way to mitigate this problem is to average the voxel time courses across subjects and apply the decomposition methods on the average time courses to estimate average group networks [110, 111]. Alternatively, in the case of PCA, one can compute the covariance matrix for each subject and applying PCA or RD (which we showed to be a solution to the non-negative sparse PCA optimization problem, Section 4.1) on the average group covariance matrix [112]. The main drawback of these averaging approaches is that the average group network may not reflect individual differences [113]. Also, these averaging approaches prohibit statistical group inference. To circumvent this limitation, group ICA has been proposed for extracting subject-specific spatial components to enable group inference [72]. Alternatively, Beckmann et al. proposed a technique called, “dual regression” for a similar purpose [114]. This technique involves concatenating the voxel time courses across subjects, applying ICA to the concatenated time course matrix to extract group spatial component maps, performing regression on each subject’s data with the group spatial components as regressors to extract subject-specific time courses, and applying regression again on the each subject’s data but with the subject-specific time courses as regressors to extract subject-specific spatial component maps. A limitation common to both methods is that all subjects are assumed to have similar spatial noise components, which is rather unlikely.

For the recent graph edge detection approaches [106-108], one way of drawing group inference is to apply a separate t-test to each element of the subjects’ covariance/precision matrices [115]. Alternatively, one may derive certain network properties based on graph theory to compare networks across subjects [116]. Since this graph edge detection approach is rather recent, the preferred methods for making group inference are yet to be established.

1.6 Brain Decoding

One of the aims of brain decoding is to exploit the collective information encoded by the signal patterns of the fMRI brain volumes to determine the experimental conditions to which these brain volumes belong. Under this classification framework, signal intensity of each
voxel is treated as a feature with brain volumes taken as samples [35, 36]. Alternatively, one may concatenate multiple brain volumes within a trial of the same experimental condition and treat each concatenated set of brain volumes as a sample [117]. Typical fMRI datasets consist of considerably more voxels (~tens of thousands) than brain volumes (~hundreds). Hence, direct application of standard classifiers, such as support vector machines (SVM) [117] and linear discriminant analysis (LDA) [118], with all brain voxels used as features will likely result in overfitting [35, 36]. To reduce the dimensionality of the feature space, a common strategy is to apply standard univariate analysis [31] to restrict the feature set to only those voxels displaying significant activation or discriminability [117]. However, this approach discards the collective discriminant information encoded by the voxel patterns, which may lead to suboptimal feature selection [119-122]. Alternatively, PCA may be applied to reduce feature dimension [123]. However, the new features created by taking a weighted sum of the original features do not allow for simple identification of relevant voxels, which hinders result interpretation [124]. More advanced methods, such as support vector machine with recursive feature selection (SVM-RFE) [125] and random forest [126], that enable multivariate feature selection have been explored, but tend to be more computationally demanding.

Recently, there is a surge of interest in applying sparsity-enforcing techniques (Section 5.1) to fMRI classification. Many of these techniques exploits the sparse property of the least absolute shrinkage and selection operator (LASSO) penalty [127] to shrink classifier weights associated with irrelevant features to exactly zero. Hence, feature selection is implicitly integrated as part of classifier learning. However, naively enforcing sparsity may result in spurious classifier weight patterns [119]. In particular, classifier weights derived from enforcing sparsity alone often deviate from how brain activity tends to distribute in localized spatial clusters [119]. To encourage spatial contiguity, spatial SVM [128] has been proposed. However, this technique assigns weights to all input voxels, which renders identification of relevant voxels nontrivial. Group LASSO has also been explored to jointly select spatially proximal voxels as well as voxels within ROIs [129]. However, group LASSO does not model the associations between features, hence may not assign similar weights to neighboring voxels to reflect how adjacent voxels tend to display similar level of brain activity.
activity [130]. Elastic net [119, 121], which encourages correlated voxels to be jointly selected, has also been employed, but suffers from similar limitations as Group LASSO. More recently, van Gerven et al. proposed a Bayesian formulation for incorporating a spatiotemporal prior, where the authors opted to model uncertainty by estimating the posterior probabilities of the classifier weights as opposed to obtaining sparse weightings through finding the maximum a posterior solution [131].
1.7 Thesis Contributions

We propose in this thesis novel directions for fMRI analysis, where informative priors are seamlessly integrated into activation detection, ROI characterization, functional connectivity analysis, and brain decoding. We demonstrate the power of such prior-informed approaches on carefully designed simulations as well as extensive sets of real fMRI data. We list below our specific contributions in advancing the current state-of-the-art. Our proposed methods enable an extensive analysis of fMRI data from studying the functional role of individual brain units and modules (i.e. voxels and ROIs) to investigating at the system level how cognitive functions are mediated through the interactions between brain modules and represented in whole-brain activity patterns. Details of these contributions are provided in Chapters 2, 3, 4, and 5, respectively. We note that we have kept the original mathematical notations used in the respective publications to ease cross-referencing. There might hence be some overlaps in notations. To disambiguate these overlaps, we have explicitly redefined the notations in each subsection.

1.7.1 Activation Detection

- We proposed incorporating functional connectivity priors estimated from task-based [P1, P2] as well as RS data [P6] into activation detection. Our proposed formulations model both local and long-range voxel interactions, which are completely neglected in existing techniques. We demonstrated on large sets of synthetic and real data that integrating connectivity priors increases detection sensitivity over the standard univariate approach.

- We proposed integrating a group prior into intra-subject activation detection [P3–P5]. Our proposed group-extended graph structure enables information coalescing between subjects without the need for a one-to-one voxel correspondence. We showed on large sets of synthetic and real data that incorporating a group prior enhances detection sensitivity over the standard univariate approach.
1.7.2 ROI Characterization

- We proposed inferring brain activation from the spatial modulations of regional activity patterns as measured using invariant spatial features [81, 82]. We validated the generality of this spatial analysis approach over large sets of real fMRI data [P7, P11]. We showed that brain activation can generally be inferred from spatial changes in ROI activity patterns. Our results thus demonstrate that signal amplitude is not the only means by which to infer brain activation.

- Applying our proposed invariant spatial features [81, 82] to specific datasets have resulted in other important findings. We showed that spatial normalization reduces discriminability of task-related activation differences in ROI analysis [P8]. Thus, spatial normalization, which is standardly performed in group activation analysis, may be distorting important spatial information encoded in the original activation statistics maps. We also demonstrated that a predominant effect of levodopa (L-dopa) on PD subjects is a spatial focusing of regional activation patterns, not signal amplitude changes [P9]. We further showed that increasing spatial extent of ROI recruitment is another strategy for PD subjects to handle increasing task difficulties [P10]. Other findings include earlier peak time in spatial response as compared to BOLD amplitude changes [P12] and that functional connectivity can be inferred from the spatial modulations of regional activity patterns [P14].

- We proposed measuring the skewness of the regional activity patterns [P13], in addition to spatial extent. Our results on real data showed that this skewness feature is more sensitive to directional changes than measuring spatial extent, and thus can play a complementary role to facilitate a more elaborate spatial analysis.
1.7.3 Functional Connectivity Analysis

- We showed that replicator dynamics is a solution of the challenging non-negative sparse PCA optimization problem [P15, P16].

- We proposed jointly integrating sparsity and group priors into intra-subject network detection. We demonstrated on synthetic and real data that enforcing sparsity simplifies network identification, while adding a group prior effectively handles the large inter-subject variability often seen in fMRI studies.

1.7.4 Brain Decoding

- We proposed a simple regularization technique for incorporating domain-specific knowledge into a wide range of sparse linear models [P17–P20].

- We validated on real fMRI data and showed how our prior-informed sparse classifiers outperform standard classifiers, such as SVM and LDA, as well as a number of existing sparse classifiers both in terms of prediction accuracy and result interpretability.
2 Prior-informed Activation Detection

The traditional way for studying functional specialization through fMRI is by invoking different stimuli and examining which areas of the brain are responding. To localize the activated brain areas, the standard approach is to analyze each voxel in isolation using GLM (Section 2.1). The activated voxels are then delineated by thresholding the resulting activation statistics maps based on GRF theory. This univariate approach has the drawbacks of neglecting voxel interactions and requiring multiple comparisons correction. One way of mitigating multiple comparisons correction is to model the activation statistics maps using Gaussian mixture models (GMM) (Section 2.2), but GMM alone does not account for voxel interactions. One may integrate GMM into an MRF framework to model local spatial correlations (Section 2.3), but longer range interactions will still be neglected. We thus propose incorporating a functional connectivity prior into GMM using a random walker (RW) formulation [P1] to model the interactions between both adjacent and distant voxels (Section 2.4). Also, since intra-subject activation statistics maps are typically very noisy due to the low SNR inherent in BOLD signals, we propose a group MRF (GMRF) [P3, P5] that coalesces information across subjects to handle noise (Section 2.5). Validations of the proposed methods were performed on both synthetic (Section 2.6) and real data (Section 2.7). In addition, recent studies suggest that an intrinsic relationship might exist between brain activity at rest and during task, which may serve as another source of information for activation detection. We thus propose a RS-informed activation model for integrating RS-connectivity in estimating activation effects [P6] (Section 2.8). This model was also tested on both synthetic (Section 2.9) and real data (Section 2.10). Our other related contributions that are not discussed in this chapter include a new sparse multivariate method that we refer to as “adaptive non-negative sparse PCA” (ANSPCA) [P2] for incorporating functional connectivity into activation detection with local spatial correlations explicitly modeled, and a group RW (GRW) [P4] that adaptively controls the amount of intra- and inter-subject spatial regularization based on similarities in the frequency spectrums of the BOLD signals within and between subjects.
2.1 Univariate Activation Detection with GLM

Let $y_j$ be an $n \times 1$ vector corresponding to the intensity time course of voxel $j$ of a subject, where $n$ is the number of time points. The activation effect, $\beta_j$, and its associating t-value, $t_j$, can be estimated using a GLM:

$$y_j = X\beta_j + \omega_j,$$

(2.1)

$$t_j = \beta_j ./ se(\beta_j),$$

(2.2)

where $\beta_j$ is an $m \times 1$ vector, $m$ is the number of experimental conditions, $\omega_j$ is assumed to be white Gaussian noise after preprocessing (Section 1.2), ./ stands for element-wise division, and $se(\beta_j)$ is an $m \times 1$ vector containing the standard error of each element of $\beta_j$ [132]. $X$ is an $n \times m$ design matrix with boxcar functions time-locked to task stimuli convolved with the canonical HRF as regressors along the columns [31]. A pictorial depiction of GLM is shown in Figure 2.1.

![Figure 2.1 Pictorial depiction of GLM.](image)

To isolate the activated voxels for each subject, the standard approach is to apply a GRF-threshold corresponding to a typical experiment-wise p-value of 0.05 to the t-map [29].
For detecting group activation, the standard approach involves warping the β-maps onto a common brain template and applying a t-test to the subjects’ βj at each voxel location to generate a group t-map. GRF thresholding is then applied for isolating the commonly activated voxels. For both intra-subject and group activation detection, spatial smoothing of the BOLD volumes is required for deriving the GRF-based threshold, but in the case of group analysis, spatial smoothing additionally serves the purpose of increasing the functional overlaps across subjects. However, spatial smoothing hinders precise localization of the functional boundaries of the active clusters.

2.2 Univariate Activation Detection with GMM

Another approach for isolating the activated voxels is to model the activation statistics maps using GMMs, which alleviates the need for multiple comparisons correction [44]. In this model, the t-values, tj’s, of an activation statistics map (intra-subject or group) are assumed to be generated from a mixture of K Gaussians with mixing coefficients πk, means μk, and variance σk2:

\[ t_j \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma_k^2), \]  

where K = 2 if classifying the voxels as either active (k = 1) or non-active (k = 2). The maximum likelihood (ML) solution of πk, μk, and σk2 can be estimated e.g. using expectation maximization [69], but the singularity problem inherent in the ML solution, i.e. μk collapsing onto one of the tj’s with σk2 approaching 0, may result in a tj being assigned infinite likelihood. One way to deal with this singularity problem is to add priors to the parameters:

\[ \mu_k \sim N(\eta, \tau^2), \quad \sigma_k^2 \sim IG(a, b), \quad \pi \sim Dir(\alpha), \]  

where IG(a,b) and Dir(α) denote inverse Gamma and Dirichlet distributions, respectively. A graphical summary of the resulting model is shown in Figure 2.2.
Figure 2.2  Graphical depiction of GMM. \( N_v \) is the number of voxels. \( z \) is a \( N_v \times K \) matrix with each element being an indicator variable set to 1 if voxel \( j \) is estimated to be in state \( k \), and 0 otherwise.

In addition to mitigating the singularity problem of the ML solution, adding priors allows us to encode our knowledge into the model. Specifically, we know that \( t_j \) of non-active voxels should theoretically be 0 and the \( t \)-threshold for active voxels is typically set between 3 and 4, based on GRF theory [29]. We can encode this prior knowledge on \( t \)-values of active and non-active voxels through \( \eta \) with \( \tau^2 \) set to 1 to model uncertainty in \( \eta \). As for \( \sigma_k^2 \), an uninformative prior is typically used by setting both \( a \) and \( b \) to 0.5 [38], since little is known about this parameter. Finally, \( \alpha \) is set to \( 1/K \) assuming equal prior class probabilities.

To estimate the parameters, \( \pi, \mu_k \), and \( \sigma_k^2 \), we use Gibbs sampling [69] by drawing samples from the following posterior conditional distributions:

\[
p(\pi | z, \alpha) = \text{Dir} \left( \alpha_1 + \sum_{j=1}^{N_v} z_{j1}, \cdots, \alpha_K + \sum_{j=1}^{N_v} z_{jK} \right),
\]

(2.5)

\[
p(\mu_k | t, z) = N \left( \frac{\sum_{j=1}^{N_v} z_{jk} t j \tau^2 + \eta \sigma_k^2}{\tau^2 \sum_{j=1}^{N_v} z_{jk} + \sigma_k^2}, \frac{\tau^2 \sigma_k^2}{\tau^2 \sum_{j=1}^{N_v} z_{jk} + \sigma_k^2} \right),
\]

(2.6)

\[
p(\sigma_k^2 | t, z, a, b) = IG \left( a + 0.5 \sum_{j=1}^{N_v} z_{jk} b + 0.5 \sum_{j=1}^{N_v} z_{jk} (t_j - \mu_k)^2 \right),
\]

(2.7)
\[ p(z_{jk} = 1 | t_j, \mu_k, \sigma_k^2, \pi_k) \propto \pi_k N(t_j | \mu_k, \sigma_k^2), \]  

(2.8)

where \( N \) is the number of voxels and \( z \) is an \( N \times K \) matrix with the \((j,k)\)th element, \( z_{jk} \), being an indicator variable set to 1 if voxel \( j \) is estimated to be in state \( k \), and 0 otherwise. A voxel \( j \) is declared as activated if \( p(z_{j1} = 1 | t_j, \mu_k, \sigma_k^2, \pi_k) > p(z_{j2} = 1 | t_j, \mu_k, \sigma_k^2, \pi_k) \).

### 2.3 Modeling Local Spatial Correlations with MRF

To model local spatial correlations, one can integrate the GMM into an MRF framework. Specifically, let \( G = (V, E) \) be a graph, where \( V \) and \( E \) are sets of vertices (voxels) and edges, respectively. Also, let \( jp \in E \) be the edge between voxels \( j \) and \( p \). The activated voxels can be delineated by minimizing the MRF energy function:

\[ E(x) = \sum_{j \in V} \theta_j(x_j) + \lambda \sum_{jp \in E} \theta_{jp}(x_j, x_p), \]  

(2.9)

where \( x_j \in \{1, \ldots, K\} \) is the label assigned to voxel \( j \), \( K \) is set to 2 to label voxels as either active or non-active, \( \theta_j(x) \) is a unary potential for modeling the activation statistics, and \( \theta_{jp}(x_j, x_p) \) is a pairwise potential for encouraging spatial smoothness as controlled by adjusting \( \lambda \). We define the unary potential as:

\[ \theta_j(x_j = k) = 1 - p(x_j = k | t_j), \]  

(2.10)

where \( t_j \) is the t-value of voxel \( j \) and \( p(x_j = k | t_j) \) is the probability of \( x_j \) being assigned label \( k \) given \( t_j \) as estimated using GMM (Section 2.2). Thus, if voxel \( j \) has a low probability of being activated, we penalize labeling \( x_j \) as active, and vice versa. \( k = 1 \) or 2 corresponding to active or non-active. As for the pairwise potential, we employ the widely-used Pott’s model [133]:

\[ \theta_{jp}(x_j, x_p) = 1 - \delta(x_j - x_p), \]  

(2.11)
where $\delta(y) = 1$ if $y = 0$ and 0 otherwise. Edges are added between every voxel $j$ and its 6-connected spatial neighbors to encourage adjacent voxels to have similar labels. $\lambda$ is set to 1 to weight the unary and pairwise potentials equally. We note that this MRF approach is applicable for both intra-subject and group analysis, since the algorithm is indifferent to the type of input t-maps. The globally optimal label configuration can be determined using graph cuts [134] with the activated voxels delineated based on the voxel labels.

### 2.4 Modeling Functional Connectivity with RW

In addition to modeling local spatial correlations, it may also be beneficial to model longer range interactions. We thus propose using RW to integrate voxel correlations (both local and long range) and activation statistics in estimating probabilistic activation maps [P1]. Under the original RW framework [135], each voxel corresponds to a graph vertex with edges between voxels weighted by their correlations to bias the paths for which a random walker may transverse. Voxels are labeled (i.e. active or non-active) based on the probability that a random walker starting at each voxel location will first reach a pre-labeled seed. This framework, however, not only requires specifying seed voxels but also does not model activation statistics. Therefore, we adopt an augmented RW formulation [136] that facilitates incorporation of activation statistics through label priors. This formulation is equivalent to adding an artificial seed vertex for each label and connecting these seeds to every vertex in the original graph with label priors being the edge weights [136]. Hence, this augmented RW formulation is analogous to MRF except the optimization is performed over real-valued probabilities instead of binary labels. Also, this augmented RW formulation has an exact, unique closed-form solution for computing the posterior label probabilities with global optimality guaranteed for an arbitrary number of labels [136]. The posterior probabilities of the voxels belonging to label class $s$, $x^s$, can be estimated by solving:

$$\left( L + \sum_{k=1}^{K} A^k \right) x^s = \lambda^s,$$

(2.12)
where $x^s$ is an $d \times 1$ vector, $d$ is the number of voxels, $K$ is the number of labels set to 2, and $s = 1$ or 2 corresponding to active or non-active, respectively. $\Lambda^s$ is a diagonal matrix containing the $d \times 1$ label prior vector, $\lambda^s$, which we set as the probabilities of the voxels belonging to class $s$ given their associating t-values as estimated using GMM (Section 2.2). $L$ is an $d \times d$ weighted graph Laplacian matrix:

$$L_{ij} = \begin{cases} -w_{ij}, & ij \in E \\ \sum_i w_{ij}, & i = j \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (2.13)$$

where $w_{ij}$ is the correlation between the intensity time courses of voxels $i$ and $j$, $ij$ denotes a graph edge between voxels $i$ and $j$, and $E$ denote the sets of edges. Graph edges are added between every pair of voxels with correlation greater than a conventional threshold of 0.4 [137]. This threshold can be empirically derived based on the assumption that time courses of non-active voxels after preprocessing are distributed as white Gaussian noise. Specifically, the correlation values between the non-active voxels likely lie between -0.4 and 0.4, as shown in Figure 2.3, where a histogram of pairwise correlations between 1,000 random Gaussian noise time courses is plotted. As evident from Figure 2.3, setting a threshold of 0.4 will remove majority of the correlations due to white noise.

![Figure 2.3 Histogram of correlation values between random Gaussian noise time courses.](image)
To explicitly encourage spatial smoothness, which may be underrepresented in the estimated voxel correlation values due to noise, we add a small constant of 0.001 to the $w_{ij}$ between every voxel and its 6-connected spatial neighbors. We have tested on large synthetic datasets (Section 2.6) that adding a constant two orders of magnitude lower than the signal correlation values promotes spatial continuity while permitting functional connectivity to remain dominant over spatial smoothness in driving the activation detection process. Upon solving (2.12) for $x^s$, identification of the activated voxels is achieved by comparing $x^1$ against $x^2$, i.e. if element $j$ of $x^1$ is greater than that of $x^2$, then voxel $j$ is declared as active. We note RW naturally permits intra-subject analysis, but not group analysis since delineating a group activation statistics map is complicated by the need for a group $L$. One possible way of generating a group $L$ is to average the subjects’ correlation matrices and then apply (2.13).

### 2.5 Modeling Group Information with GMRF

The methods described thus far in this chapter, i.e. GLM, GMM, MRF, and RW, all operate on a single activation statistics map at a time in detecting the activated voxels. In the case of intra-subject analysis, the derived activation statistics maps might be very noisy even with functional connectivity integrated due to the low SNR inherent in BOLD signals. In the case of group analysis, averaging over subjects should reduce noise, but the considerable amount of functional inter-subject variability [138] renders the one-to-one voxel correspondence assumption questionable. Therefore, we propose a group MRF that does not require a one-to-one voxel correspondence to integrate group information into intra-subject activation detection [P3, P5]. Modeling each subject’s activation statistics map as an MRF, we extend the neighborhood system to other subjects so that information can be coalesced across subjects. The intuition behind this approach is that, although the existence of a one-to-one functional correspondence between voxels is very unlikely, true brain activation should presumably appear in similar proximal locations across subjects [138], whereas false positives are more likely to be randomly scattered across the brain with little cross-subject consistency. Thus, regularizing inter-subject neighbors reinforces brain areas that are consistently recruited across subjects, while suppressing false positives. Hence, even if intra-
subject neighbors are deemed unreliable due to noise, additional neighbors from other subjects are available to better approximate the state of each voxel. Since only voxels that are spatially *proximal* to each other are hypothesized to be in similar state, our approach alleviates the need for finding an exact one-to-one voxel correspondence across subjects. Instead, a relaxed condition of requiring only the brain structures of the subjects to be approximately aligned is in place.

### 2.5.1 GMRF Formulation

Let $G_i = (V_i, E_i)$ be a graph where $E_i = e_i \cup \epsilon_{-i}$. $V_i$ and $e_i$ are the sets of graph vertices (voxels) and edges of subject $i$, respectively, and $\epsilon_{-i}$ is the set of edges between subject $i$ and all other subjects. Also, let $j \in e_i$ denotes the intra-subject edge between voxels $j$ and $p$ and $jq \in \epsilon_{-i}$ be the inter-subject edge between voxels $j$ and $q$. Our proposed group-extended graph structure for a single subject $i$ is shown in Figure 2.4.

![Figure 2.4 Proposed GMRF graph structure. Each subject’s activation statistics map is modeled as an MRF with its neighborhood system extended to other subjects. $N_s$ is the number of subjects in the group.](image)

If we connect the voxels of every subject in an analogous manner, yielding a “GMRF”, all subjects’ activation statistics maps can be simultaneously and collaboratively delineated by minimizing a single MRF energy function:
\begin{equation}
E(x) = \sum_{i=1}^{N_s} \sum_{j \in V_i} \theta_j(x_j) + \lambda \left( \sum_{jp \in E_i} \theta_{jp}(x_j, x_p) + \gamma \sum_{jq \in E_{i-h}} \theta_{jq}(x_j, x_q) \right),
\end{equation}

where $x_j \in \{1, \ldots, K\}$ is the label assigned to voxel $j$, $K$ is set to 2 to label voxels as active or non-active, $N_s$ is the number of subjects, $\theta_j(x_j)$ denotes a unary potential, and $\theta_{jp}(x_j, x_p)$ and $\theta_{jq}(x_j, x_q)$ are intra- and inter-subject pairwise potentials. $\lambda$ is used to adjust the contribution of the pairwise potentials while $\gamma$ balances the contribution from the intra- and inter-subject pairwise potentials. We define the unary potential exactly the same as (2.10). As for the intra-subject and inter-subject pairwise potentials, we employ the Pott’s model [133]:

\begin{equation}
\theta_{jm}(x_j, x_m) = 1 - \delta(x_j - x_m),
\end{equation}

where $\delta(y) = 1$ if $y = 0$ and 0 otherwise, and $m = p$ or $q$. Edges are added such that they link every voxel of subject $i$ to its 6-connected intra-subject spatial neighbors as well as its $c$ closest inter-subject spatial neighbors for every subject pair $(i, h)$, $i \neq h$. To define inter-subject neighbors, brain structures of all subjects are first aligned (Section 2.5.2). The resulting MRF energy (2.14) can be minimized using graph cuts [134] to find the globally optimal labeling for simultaneous delineation of all subjects’ activation statistics maps.

**2.5.2 Implementation Details**

To properly define inter-subject spatial neighbors, we need to first align the subjects’ brain structures. The conventional strategy is to perform whole-brain spatial normalization, but the large anatomical inter-subject variability renders accurate whole-brain registration difficult [61], especially for diseased subjects. Another approach is to extract anatomical ROIs and perform alignment at the region level for which highly accurate techniques are available [139]. This alignment approach has several advantages over whole-brain registration. First, no brain structures will be mistakenly taken as part of another structure. Second, the improvement in alignment has shown to enhance activation localization [140]. We have also shown that this region level alignment approach better preserves the spatial information in
the regional activation statistics patterns than whole-brain spatial normalization, as discussed in Section 3.2. Moreover, computational cost is substantially reduced.

From an implementation perspective, GMRF essentially creates a “mega-subject” by combining the voxels of all subjects into a single set with edges added between intra- and inter-subject spatial neighbors. We set $\lambda$ to $1/N_s$ to scale the cumulative pairwise potential over subjects to a level similar to that of a single subject, which ensures that the pairwise potentials will not dominate over the unary potential. The choice of $\gamma$ depends on the degree of inter-subject variability, which is unknown. We thus set $\gamma$ to 0.5 to weight subject $i$’s own information more than that of other subjects. As for the parameter $c$, we set it to 3 in agreement with how a one-to-one voxel correspondence unlikely exists between subjects.

### 2.6 Synthetic Task Data Experiments

#### 2.6.1 Data Generation

We generated 200 synthetic datasets, each consisting of ten subjects. The activation pattern comprised three clusters (Figure 2.5). The signal intensity of the active voxels (marked with circles in Figure 2.5) was set to decrease exponentially as a function of distance from the respective activation centroid of each cluster. The smaller cluster was added to test whether the compared methods can detect spatially isolated active voxels that are consistently present across subjects. The synthetic time courses of the active voxels were generated by convolving a box-car function, having the same stimulus timing as in our real data experiment (Section 2.7.1), with a canonical HRF and adding Gaussian noise. We also added low frequency temporally-correlated noise to emulate confounds, such as scanner drifts, aliased artifacts from cardiac and respiratory cycles, and mind wandering. Furthermore, we randomly varied the locations of the two larger clusters across subjects to introduce inter-subject variability in location of the active regions. The maximum cluster misalignment was set to two voxels (~30% of the cluster width) in both the vertical and horizontal directions. Two SNR levels were tested: 0.2 and 0.5, which are typical bounds for fMRI studies [21, 22]. We defined SNR as the standard deviation of the signal over that of noise as in [141]. We
note that only voxels near the cluster centers attained SNRs close to the maximum value of 0.2 or 0.5. SNR for voxels near the cluster borders are well below the maximum SNR values. 100 synthetic datasets were generated at each maximum SNR level.

### 2.6.2 Results and Discussion

We applied GLM, GMM, MRF, RW, and GMRF to the synthetic datasets to compare the effects of different priors on intra-subject activation detection. The qualitative results of an exemplar synthetic dataset (three arbitrary subjects included) at an SNR of 0.35 (i.e. midpoint between the typical SNR bounds of 0.2 and 0.5 [21, 22]) are shown in Figure 2.5. GLM missed many of the active voxels, whereas GMM detected more active voxels but also included many isolated false positives. Enforcing intra-subject spatial continuity using MRF reduced the amount of false positives, but failed to detect some of the active voxels near the cluster borders. The smallest cluster was also neglected due to lack of intra-subject active neighbors. Incorporating functional connectivity using RW led to more detections near the cluster borders than MRF since RW adapts the amount of spatial regularization based on the level of correlations between the voxels, i.e. active and non-active voxels are presumably uncorrelated hence less encouraged to be assigned the same class label. Employing RW also resulted in fewer isolated false positives than GMM, though an ample amount remained due to noise-induced false correlations between the active and non-active voxels, which were mistakenly taken as long-range interactions. Integrating group information using GMRF correctly detected most of the active voxels with much stricter control over false positives than GMM and RW. Also, compared to MRF, GMRF detected the smallest cluster and more active voxels at the cluster borders, thus demonstrating the benefit of adding inter-subject neighbors to better disambiguate the state of the voxels.

To quantify and compare the various methods’ performances, we computed the average true positive rate (TPR) and average false positive rate (FPR) over the 100 synthetic datasets for each SNR level. We also computed the average Dice similarity coefficient (DSC) [142] to provide a single, overall measure of performance:
\[ DSC = \frac{2TP}{2TP + FP + FN}, \]  

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. The quantitative results are summarized in Figure 2.6. At an SNR of 0.2, GLM achieved the lowest FPR at the expense of attaining the worst TPR. GMM obtained \~30\ times higher TPR than GLM, but a horrific FPR of 0.31. Incorporating functional connectivity into GMM using RW further increased TPR and mildly decreased FPR. Encouraging only spatial continuity using MRF reduced FPR by \~3\ times compared to RW, but resulted in \~2\ times lower TPR. Our results thus show that modeling functional connectivity using RW increases detection sensitivity but has less control over false positives compared to MRF due to noise-induced correlations. In contrast, integrating group information using GMRF decreased FPR by an order of magnitude compared to GMM, MRF, and RW, while maintaining a competitive TPR that was only \~1.2\ times lower than the highest TPR as obtained with RW, hence resulting in the best overall DSC among the compared methods. These results suggest that relying on the information within a single dataset of a subject might be insufficient to obtain satisfactory intra-subject activation detection at low SNR, even with spatial regularization or incorporation of a functional connectivity prior due to the lack of reliable intra-subject neighbors and noisy signal correlation estimates. At an SNR of 0.5, adequate signal appeared to be present for MRF and RW to be more effective, resulting in \~20\ times and \~4\ times decrease in FPR, respectively. Nevertheless, even at higher SNR where more reliable information was available within each subject, the addition of group information using GMRF still improved results over the compared methods with TPR increased to a level similar to that of RW and a FPR close to 0, hence achieving the highest DSC overall.
Figure 2.5 Synthetic data results. Results at an SNR of 0.35 for three exemplar subjects of a synthetic data set shown. Circle = ground truth activated voxels. Dots = detected activated voxels. Note: Data were spatially smoothed only for the case of GLM as required for deriving a GRF threshold.
Figure 2.6 TPR, FPR, and DSC at two typical SNR levels. GMRF obtained the highest DSC for both SNR levels, demonstrating a better balance between TPR and FPR than the other compared methods. Note: GLM achieved FPRs of approximately 0 at the expense of lower TPRs.

In addition to the presented test case, we also tested the methods above under a couple of other scenarios. First, we tested the case where location of the clusters were consistent across subjects but the placement of the activation centroids were varied and permitted to be anywhere within the clusters. This scenario emulates the situation where the active regions completely overlap between subjects but the location at which the BOLD signal concentrates varies. Hence, it will appear that there is little overlap between the active regions of the subjects. Results were similar to those described above. Details of which can be found in [P3]. Second, we tested the case where artificial activations are injected into real ROIs. The general trends of the results were again similar to those above. Details can be found in [P4, P5]. We note that some of the activation detection results presented in this thesis slightly differ from those in our original publications due to a couple of changes made. First, we now define SNR as the standard deviation of the signal over that of noise [141], as opposed to squared mean of the signal over variance of the noise. Second, in detecting activation with GLM, spatial smoothing is now applied to the voxel time courses instead of on the activation statistics maps. In [143], it was mentioned that spatially smoothing the activation statistics maps provides similar results as spatially smoothing the voxel time courses. However, in our synthetic experiments, we obtained higher sensitivity with the latter approach, and have thus adopted this latter approach in applying GLM to our synthetic and real data (Section 2.7).
2.7 Real Task Data Analysis

2.7.1 Materials

Functional MRI data were provided by Dr. Martin McKeown of the Pacific Parkinson Research Center. 10 PD patients off and on medication (4 men, 6 women, mean age 66 ± 8 years) and 10 age-matched healthy controls (3 men, 7 women, mean age 57.4 ± 14 years) were recruited. Each subject used their right hand to squeeze a bulb with sufficient pressure to maintain a bar shown on a screen within an undulating pathway. The pathway remained straight during baseline periods and became sinusoidal at a frequency of 0.25 Hz (slow), 0.5 Hz (medium) or 0.75 Hz (fast) during the time of stimulus. Each session lasted 260 s, alternating between baseline and stimulus of 20 s duration.

Functional MRI was performed on a Philips Gyroscan Inter 3.0 T scanner (Philips, Best, Netherlands) equipped with a head-coil. T2*-weighted images with BOLD contrast were acquired using an echo-planar (EPI) sequence with an echo time of 3.7 ms, a repetition time of 1985 ms, a flip angle of 90°, an in plane resolution of 128×128 pixels, and a pixel size of 1.9×1.9 mm. Each volume consisted of 36 axial slices of 3 mm thickness with a 1 mm gap. A 3D T1-weighted image consisting of 170 axial slices was further acquired to facilitate anatomical localization of activation.

Slice timing and motion correction were performed using Brain Voyager (Brain Innovation B.V.). Further motion correction was performed using motion corrected independent component analysis (MCICA) [144]. The voxel time courses were high-pass filtered to account for temporal drifts and temporally whitened using an AR(1) model [28]. No whole-brain spatial normalization or spatial smoothing was performed. Anatomical delineation of the ROIs, including the primary motor cortex (M1), supplementary motor cortex (SMA), prefrontal cortex, caudate, putamen, globus pallidus, thalamus, cerebellum, and anterior cingulate cortex, was performed by an expert based on anatomical landmarks and guided by a neurological atlas. The segmented ROIs were resliced at the fMRI resolution for extracting the preprocessed voxel time courses within each ROI.
2.7.2 Results and Discussion

For testing the various activation detection methods, we selected the left M1 as the region of interest, since this region is known to activate during right-hand motor movements (Figure 2.7). For applying GMRF, the left M1 of all subjects were non-rigidly aligned using a point set registration algorithm called, “coherent point drift” [139], which treats the point sets as Gaussian mixtures and registers the point sets by aligning the mixture distributions. Enhanced robustness to noise and outliers has been shown with this algorithm over standard techniques, such as iterative close point [139]. We note that since the subjects were required to perform motor squeezing during both baseline and task periods, the activation differences in the left M1 between these two conditions are expected to be subtle.

Figure 2.7 Homunculus. Hand area of M1 is circled in Homunculus diagram. Approximate corresponding area in the aligned left M1 is indicated. Homunculus courtesy of thebrain.mcgill.ca.

Results obtained with GLM, GMM, MRF, RW, and GMRF on the healthy controls’ data are shown in Figure 2.8. GLM missed many of the voxels in the hand area (Figure 2.7). GMM detected more voxels in the hand area, but also included many isolated false positives. MRF reduced the amount of isolated false positives, but mistakenly declared adjacent non-hand areas as active, and RW led to similar results. In contrast, GMRF correctly detected the hand area in all subjects. Our results thus demonstrate the benefits of incorporating group information, especially for tasks where we expect weak brain activations as in our study.
Results on the PD data are shown in Figure 2.9. The top and bottom rows correspond to PD patients off and on medication, respectively. GLM detected part of the hand area in most subjects, but also falsely declared islands of activation outside the hand area. GMM declared most voxels along the left M1 as active, which likely included many false positives. MRF detected spatially-contiguous clusters of voxels in the hand area, but also declared the nearby hip and leg areas, which should not be active unless the subjects’ hip and leg moved in sync with the stimulus. Using RW reduced the amount of false positives compared to GMM, but ample false positives remained, which were likely due to correlations arising from factors, such as physiological confounds, MR reconstruction, and head motion, mistakenly taken as voxel interactions. In contrast, GMRF consistently detected the hand area in all subjects without falsely declaring the hip and leg areas. In addition, the GMRF results suggest a very interesting trend between PD on and off medication. Specifically, PD off medication seemed to require recruiting a wider area of the left M1, which normalized back to an extent similar to the controls upon medication. Such spatial focusing effect of L-dopa medication has also been observed in past cognitive studies [145], thus confirming the validity of our results.
Figure 2.9 Real data results for PD subjects. 5 arbitrary exemplar subjects shown. Top and bottom rows of each sub-figure correspond to PD off and on medication, respectively.
2.8 Resting State-informed Activation Detection

The discovery of signal structure in ongoing brain activity in the absence of external stimulus has ignited enormous research interest [20, 88, 90, 92, 97]. Beginning with the seminal work by Biswal et al. [86], where coherent spontaneous fluctuations in neural activity were observed between brain regions in the somatomotor system, a huge body of work studying these spontaneous brain activities has followed [146]. For instance, a pioneer study by Raichle et al. investigating the baseline state of the human brain demonstrated a high degree of functional connectivity between the precuneus, posterior cingulate cortex, medial prefrontal cortex, and medial, lateral and inferior parietal cortex during rest and deactivation of these same regions during goal-directed tasks [147]. This set of brain regions has been coined the “default mode network” [147], which is one of the most widely-studied RS networks [148]. Others have found high signal correlations between the dorsolateral prefrontal cortex, inferior parietal cortex, and SMA at rest [149]. These regions are known to be associated with increased alertness during task performance, and have been termed the “task-positive network” [149]. There are many other RS networks that exhibit high resemblance to those engaged during various tasks [90], and all these findings are suggesting potential relationships between brain activity during task and rest, which may serve as an additional source of information for detecting brain activation. In particular, task-based fMRI data typically display low SNR, especially in patients due to difficulties in performing certain tasks [20]. Since acquiring RS data requires minimal task demands and RS data are less susceptible to behavioural confounds [20], extracting connectivity priors from RS data to inform activation detection may enhance detection sensitivity, which is particularly beneficial for studying diseased populations. We thus proposed to integrate RS-connectivity and task-evoked responses under a unified generative model [P6].
2.8.1 RS-informed Activation Model

Let $Y$ be an $d \times n$ matrix containing the intensity time courses of $d$ voxels or $d$ brain regions of a subject. Our proposed model can be summarized as follows:

$$ Y \sim N(AX, V_1) = \frac{1}{\sqrt{2\pi V_1}} \exp \left( -\frac{1}{2} \text{tr} \left( (Y - AX)^T V_1^{-1} (Y - AX) \right) \right), $$

(2.17)

$$ A \sim MN(M, V_2, K) = \frac{\sqrt{K}^d}{\sqrt{2\pi\alpha} V_2^{m/2}} \exp \left( -\frac{1}{2} \alpha \cdot \text{tr} \left( (A-M)^T V_2^{-1} (A-M)K \right) \right), $$

(2.18)

where $X$ is an $m \times n$ matrix with $m$ regressors along the rows for modeling stimulus-invoked response and confounds [31]. $A$ is the $d \times m$ activation effect matrix to be estimated (Section 2.8.2). $V_1$ is the $d \times d$ covariance matrix of $Y$, $V_2$ is an $d \times d$ covariance matrix modeling the correlations between the activation effects of the $d$ brain regions, and $K$ is an $m \times m$ covariance matrix modeling the correlations between the experimental conditions. $MN(M,V_2,K)$ denotes the matrix normal distribution, which serves as the conjugate prior of (2.17) [150] with $\alpha$ controlling the degree of influence of this prior on $A$ (Section 2.8.4). To ensure that our estimate of $A$ is invariant to affine transformations on $Y$ and $X$, we set $M$ to $0_{d \times m}$ and $K$ to $XX^T$ [150]. Setting $M$ to $0_{d \times m}$ also ensures that the activation effect estimates will not be biased towards non-zero values, which could induce false detections. $V_1$ and $V_2$ are assumed to be known. Compared to the model in [150] where $V_1$ and $V_2$ are assumed to be equal, we show that exact closed-form solutions for the posterior estimate of $A$ and the model evidence can be derived even for the more general case with $V_1$ and $V_2$ being distinct. Permitting distinct $V_1$ and $V_2$ accounts for how $Y$ and $A$ might have different correlation structures. Also, we hypothesize that brain regions that are functionally correlated at rest are more likely to display similar levels of brain activity during task performance. We thus set $V_2$ to the covariance estimates learned from RS data (Section 2.8.3). We assume $V_1 = I_{d \times d}$ as conventionally done for analytical simplicity, and defer learning $V_1$ from data for future work.
2.8.2 Posterior Activation Effects Estimation

To estimate $A$, we first derive the joint distribution of $Y$ and $A$ by taking the product of (2.17) and (2.18) and isolating the terms involving $A$ from those involving $Y$:

$$P(Y, A | X, V_1, V_2) \propto \exp \left( -\frac{1}{2} \text{tr} \left( (Y - AX)^T V_1^{-1} (Y - AX) + \alpha X^T A^T V_2^{-1} AX \right) \right)$$

$$= \exp \left( -\frac{1}{2} \text{tr} \left( V_1^{-1} + \alpha V_2^{-1} \left( AXX^T A^T - 2 \left( V_1^{-1} + \alpha V_2^{-1} \right) V_1^{-1} YX^T A^T \right) + V_1^{-1} YY^T \right) \right). \tag{2.19}$$

Since the terms involving $A$ take a quadratic form, the posterior distribution of $A$ is again a matrix normal distribution, as expected from the conjugacy of (2.17) and (2.18). By completing the square, the maximum a posteriori mean of $A$ can be derived:

$$M_A = \left( V_1^{-1} + \alpha V_2^{-1} \right)^{-1} V_1^{-1} YX^T \left( XX^T \right)^{-1}. \tag{2.20}$$

Computing $M_A$ requires inverting $V_1$ and $V_2$, which is unstable if $V_1$ and $V_2$ are set as sample covariance estimated from data with more brain regions than time points. In this work, we employ and compare two state-of-the-art techniques, namely graphical LASSO (GL) [106] and oracle approximating shrinkage (OAS) [151], for obtaining a well-conditioned $V_2$ with $V_1$ assumed to be $I_{d \times d}$.

2.8.3 Functional Connectivity Estimation

**Graphical LASSO.** Given a sample covariance matrix, $S$, computed from data that are assumed to follow a centered multivariate Gaussian distribution, one can estimate a well-conditioned sparse inverse covariance matrix, $\hat{\Lambda}$, by minimizing the negative log-likelihood of the data distribution over the space of positive definite (p.d.) matrices while imposing an $l_1$ penalty on $\hat{\Lambda}$ [106]:

$$\hat{\Lambda} = \arg \min_{\Lambda > 0} \text{tr} (\Lambda S) - \log \det(\Lambda) + \lambda \|\Lambda\|_1, \tag{2.21}$$
where \( \| \cdot \|_1 \) is the element-wise \( l_1 \) norm and \( \lambda \) controls the level of sparsity. Having a sparse inverse covariance matrix simplifies interpretation, since \( \hat{\lambda}_{ij} = 0 \) implies variables \( i \) and \( j \) are conditionally independent given all other variables. In the context of fMRI, \( \hat{\lambda}_{ij} = 0 \) implies brain regions \( i \) and \( j \) are not connected. To optimize (2.21), we employ the Two-Metric Projection method for its efficiency [152].

**Oracle Approximating Shrinkage.** Assume the data for estimating the ground truth covariance, \( \Sigma \), is generated from a multivariate Gaussian distribution. The most well-conditioned covariance estimate of \( \Sigma \) is \( F = \text{tr}(S)/d\cdot I_d \times d \) [151]. The idea of OAS is to shrink the ill-conditioned sample covariance, \( S \), towards \( F \) so that a well-conditioned covariance estimate, \( \hat{\Sigma} \), can be obtained. Specifically, OAS optimizes the following cost function [151]:

\[
\hat{\rho} = \arg \min_{\rho} E \left[ \left\| \hat{\Sigma} - \Sigma \right\|_F^2 \right], \quad \text{s.t.} \quad \hat{\Sigma} = (1 - \rho)S + \rho F,
\]

(2.22)

where \( \rho \) controls the amount of shrinkage with the optimal value given by [151]:

\[
\hat{\rho} = \left( 1 - \frac{2}{d} \right) \frac{\text{tr}(S^2) + \text{tr}^2(S)}{n + 1 - \frac{2}{d}} \frac{d}{\text{tr}(S^2) - \frac{d}{d} \text{tr}^2(S)}. \quad (2.23)
\]

Thus, no parameter selection is required and the inverse covariance matrix can be obtained by inverting \( \hat{\Sigma} \) with stable inversion warranted.

**2.8.4 Hyper-parameters Estimation**

A critical hyper-parameter in our model is \( \alpha \), which controls the degree of influence of the connectivity prior on the activation effect estimates. To set \( \alpha \), we first derive the model evidence by integrating \( P(Y,A|\alpha,V_1,V_2) \) over \( A \):
\[
P(Y | \alpha, V_1, V_2) = \frac{\left( V_1^{-1} + \alpha V_2^{-1} \right)^{-1/2}}{|2\pi V_1|^{1/2} |\alpha^{-1} V_2|^{1/2}} \exp \left( -\frac{1}{2} tr \left( Y Y^T \ldots \right) \right) \left( -Y X^T \left( X X^T \right)^{-1} X Y^T V_1^{-1} \left( V_1^{-1} + \alpha V_2^{-1} \right)^{-1} \right) \right).
\] (2.24)

Since \( V_2^{-1} \) estimated from GL or OAS is p.d., \( V_2^{-1} \) can be decomposed into \( Q \Gamma Q^T \) where \( Q \) are the eigenvectors of \( V_2^{-1} \) and \( \Gamma \) contains the eigenvalues of \( V_2^{-1} \) along the diagonal. By exploiting this property of p.d. matrix and that \( V_1^{-1} \) is assumed to be \( I = QIQ^T \), the log of (2.24) can be simplified into the following expression:

\[
\ln P(Y | \alpha, V_1, V_2) = -\frac{m}{2} \sum_{i=1}^{d} \left( \ln(1 + \alpha \gamma_i) - \ln(\alpha \gamma_i) - \frac{1}{m + \alpha \gamma_i} \right) + C, \tag{2.25}
\]

\[
B = Q^T Y X^T \left( X X^T \right)^{-1} X Y Q, \tag{2.26}
\]

where \( \gamma_i \) is the \( i^{th} \) eigenvalue of \( V_2^{-1} \), and terms that do not depend on \( \alpha, V_1, \) or \( V_2 \) are grouped into \( C \). Since (2.25) is a single variable function of \( \alpha \), the optimal \( \alpha \) at which (2.25) is maximized can be efficiently determined using any generic optimization routines.

To optimize the choice of \( \lambda \) for the case where \( V_2^{-1} \) is estimated using GL, we define a grid of \( \lambda \) values at which the percentage of non-zero elements in \( V_2^{-1} \) roughly ranges from 10% to 90%. For each \( V_2^{-1} \) associated with a given \( \lambda \), we determine the optimal \( \alpha \) at which (2.25) is maximized and compute the log model evidence. The \( \lambda \) associated with the largest log model evidence is then taken as the optimal \( \lambda \).
2.9 Synthetic RS/Task Paired Data Experiments

2.9.1 Data Generation

We generated 200 synthetic datasets based on our model described in Section 2.8.1. Each dataset consisted of 10 subjects. Each subject’s data comprised 100 regions with the first 20 regions set to be mutually correlated and activated across all subjects. The next 20 regions were set to be mutually correlated but not activated to test whether our model will falsely declare correlated regions as activated. The remaining regions were not correlated nor activated. To simulate the described scenario, we first generated 100 signal time courses, where the first 20 time courses were sine functions with Gaussian noise added. The next 20 time courses were cosine functions with Gaussian noise added. The remaining 60 time courses were simply Gaussian noise. We then computed the empirical covariance of these signal time courses, which was taken as the group covariance matrix of the activation effects and the resting state networks, $\Omega^g$. A random p.d. matrix was added to $\Omega^g$ to introduce inter-subject variability into the individual covariance matrix of each subject, $\Omega^i$. The degree of variability was controlled by restricting the Kullback-Leibler divergence of $N(0,\Omega^g)$ and $N(0,\Omega^i)$ to be $\sim 1.5$. For each subject $i$, 100 RS time courses of length $N_t$ were drawn from a multivariate Gaussian with zero mean and covariance $\Omega^i$. We set $N_t$ to 25 to emulate the typical situation where the number of regions exceeds the number of time points. To generate task data, samples of $A$ were independently drawn from (2.18) for each subject $i$ with $V_2$ set to $\Omega^i$. To simulate activation, the means of the first 20 regions in (2.18) were set to a small positive number, $\delta$, that depended on SNR, $\delta^2/\sigma^2$, at which we were testing. Random Gaussian noise was added to $\delta$ to further introduce inter-subject variability. The resulting $A$ were then used to draw samples of $Y$ from (2.17) with $V_1$ set to $\sigma^2I$ and $X$ being boxcar functions with 25 time points convolved with the canonical hemodynamic response function [24]. Two typical SNR levels were tested: 0.2 and 0.5 [21, 22], with 100 synthetic datasets generated at each SNR level.

2.9.2 Results and Discussion

For validation, we compared the sensitivity of our model with connectivity prior estimated from OAS and GL against the ridge regression model, i.e. $V_2$ set to $I$, and the standard
univariate model [31] in detecting group activation. We denote these models as OAS-CM, GL-CM, R-UM, and S-UM, respectively. Since the ground truth activated brain regions are unknown for real data, we employed the max-t permutation test [40] to enforce strict control on FPR so that we can safely base our validation on the number of regions detected with significant activation differences. We note that for each permutation, the entire activation effect map of each subject was multiplied by either 1 or -1 chosen at random. Hence, the spatial covariance structure of the activation effect maps was preserved. Also, we emphasize that our validation criterion is independent of the criterion used for optimizing the model parameters. This independence mitigates bias from being introduced during parameter estimation.

The mean receiver operating characteristics curves averaged over the synthetic datasets for each model are shown in Figure 2.10. At all FPR and both SNR levels, OAS-CM and GL-CM achieved higher TPR than R-UM and S-UM, thus confirming that given the activation effects are inherently correlated, our model can exploit this information to improve detection. We note that there is less need for imposing a prior (which introduces a bias) at higher SNR since more signals are available to estimate $A$, hence the decrease in sensitivity for OAS-CM and GL-CM.

![Figure 2.10](image)

**Figure 2.10** Synthetic data results. OAS-CM (red) and GL-CM (blue) achieved higher TPR for all FPR and both SNR levels than R-UM (green) and S-UM (black).
2.10 Real RS/Task Paired Data Analysis

2.10.1 Materials

Functional MRI data were provided by Dr. Jean Baptiste Poline of INRIA. 65 healthy subjects were recruited and scanned at multiple imaging centers. Each subject performed 10 language, computation, and sensorimotor tasks over a period of ~5 min (140 brain volumes) similar to those in [153]. RS data of ~7 min duration (187 brain volumes) were also collected. 3T scanners from multiple manufacturers were used for acquiring the data with TR = 2200 ms, TE = 30 ms, and flip angle = 75°. Standard preprocessing, including slice timing correction, motion correction, temporal detrending, and spatial normalization, were performed on the task-based data using the SPM8 software. Similar preprocessing was performed on the RS data except a band-pass filter with cutoff frequencies at 0.01 to 0.1 Hz was applied to isolate the signal of interest [88]. Signals from white-matter and cerebrospinal fluid voxels were regressed out from the gray-matter voxels.

To ensure stable sparse inverse covariance estimation using GL [106], we reduced the dimension of the data by grouping the voxels into 1000 parcels. Specifically, we concatenated the RS voxel time courses across subjects and applied the functional parcellation technique in [130] to generate a group parcellation map. Each subject’s brain images (in normalized space) were then parcellated using the group parcel labels. The mean voxel time courses of each parcel from rest and task were taken as the input to our model. To account for scanner variability across imaging centers, we normalized the parcel time courses by subtracting the mean and dividing by the standard deviation.

2.10.2 Results and Discussion

To test the generality of our model, we examined 21 contrasts between the 10 experimental conditions. Contrasts included computation vs. sentence processing task, auditory vs. visual task among others. Shown in Figure 2.11(a) is the percentage of parcels detected with significant activation differences averaged over contrasts for p-value thresholds ranging from 0 to 0.5. Incorporating a connectivity prior using OAS-CM and GL-CM outperformed R-UM and S-UM even under the simplifying yet common assumption of $V_1 = I_d \times d$. We note that
adding a shrinkage prior to control overfitting, as in the case of R-UM, only improved detection mildly. Thus, our results suggest an intrinsic relationship between the correlation structure of activation effects and RS-connectivity, and this relationship is consistent across subjects, hence the improved group activation detection. Qualitatively, incorporating a RS-connectivity prior led to more detections of bilateral activation in brain regions implicated for the contrasts examined, examples of which are shown in Figure 2.11(b) and (c).

![Figure 2.11](image)

**Figure 2.11** Real data results. (a) Percentage of parcels with significant activation differences averaged across contrasts vs. p-value thresholds. (b) Parcels detected by contrasting computation against sentence processing task, and (c) auditory against visual task. Red = detected by only OAS-CM. Purple = detected by both OAS-CM and GL-CM. Blue = detected by all methods.
2.11 Summary

In this chapter, we contrasted the effects of different priors on activation detection. The strengths and weaknesses of the current state-of-the-art are summarized in Table 2.1.

Table 2.1  Summary of activation detection methods. GA = group analysis. MCC = multiple comparisons correction. SS = spatial smoothing. VI = voxel interactions.

<table>
<thead>
<tr>
<th>Priors</th>
<th>Existing Methods</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Local VI</td>
<td>Long-range VI</td>
</tr>
<tr>
<td>Univariate</td>
<td>GLM with GRF threshold [31]</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>MM [43-45]</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Spatial</td>
<td>MRF [48-50]</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayesian [46, 47, 55-59]</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Wavelet [51, 53, 54]</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

As evident from Table 2.1, a common missing attribute in all of the existing methods is that they do not model long-range voxel interactions, despite distant voxels also interact with each other. We thus proposed integrating a connectivity prior to model both local and long-range voxel interactions. We showed that incorporating a connectivity prior increases detection sensitivity in intra-subject activation over using univariate and spatial priors, but is confounded by noise-induced correlations. The reason was likely due to limited discernible signals within a single dataset of a subject. As an additional source of information, we proposed incorporating a group prior using a novel graph-based approach that enables information coalescing between subjects without the need for a one-to-one voxel correspondence. We showed that integrating a group prior with our proposed graph-based approach improves detection sensitivity over using univariate and spatial priors, and provides stricter control over false positives than using a connectivity prior estimated from the same task-based datasets employed for activation statistics estimation. In addition, as an alternative source of information, we showed that extracting a connectivity prior from RS data provided enhanced sensitivity in group activation detection over the standard univariate approach.
3 Spatial Characterization of ROI Activation

A major limitation of the standard approach for fMRI group analysis is the need for drawing a one-to-one voxel correspondence across subjects. Due to the large inter-subject variability, spatial normalization tends to introduce errors into the analysis. An approach for mitigating mis-registration errors is to skip spatial normalization in all and perform group inference at the ROI level by comparing certain characteristics derived from regional activation statistics across subjects. Conventional features include thresholded mean activation statistics and the percentage of activated voxels [32, 33], which are solely based on the amplitude of the activation statistics without accounting for how these statistics are spatially distributed. As evident from recent brain decoding studies [35, 80], the spatial patterns of brain activity also encode important cognitive information. We thus proposed, during my Master studies, to use invariant spatial features to capture this spatial information [81] (Section 3.1). We also explored applying these spatial features directly to the BOLD signals to infer brain activation based on the spatial modulations of regional BOLD distributions [82]. In this thesis, we further validated our spatial characterization approach [81] on extended real data [P7] and investigated the benefits of adopting an ROI approach for spatial analysis by applying our proposed invariant spatial features to regional activation statistics with and without spatial normalization [P8] (Section 3.2). We also compared these spatial features against the traditional amplitude-based features in studying the effects of L-dopa on PD [P9] (Section 3.3). Furthermore, we applied our spatiotemporal characterization approach [82] on a large cohort of 65 subjects undergoing various experimental conditions to demonstrate its generality [P11] (Section 3.4). Our other related works not described in this chapter include investigation into the compensatory mechanisms in brain activations of PD patients using invariant spatial features [P10], comparisons between the temporal characteristics of spatial and BOLD responses [P12], proposal of a new spatial feature for measuring spatial skewness in BOLD signal patterns [P13], and inferring functional connectivity based on spatial modulations of regional BOLD distributions [P14].
3.1 3D Moment-based Invariant Spatial Features

Our proposed invariant spatial features [81] are derived based on 3D moments:

\[
m_{pqr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q z^r \rho(x,y,z) \, dx \, dy \, dz ,
\]

where \( n = p + q + r \) is the order of the 3D moment, \((x,y,z)\) are the coordinates of a voxel, and \( \rho(x,y,z) \) is the normalized t-value of a voxel located at \((x,y,z)\) within an ROI, as estimated using GLM. We restrict our analysis to only positive t-values due to the presently unclear interpretations of negative t-values [154]. The regional t-maps are normalized such that the t-values sum to one to decouple the effects of any overall amplitude changes.

Naively applying 3D moments to characterize the spatial distribution of the regional activation statistics can be problematic, since brain size and orientation differences across subjects would be inappropriately interpreted as spatial changes, which greatly amplifies functionally-irrelevant inter-subject variability. Therefore, metrics invariant to similarity transformations, i.e. rotation, translation, and scaling, are desired. Translational invariance can be obtained by using central moments defined as:

\[
\mu_{pqr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q (z - \bar{z})^r \rho(x,y,z) \, dx \, dy \, dz ,
\]

where \( \bar{x}, \bar{y}, \) and \( \bar{z} \) are the centroid coordinates of \( \rho(x,y,z) \):

\[
\bar{x} = \frac{m_{100}}{m_{000}}, \quad \bar{y} = \frac{m_{010}}{m_{000}}, \quad \bar{z} = \frac{m_{001}}{m_{000}},
\]

and scale invariance can be obtained by normalizing the moments:
\[ \eta_{pqr} = \frac{\mu_{pqr}}{\mu_{000}^{\frac{p+q+r+1}{3}}}. \]  

(3.4)

To obtain rotational invariance, the 3D moments need to be combined in specific ways. For all experiments in Sections 3.2 to 3.4, we focused on an invariant spatial feature [155]:

\[ J_1 = \eta_{200} + \eta_{020} + \eta_{002}, \]  

(3.5)

that measures the spatial extent of regional activation statistics, and refer readers to [81] and [P7] for other 3D moment invariants that we derived.

Figure 3.1  Spatiotemporal ROI analysis pipeline.

To extend the analysis to the spatiotemporal domain, we proposed computing \( J_1 \) of every regional BOLD volume to generate spatial feature time courses, i.e. \( \rho(x,y,z) \) in (3.2) is now the signal intensity of the voxel located at \( (x,y,z) \) at a given time point. We can then directly apply GLM to the resulting spatial feature time courses, as in the standard approach for activation detection [31], to identify activated ROIs [82], as summarized in Figure 3.1.
3.2 Effects of Spatial Normalization

The standard approach for detecting group activation [31] entails whole-brain spatial normalization, where each subject’s brain images are warped to a common template to generate a one-to-one voxel correspondence. The assumption is that aligning anatomical structures would correspondingly align the functional regions across subjects. However, due to the large anatomical and functional inter-subject variability, spatial normalization is prone to mis-registration errors. For instance, as discussed in a review by Uylings et al. [61], the location of Brodmann areas can vary considerably across normal individuals even after non-linear elastic registration. This issue of anatomical variability is further magnified in studies investigating the effects of age [156, 157] and neurological diseases [158], where the shape and size of the subjects’ brains may significantly depart from the commonly used templates [60, 159]. To circumvent this issue, Brett et al. proposed masking out regions of the brain that differ from the template so that these problematic regions would not affect the registration process [158]. However, manual delineation of such regions is quite laborious and automating the procedure is a highly challenging problem on its own.

To reduce the amount of mis-registration, region-level spatial normalization has been proposed and shown to provide higher sensitivity to activation detection compared to whole-brain spatial normalization [140, 160, 161]. Under this approach, an ROI is first segmented for each subject in its native space. The segmented ROIs are then aligned and averaged across subjects to generate an ROI template. The ROI of each subject is then warped onto the template to generate voxel correspondence. This region-level spatial normalization approach is especially important for analyzing smaller brain regions such as the thalamus and other subcortical structures, which tend to be mis-aligned in whole-brain spatial normalization. The reason is that smaller regions contribute considerably less to the objective function in driving the whole-brain registration process compared to larger brain structures. Hence, adopting a region-level spatial normalization approach, where each ROI is separately aligned, is likely to reduce anatomical mis-registration. However, functional mis-alignments between subjects remain an issue. Also, any spatial normalization will certainly distort the activation statistics patterns, which may affect the outcomes of the analysis.
An alternative approach for drawing group inference is to specify ROIs for each subject and compare certain properties of their regional activation statistics. Although this ROI analysis approach does not localize the active voxels, it addresses a more hypothesis-driven question of whether a particular brain region is commonly activated. A benefit of this approach is that it mitigates spatial distortions introduced by spatial normalization. We quantitatively assess the benefits of retaining the original activation statistics patterns by comparing the task discriminability of $J_1$ (3.5) for ROIs extracted 1) from the subject’s native space, 2) using whole-brain spatial normalization, and 3) using region-level spatial normalization [P8].

3.2.1 Assessment Methods

We use the fMRI data of the ten healthy subjects described in Section 2.7.1 to assess the effects of spatial normalization. We restrict our attention to four motor-related ROIs, namely the bilateral thalamus and the bilateral cerebellum, which are expected incur task-related activation changes. The respective sizes of the thalamus and cerebellum render these ROIs representative for testing the impact of spatial normalization on smaller and larger ROIs. The activation statistics, i.e. t-maps, are computed using the standard GLM [31]. To proceed, we first extract ROI activation statistics using three different procedures as summarized in Figure 3.2. For the warp-free case, each ROI is manually segmented from the whole-brain anatomical image by an expert based on anatomical landmarks and guided by a neurological atlas [159]. The segmented ROI is then used to extract the corresponding ROI activation statistics from the whole-brain activation statistics map without any spatial normalization. For the whole-brain spatial normalization case, the whole-brain anatomical image is warped to the Montreal Neurological Institutes (MNI) template using SPM 5 [30]. The same warp is then applied to the whole-brain activation statistics map with the activation statistics of each ROI extracted using predefined labels associated with the MNI template. Lastly, for the region-level spatial normalization case, the segmented anatomical ROIs from the warp-free procedure are rigidly aligned and averaged across subjects to generate an ROI template for each brain structure. This method for generating an ROI template is similar to the approach proposed by Stark and Okado [140], which provides a template that more closely matches the ROI shapes of the subjects under study. The anatomical ROI of each subject is then warped onto the ROI template with the same warp applied to the ROI activation statistics.
Figure 3.2 Extraction of ROI activation statistics. (a) Under the warp-free scheme, ROI activation statistics were extracted for each subject by masking out the appropriate regions from the whole-brain activation map using manually defined labels. (b) The whole-brain warping scheme involved warping the anatomical brain image of each subject to a template, applying the same warp to the whole-brain activation map, and extracting the ROI activation statistics using the pre-defined labels of the template. (c) The region-level warping scheme required creating an ROI template and warping of each subject’s ROI to this template with the same warp applied the corresponding activation statistics.
For each ROI activation statistics extraction procedure, we compute $J_1$ for each ROI and apply a paired t-test to contrast the activation differences between the fast and slow conditions of the experiment described in Section 2.7.1. Significance is declared at a family-wise p-value of 0.05 (with Bonferroni correction to account for the number of ROIs). The corrected critical p-value is 0.0127, which corresponds to a t-value of 3.10. For comparison, we also calculated the mean ROI activation statistics without spatial normalization.

### 3.2.2 Results and Discussion

Employing the traditionally-used mean activation statistics detected no significant task-related activation differences in any of the ROIs (left thalamus t-value = 0.213, right thalamus t-value = 0.160, left cerebellum t-value = 1.613, and right cerebellum t-value = 0.356). In contrast, $J_1$ without spatial normalization detected significant activation differences in all four ROIs as summarized in Figure 3.3, thus suggesting that important task-related information was encoded in the spatial distribution of the ROI activation statistics. $J_1$ with region-level spatial normalization detected the bilateral thalami and the left cerebellum but not the right cerebellum. $J_1$ with whole-brain spatial normalization detected only the left thalamus and at reduced discriminability compared to $J_1$ without spatial normalization. The quantitative results thus suggest that $J_1$ was more discriminant to task-related spatial changes under the warp-free settings than after whole-brain spatial normalization. $J_1$ without spatial normalization also appeared more discriminant than after region-level spatial normalization but to a lesser extent, except for the left thalamus. We suspect that this exception was due to an increase in functional overlaps across subjects upon region-level warping, which would increase the inter-subject consistency in the spatial activation patterns, hence the enhanced discriminability. However, this benefit of region-level spatial normalization was not evident for the other ROIs, thus might not be generalizable. To determine the reasons for the differences in discriminability between the compared methods, we generated 3D renderings of the ROI activation statistics under each of the three ROI activation statistics extraction schemes to visualize the effects of spatial normalization.
Figure 3.3  Discriminability comparisons between the contrasted procedures for extracting ROI activation statistics. The blue, magenta, and yellow bars correspond to warp-free analysis, region-level warping, and whole-brain warping, respectively. The green dash line is the t-threshold for a family-wise p-value of 0.05 with Bonferroni correction. Significant activation differences were detected in all ROIs for J₁ in the warp-free case. Region-level warping resulted in reduced discriminability (except in the left thalamus) and a loss of significance in the right cerebellum. Whole-brain warping led to further loss of significance in the right thalamus and the bilateral cerebellum.

3.2.2.1  Effects of Spatial Normalization on Smaller ROIs

The top two rows of Figure 3.4 show the 3D rendering of the warp-free ROI activation statistics within the right thalamus under the slow condition for all ten healthy subjects. As apparent from the figure, the locations of the more highly activated voxels varied across subjects and similar level of functional inter-subject variability was also observed for the fast condition (Figure 3.5). Hence, even if the ROIs were anatomically aligned perfectly, group activation differences might be difficult to detect using the standard univariate approach [31] due to functional misalignment. Instead, comparing Figures 3.4 and 3.5, a wider area of the right thalamus appeared to be consistently recruited in all subjects under the fast condition. Thus, measuring the spatial extent of ROI activation should provide higher sensitivity than examining each voxel in isolation for the type of activation changes observed.
Figure 3.4 ROI activation statistics in the right thalamus under the slow condition (Section 2.7.1). The top, middle, and bottom two rows correspond to warp-free, whole-brain warped, and region-level warped ROI activation statistics.
Figure 3.5 ROI activation statistics in the right thalamus under the fast condition (Section 2.7.1). The top, middle, and bottom two rows correspond to warp-free, whole-brain warped, and region-level warped ROI activation statistics.
One might argue that the lack of functional overlap was due to ROI misalignment, and conventional whole-brain spatial normalization should be performed before any conclusions could be drawn. The middle two rows of Figure 3.4 show the activation maps of the right thalamus extracted using the whole-brain spatial normalization scheme. Consistent with Crivello et al.’s findings [162], the amount of functional overlap did not seem to increase after whole-brain spatial normalization. Worse yet, whole-brain spatial normalization appeared to have pooled neighboring regions as part of the right thalamus. For instance, the more activated voxels were originally concentrated on the left side of the right thalamus in Figure 3.4(f), but an extra cluster emerged in the lower right region after whole-brain spatial normalization in Figure 3.4(p). This mis-registration problem was more pronounced in Figure 3.4(t), where the entire upper region of the right thalamus appeared to be active when only a few voxels may actually be activated (Figure 3.4(j)). Moreover, as evident by comparing Figure 3.4(c) and (m), whole-brain spatial normalization might have so heavily distorted the ROI activation patterns that the patterns no longer resemble the original data. Since both problems of mis-registration and spatial distortions were inherent in the whole-brain spatial normalization process, similar artifacts were found in the ROI activation maps for the fast condition (middle two rows of Figure 3.5) and in the left thalamus. Thus, although significant activation differences were detected in the left thalamus under whole-brain spatial normalization scheme, results should be interpreted with caution since the detected differences could have partly arise from the pooling of neighboring active regions as opposed to the true differences observed in the warp-free data.

A plausible explanation for the observed mis-registration and spatial distortions under the whole-brain spatial normalization scheme is that the thalamus is smaller than many other regions in the brain, hence may have little influence on the spatial normalization process. One way to reduce the amount of mis-registration is to perform region-level spatial normalization, which mitigates the problem of pooling neighboring regions. The bottom two rows of Figures 3.4 and 3.5 show the activation statistics of the right thalamus after region-level spatial normalization. Compared to the original warp-free data (top two rows of Figures 3.4 and 3.5), the overall spatial patterns appeared to be conserved although slightly distorted and shifted. The amount of functional inter-subject variability also seemed to have mildly
decreased, but this gain might have been offset by the smearing and spatial distortions introduced, which led to reduced discriminability (Figure 3.3). For the case of the left thalamus, the decrease in functional variability actually appeared to have outweighed the artifacts introduced, though this trend was not consistent for the other ROIs examined. Compared to whole-brain spatial normalization, a reduced amount of mis-registration and spatial distortions was definitely evident, and thus the increase in discriminability.

3.2.2.2 Effects of Spatial Normalization on Larger ROIs

The top two rows of Figures 3.6 and 3.7 show the warp-free activation statistics of the right cerebellum for the slow and fast conditions, respectively. As can be seen from the figures, considerable functional inter-subject variability was present in the spatial activation pattern. Also, comparing Figures 3.6 and 3.7, the spread of activation only mildly increased under the fast condition. For instance, only the middle region of the right cerebellum (Figure 3.6(e) and (j)) became more active (Figure 3.7(e) and (j)). Thus, any distortions introduced by spatial normalization might remove these subtle spatial differences.

Since larger ROIs have more influence on whole-brain spatial normalization process, the cerebellum should theoretically be less misaligned than the thalamus. Examining the middle two rows of Figure 3.6 supports this hypothesis. Although spatial distortion was apparent and neighboring regions might have been pooled into the border of the right cerebellum (e.g. top of right cerebellum in Figure 3.6(m)), correspondence in the overall spatial pattern could roughly be drawn between the warp-free and whole-brain warped activation maps. Nevertheless, region-level spatial normalization again resulted in less spatial distortions than whole-brain spatial normalization as shown in the bottom two rows of Figures 3.6 and 3.7. Correspondence in the overall spatial patterns could be easily identified between the warp-free and region-level warped cases (e.g. Figure 3.6(j) and (η)), but ample amount of smearing was clearly present (e.g. Figure 3.6(i) and (γ)). Since only mild spatial differences were observed between the slow and fast conditions in the original warp-free activation maps, this smearing effect seemed to have reduced task discriminability, leading to a loss of significance. Thus, our results again suggest the importance of retaining the original ROI activation statistics patterns in fMRI spatial analysis.
Figure 3.6 ROI activation statistics in the right cerebellum under the slow condition (Section 2.7.1). The top, middle, and bottom two rows correspond to warp-free, whole-brain warped, and region-level warped ROI activation statistics.
Figure 3.7 ROI activation statistics in the right cerebellum under the fast condition (Section 2.7.1). The top, middle, and bottom two rows correspond to warp-free, whole-brain warped, and region-level warped ROI activation statistics.
3.3 Effects of L-dopa on Parkinson’s Disease

Despite the clinical benefits of L-dopa on the motor symptoms of PD, neuroimaging studies have revealed conflicting results for the effects of L-dopa on brain activity. Resting state PET studies of PD subjects have shown a pattern of hypermetabolism in lentiform and thalamic areas and hypometabolism of cortical motor regions, which is improved by L-dopa [163-165]. Earlier studies, however, showed no effect of L-dopa on regional cerebral glucose metabolism [166]. During voluntary movement, studies have also produced conflicting results. While some have demonstrated hypoactivity of the SMA [167], others have found increased activity of caudal portions of the SMA [168]. Similarly, Haslinger et al. [169] showed hyperactivity in M1, while Buhmann et al. [170] found hypoactivity in M1. Nonetheless, both studies showed that L-dopa “normalized” brain activity.

A possible explanation of these conflicting results is that the aforementioned studies have focused on examining the amplitude of ROI activation statistics, and neglected possible characteristic changes in the regional spatial patterns. The assumption that only amplitude is modulated by disease can be overly simplistic, since a “spatially diffused” and a “spatially focused” pattern may have similar mean ROI activation statistics. We hypothesize that one of the effects of L-dopa may in fact be a “re-focusing” of brain activity in specific brain regions. We test this hypothesis by measuring the spatial extent of regional activation statistics in the PD subjects off and on L-dopa medication (Section 2.7.1) using $J_1$ (3.5) [P9].

3.3.1 Statistical Analysis

For each ROI and experimental condition (Section 2.7.1), we computed $J_1$ of the regional activation statistics of the PD subjects off and on medication and the age-matched controls. For comparison, we also computed the mean ROI activation statistics thresholded at a t-value of 1.96. A 2×3 repeated-measures design analysis of variance (ANOVA) was performed on the PD data to analyze the main effects of medication, movement frequency, and any interactions between these two factors. A paired-test was also performed to examine the effects of medication at each task frequency. Statistical significance is declared at a family-wise p-value of 0.05 with FDR correction [52].
3.3.2 Results and Discussion

Applying an ANOVA on the mean ROI activation statistics detected significant effects of L-dopa on the bilateral M1 and the contralateral SMA. Frequency had a significant main effect in left putamen, bilateral M1, and bilateral SMA. Assessing the effects of L-dopa at each frequency separately detected no significant changes within any ROIs at the slow frequency. At the medium frequency, only the left M1 demonstrated a significant decrease in amplitude. At the fastest frequency, significant decreases in amplitude were detected in the contralateral SMA, ipsilateral putamen, thalamus, cerebellum, and M1.

Using $J_1$, significant main effects of L-dopa were detected in bilateral cerebellum, bilateral M1, bilateral SMA, and ipsilateral prefrontal cortex. A significant interaction in the left prefrontal cortex was found. No main effect of frequency was detected, but frequency specific effects of L-dopa were found in the ipsilateral cerebellum and SMA, and contralateral primary motor cortex at the slow frequency. At medium frequency, significant medication effects were detected in the contralateral thalamus, ipsilateral cerebellum, and bilateral M1 and SMA. At the highest frequency, significant effects were detected in bilateral prefrontal cortex and SMA as well as contralateral cerebellum and ipsilateral M1.

To visualize the effects of L-dopa, we plotted the t-values within the right M1 of an exemplar PD subject off and on L-dopa Figure 3.8. Comparing the spatial distribution of t-values, L-dopa appears to ‘refocus’ the area of activation. This trend was consistently found in all PD subjects as shown in Figure 3.9, where the spatial extent of regional activation statistics in PD subjects off and on medication as compared to healthy age-matched controls within the left SMA and right M1 (two regions where both amplitude and spatial distribution of activation was significantly affected by L-dopa) are plotted. The spatial extent of activation statistics within these two regions was reduced to a similar level as controls upon medication, thus demonstrating the normalizing effect of L-dopa.
Figure 3.8  Spatial focusing effect of L-dopa seen in a typical PD subject. The t-maps of right M1 shown with t-values normalized between [0,1] to emphasize the spatial distribution changes. (a) A much wider recruitment of the right M1 was required to perform the same motor task prior to L-dopa medication. (b) The ROI activation pattern appears to focus upon medication.

Figure 3.9  Normalization of the spatial extent of ROI activation statistics by L-dopa. (a) The spatial extent of activation statistics in the left SMA is plotted. The arrows represent the effects of L-dopa on the spatial extent of ROI activation statistics in each PD subject. The feature values for the normal subjects are shown for comparison. (b) Normalization of spatial extent of activation statistics in the right M1. L-dopa appeared to refocus the regional activation in PD with the resultant spatial extent becoming more similar to that of the healthy control subjects.
Based on our results, a predominant effect of L-dopa appeared to be a change in the spatial extent of activation in both cortical and subcortical regions. In fact, the spatial effects of medication were significant in the contralateral M1 and ipsilateral cerebellum even at the lowest frequency, whereas medication-related changes in amplitude, as estimated with mean ROI activation statistics, were only evident at the higher frequencies. Similar findings have been observed in cognitive studies [145], and we are demonstrating here that such “focussing” effect of L-dopa is also present for motor tasks. This important brain characteristic would have been missed if only amplitude-based features were employed.

3.4 Large Scale Real Data Experiment

We have previously demonstrated that spatial modulations in regional BOLD distributions can be used to infer brain activation for motor tasks [82]. The question is thus whether this is generally true for all different tasks. To address this question, we apply our spatiotemporal characterization approach [82] to fMRI data (Section 2.10.1) collected from a large cohort of 65 subjects [P11]. Each subject performed 10 tasks: 1) passive viewing of flashing horizontal checkerboards, 2) passive viewing of flashing vertical checkerboards, 3) pressing a button with the left thumb button according to visual instructions, 4) pressing a button with right thumb according to visual instruction, 5) pressing a button with left thumb according to auditory instruction, 6) pressing a button with right thumb according to auditory instruction, 7) silently reading short visual sentences, 8) listening to short sentences, 9) solving silently visual subtraction problems, and 10) solving silently auditory subtraction problems [153]. This dataset thus encompasses tasks, such as auditory perception, visual perception, motor actions, reading, language comprehension, and mental calculation, and hence particularly suitable for testing the generality of our spatiotemporal characterization approach.

3.4.1 Statistical Analysis

The overall analysis pipeline is summarized in Figure 3.1. After extracting ROIs using the AAL atlas [171], we normalized the regional BOLD distributions such that, at each time point, the intensity of the voxels is non-negative and sums to one. This step decouples
amplitude changes from spatial changes. \( J_1 \) was then computed at each time point to generate a spatial feature time course. Standard GLM [31] was employed to estimate the activation effects. To identify ROIs with significant activation changes, a non-parametric permutation test with strict control over false positives was used [40]. Statistical significance was declared at an experiment-wise \( p \)-value of 0.05. For comparison, the same analysis was performed using mean intensity within parcels as feature time courses. We employed parcel-level analysis since the mean intensity over the large AAL ROIs might be too coarse a measure for representing the ROI response. Parcels were functionally defined [130] based on resting-state data of the subjects (Section 2.10.1).

### 3.4.2 Results and Discussion

21 contrasts between the 10 experimental conditions were examined. Contrasts included auditory vs. visual task, language vs. computation task among others. Figure 3.10 shows the percentage of AAL ROIs detected with significant activation differences using \( J_1 \). On average, significant activation differences were found in 9% of the AAL ROIs across contrasts. Thus, task-related spatial modulations of the BOLD distribution appeared to be present for various tasks, and not specific to only motor tasks. Also, a moderate number of ROIs detected with \( J_1 \) were different from those found using mean intensity, as shown in Figure 3.11(a) and (b). This observation suggests that spatial characterization of the regional BOLD response can play a complementary role to traditional intensity-based analysis.
Figure 3.10 Percentage of ROIs detected with significant activation differences using $J_1$. L = left, R = right, H = horizontal, V = vertical.

Figure 3.11 Activation detection comparisons between $J_1$ and mean intensity. Red = detected by $J_1$. Blue = detected by mean intensity. (a) Auditory vs. visual task. (b) Sentence vs. computation task.
3.5 Summary

We validated the feasibility of inferring brain activation from spatial changes in regional activity patterns as measured with our proposed invariant spatial features. We showed that task-related spatial modulations are generally present for different tasks. This important finding indicates that analyzing signal amplitude is not the only means for detecting brain activation. Another point to emphasize is that our spatial characterization approach takes into account the collective information encoded by both the strongly and weakly responding voxels. From a spatial pattern analysis perspective, weakly responding voxels might just be as relevant, yet are often discarded in standard univariate analysis. Analyzing fMRI data at the ROI level thus opens up a new dimension – the spatial dimension, in complementing traditional univariate approaches.
4 Group-wise Functional Network Analysis

The standard seed-based approach infers brain networks by comparing the intensity time courses of all voxels in the brain against that of a seed. Joint voxel interactions are thus ignored. Also, the detectable networks are restricted by the choice of seeds. Cluster analysis and decomposition techniques can alleviate these limitations, but identifying common networks across subjects using these approaches is very challenging, especially in the present of strong noise as inherent in fMRI data [108]. If each subject is analyzed separately, the less pronounced commonalities are likely to be obscured by noise. To deal with the observed inter-subject variability, we propose a new method, group replicator dynamics (GRD) [P15, P16], to exploit our prior knowledge of how subjects belonging to the same population likely recruit similar networks [108]. GRD is based on RD [70], which we show to be a simple solution to the NSPCA problem (Section 4.1). Exploiting the sparse property of RD thus eases the diffused weighting problem of PCA [102, 103] and enables explicit modeling of the sparse connections between spatially-remote brain regions [172]. To facilitate statistical group inference, we propose learning each subject’s network individually using RD (as opposed to applying RD to the average similarity matrix of the group [112]) but with group information integrated into each subject’s RD process (Section 4.2). This strategy encourages the individual networks to evolve towards the common network of the group. This results in sparse networks comprising the same brain regions across subjects yet with subject-specific regional weightings that reflect individual differences. Our approach thus enables sharing of statistical strength among subjects, while permitting inter-subject variability to be modeled. Statistical group inference can thereby be performed. We validated GRD on synthetic data (Section 4.3) and applied it to study network differences between tremor-dominant PD and akinetic-rigid PD subjects (Section 4.4).
4.1 Replicator Dynamics

RD is a well-known concept that originated from theoretical biology for modeling the evolution of interacting, self-replicating entities [70]. Under the RD model, only the fittest entities survive over time. Let \( w(k) \) be a vector with the \( j \)th element being the probability that allele \( j \) remains in the gene pool during the \( k \)th generation and \( C \) be a ‘fitness’ matrix with each element reflecting the fitness of a genotype (a pairwise combination of alleles), then based on the theory of natural selection, \( w(k) \) can be estimated as follows [70]:

\[
  w(k + 1) = \frac{w(k) \cdot Cw(k)}{w(k)^T Cw(k)},
\]

where \( \cdot \) represents element-wise multiplication. According to the fundamental theorem of natural selection [70], if \( C \) is real, symmetric, and non-negative, then the function \( w(k)^T Cw(k) \), which corresponds to the average fitness of the surviving alleles, is strictly increasing with increasing \( k \) along any non-stationary trajectory \( w(k) \), and any such trajectory converges to a stationary point \( w_s \). Furthermore, \( w_s \) is asymptotically stable under (4.1) if and only if \( w_s \) is a strict local maximizer of \( w^T Cw \). A proof of this theorem is provided in [173]. Based on this theorem, (4.1) maximizes the same objective function as PCA, i.e. \( w^T Cw \), but with a different set of constraints. Specifically, the non-negative constraint on \( C \) restricts \( w \) to be non-negative. Also, since the sum of the elements in the numerator equals the value at the denominator in (4.1), elements of \( w \) are enforced to sum to one. Hence, (4.1) is a local maximizer of the following optimization problem:

\[
  \max_{w} w^T Cw, \quad s.t. \|w\|_1 = 1, w \geq 0,
\]

where the \( w \) is constrained to lie on the standard simplex, i.e. \( \|w\|_1 = 1, w \geq 0 \). As shown in the seminal paper by Tibshirani [127], constraining the \( l_1 \) norm induces sparsity. Thus, (4.1) provides an elegant solution to the challenging problem of NSPCA [102, 103] that can be easily implemented in just a few lines of code. Also, since only vector-matrix multiplications
are involved, the complexity at each iteration is $O(n^3)$, which renders RD computationally attractive against other NSPCA methods [103]. In the context of functional connectivity analysis, a brain region is analogous to an allele with its fitness reflected by the similarity between its time course and those of other ROIs. The most correlated set of ROIs can thus be found by iterating (4.1), where the non-zero elements in $w$ indicate the ROIs belonging to the dominant network.

### 4.2 Group Replicator Dynamics

If we apply (4.1) separately for each subject, the detected networks will likely be different across subjects [113]. However, if a core functional network that is common across subjects exists, we can exploit group information to identify this core network. The underlying assumption is that subjects belonging to the same group presumably use the same core functional network, comprising the same brain regions, to perform the same task. However, the common core network may not necessarily be the most coherent network for each subject, potentially due to different processing strategies or simply due to noise [113].

#### 4.2.1 GRD Formulation

To identify the common core networks, we propose integrating group information in learning each subject’s network by maximizing the following objective function:

$$
\hat{W} = \arg \max_{W} J(W) - \lambda E(W) \quad \text{s.t.} \quad \left\| w^i \right\|_1 = 1, \ w^i \geq 0, \ i = 1,..., N_s, \quad (4.3)
$$

$$
J(W) = \sum_{i=1}^{N_s} w^i w^i \quad E(W) = \ln \sqrt{(2\pi)^{N_r} |\Sigma|}, \quad (4.4)
$$

where $W$ is an $N_r \times N_s$ matrix with the weight vector $w^i$ of each subject $i$ along the columns, $N_r$ is the number of ROIs, and $N_s$ is the number of subjects. $J(W)$ corresponds to the sum of average network correlations across subjects (recall that $w^T C w$ is the average fitness of the surviving alleles in the original context of RD, Section 4.1), and $E(W)$ is the group entropy of
the weight vectors assuming the $w^i$'s are instances of a random variable, $w$, that follows a multivariate normal distribution [174] with $\lambda$ governing the degree of group support. $\Sigma$ is the covariance of $w$. Since entropy measures the uncertainty in predicting any instance of $w$ given the other instances, e.g. predicting $w^p$ given $w^q$, minimizing the entropy increases the mutual similarity of the $w^i$'s based on commonalities shared between the subject’s network.

4.2.2 GRD Optimization

To maximize (4.3), we propose the following optimization strategy. Let $w^i(k)$ be the weight vector of subject $i$ during the $k^{th}$ iteration. At each iteration $k$, we first update each $w^i(k)$ individually using RD:

$$w(k+1) = \frac{w(k) * Cw(k)}{w(k)^T Cw(k)}, \quad (4.5)$$

We then adjust $w^i_{RD}(k)$ to reduce group entropy:

$$w^i(k+1) = w^i_{RD}(k) + \lambda \delta^i(k), \quad (4.6)$$

where $\delta^i(k)$ is the adjustment on $w^i_{RD}(k)$ derived based on gradient descent [175]. Specifically, let $W_{RD}(k)$ be an $N_r \times N_s$ matrix with $w^i_{RD}(k)$ of each subject along the columns, and $W_c(k)$ be the same as $W_{RD}(k)$ but with the subject mean removed from each row. We approximate $\Sigma$ in (4.4) with the sample covariance, $W_c W_c^T / (N_s - 1)$, and take the matrix derivative of $E(W_c)$ with respect to $W_c$:

$$\frac{dE(W_c)}{dW_c} \propto \left(W_c W_c^T \right)^{-1} W_c. \quad (4.7)$$

Based on (4.7), all $w^i_{RD}(k)$ can be simultaneously adjusted:

$$W(k+1) = W_{RD}(k) + \lambda \delta(k), \quad (4.8)$$
\[ \delta(k) = -(W_c(k)W_c(k)^T)^{-1}W_c(k), \]  

(4.9)

where \( \lambda \) corresponds to the step size in the context of gradient descent optimization (see Section 4.2.4), and \( \delta_i(k) \) in (4.6) is the \( i \)-th column of \( \delta(k) \). After applying (4.8), (4.5) is applied again in the next iteration, \( k+1 \), and the same procedure is repeated until convergence. We note that provided the step size, \( \lambda \), is not “too large” as discussed in Section 4.2.4, applying (4.5) will project the weight vectors back onto the standard simplex, which ensures the constraints in (4.3) are satisfied. Since (4.5) searches for the network with highest average correlation within each subject, while (4.8) encourages networks to be similar across subjects, the combined effect of (4.5) and (4.8) results in networks with high average correlations that comprise the same ROIs across subjects, i.e. \( w_i(k) \)'s having the same sparsity pattern, but with subject-specific ROI weightings.

For detecting subsequent networks, one may remove ROIs detected as the dominant network and reapply the procedures above [112]. Alternatively, if we consider ROIs as nodes of a graph and elements of \( C_i \) as edge weights, we can instead remove only those edges present in the dominant network, which enables the same ROI to be in multiple networks. Both of these approaches enable all functional networks in the brain to be potentially discovered in a hierarchical manner, in contrast to the seed-based approach where the detectable networks are restricted by the choice of seed [98]. We have adopted the latter approach to examine which ROIs serve as communication channels between the brain networks.

We note an interesting analogy between GRD and simulated annealing [176]. Specifically, applying (4.8) at each iteration is analogous to perturbing the RD solution of each subject to avoid local maxima. The difference is that the perturbation in GRD is informatively designed based on group consistency, as opposed to being random. Also, as the RD process of each subject evolves towards the common group network, the amount of perturbation will automatically decrease. Thus, explicitly designing an annealing schedule is not required. Furthermore, this designed perturbation renders GRD robust to initialization, as we show later in Section 4.3.
4.2.3 Group Inference

To test if a detected network is statistically significant, we first compute the average correlation of the brain regions in the detected network of each subject, \( c_i = w_i^T C^i w_i \). We then apply a one-sample t-test to the Fisher’s Z-transform of the \( c_i \)'s, which we denote as \( z_i \), against an estimated null mean to determine the probability that the \( z_i \)'s are drawn from a null distribution generated from random noise. To estimate the mean of the null distribution, we temporally permute all preprocessed ROI time courses \( N = 1000 \) times, calculate the mean \( z \) averaged across subjects for each permutation, and take the average mean \( z \) over permutations as the null mean. Details regarding the general theories behind application of permutation tests for drawing statistical inference can be found in [40].

Since a network is characterized by its connections, the average connection strength, as (i.e. \( c_i = w_i^T C^i w_i \)), which GRD maximizes, would be a representative characteristics of a network. However, comparing \( w_i^T C^i w_i \) across subjects is legitimate only if the same set of ROIs is detected as the common core functional network of the subjects. Otherwise, each estimated \( c_i \) would represent the average correlation of a different network and declaring the presence of a significant common group network would be invalid. Also, for the same reason, comparing average network correlations would not be appropriate for statistically quantifying network differences between groups, since network of different groups may comprise different ROIs. Therefore, for statistically comparing the networks between two groups, we employ a permutation test [177]. In brief, we first compute the Euclidean distance (ED) between the mean weight vectors of the two groups. We then permute the group labels of the weight vectors and count the number of times, \( q \), that the ED of the permuted groups exceeds the original ED. The p-value is then computed as \( q/N \), where \( N \) is the number of permutations set to 100,000.
4.2.4 Implementation Details

Functional connectivity analysis is typically performed at the voxel level [84] with subject correspondence drawn by whole-brain spatial normalization. However, voxel time courses are often very noisy and whole-brain spatial normalization is prone to mis-registration [61]. Hence, the implicit assumption that a perfect voxel correspondence is established upon whole-brain spatial normalization is quite tenuous. We thus instead define ROIs based on anatomy and set $C^i$ as the pairwise correlations between the average intensity time courses of the ROIs. The underlying assumption is that the same anatomical ROIs pertain to similar functions across subjects. To ensure $C^i$ is non-negative as required for employing RD, we take the absolute value of the correlations [112] assuming positive and negative correlations are equally important. Nevertheless, networks with only positively correlated ROIs can be separately identified by nulling out the negative elements in $C^i$. To avoid self connections, the main diagonal of $C^i$ is set to zero. Furthermore, to avoid bias during initialization, $w^i(0)$ is set to $1/N_r$. Convergence of GRD is defined as having all elements of $w^i(k)$ change by $< 10^{-4}$.

$W_c(k)W_c(k)^T$ in (4.9) is typically singular since there are usually more ROIs than subjects, i.e. $N_r \gg N_s$. Therefore, we add a small positive constant, $\alpha$, to the main diagonal of $W_c(k)W_c(k)^T$ to regularize the inversion. Adding $\alpha$ to $W_c(k)W_c^T(k)$ reduces the influence of the off-diagonal elements of $W_c(k)W_c^T(k)$. These off-diagonal elements encode information regarding which pairs of ROIs are consistently selected (given non-zero weights) across subjects. Hence, one would ideally want to add as small an $\alpha$ value as possible such that all eigenvalues of $W_c(k)W_c^T(k)$ are positive to ensure stable inversion of $W_c(k)W_c^T(k)$, while the influence of the off-diagonal elements is preserved. Searching for such an $\alpha$ value at every iteration $k$, however, can be computationally expensive. We thus opt to fix $\alpha$ at 0.1, which is an order of magnitude lower than the value bounds of $W_c(k)W_c^T(k)$, i.e. (-1,1). We tested on a wide set of synthetic cases with different number of ROIs and noise levels, and found that selecting an $\alpha$ value as such provides a good balance between ensuring stable inversion of $W_c(k)W_c^T(k)$ and preserving the influence of the off-diagonal elements, as confirmed by how the detected networks perfectly matched the ground truth in our synthetic examples shown in Section 4.3.
As for $\lambda$ in (4.8), which corresponds to step size in the context of gradient descent optimization, there are two standard methods for setting this parameter [175]. Specifically, one may employ exact line search [175] to find the $\lambda$ that minimizes the group entropy at every iteration, but this can be very computationally expensive. Therefore, we instead adopted the backtracking line search strategy [175]. In brief, starting from an $\lambda$ of 1, the idea is to reduce $\lambda$ until the Armijo condition is satisfied:

$$E(W_{RD}(k) + \lambda \delta(k)) \leq E(W_{RD}(k)) + \gamma \lambda \nabla E(W_{RD}(k))^T \delta(k),$$  (4.10)

where $\gamma \in (0,0.5)$. This condition guarantees sufficient decrease in group entropy at each iteration. For a typical $\gamma$ of 0.01 [175], we found that setting $\lambda$ to 0.1 satisfies the Armijo condition at every iteration for all synthetic cases and real data tested. We thus opt to fix $\lambda$ to 0.1 as opposed to checking the Armijo condition at every iteration to find a potentially larger $\lambda$. The robustness of GRD as we vary $\lambda$ is discussed in Section 4.3.

4.3 Synthetic Data Experiments

4.3.1 Data Generation

We validated GRD under two test scenarios. 500 synthetic datasets were generated for each test. Scenario 1 consisted of 10 subjects and 100 ROIs, as summarized in Figure 4.1(k). ROIs 1 to 10 were designed to be mutually correlated and constituted the primary group network (i.e. 10% of the ROIs belong to the primary group network). These ROIs were also set to be connected to ROI 21 for subject 1, ROI 22 for subject 2, etc. to introduce inter-subject variability. We note that these extra subject-dependent ROIs are not supposed to be part of the primary group network. Furthermore, we set ROIs 11 to 20 to be mutually correlated but with lower correlations except for subjects 1 and 2 to test whether GRD will get trapped in a local maximum when a secondary network is present and the primary group network is not the primary network for some of the subjects. Sinusoidal signals were used as synthetic ROI time courses to emulate the hypothesized response of the continual task employed in our clinical study, which is likely to be sinusoidal (Section 4.4.1):
where $f_p(t)$ and $f_s(t)$ are time courses of ROIs belonging to the primary and secondary group networks, respectively. $w_p$ and $w_s$ represent random Gaussian noise with $w_p < w_s$ (except for subjects 1 and 2) so that ROIs of the primary network have higher mutual correlations. $s$ is a uniform random variable with values between 0.7 and 1. We note that $f_p(t)$ is deliberately designed to be mildly correlated with $f_s(t)$ to test whether GRD can correctly separate the two networks. For ROIs that do not belong to the primary and secondary network, random Gaussian noise is used as ROI time courses.

### 4.3.2 Results and Discussion

Results obtained by applying GRD to the synthetic data of scenario 1 are plotted in Figure 4.1. For comparison, we applied PCA, RD, and ICA separately to each subject’s data. Group ICA was also tested. For ICA and Group ICA, the spatial component associated with the temporal component that has highest correlation with $\sin(0.5t)$ was taken as $w^i$, i.e. $\sin(0.5t)$ is taken as the hypothesized response of ROIs in the primary group network.

Using PCA (red “.”), higher weights were assigned to ROIs 1 to 10 and the extra subject-dependent ROI for subjects 3 to 10, and ROIs 11 to 20 for subject 1 and 2. However, if we examine each dataset separately, weights assigned to ROIs of the primary and secondary group networks were in fact very similar with higher weights occasionally assigned to ROIs of the secondary group network. Defining a threshold to separate the primary and secondary group networks is thus very difficult. Also, non-zero weights were assigned to ROIs with random noise as time courses. Since PCA does not account for group information, detecting the extra subject-dependent ROI and having the primary network of subjects 1 and 2 deviating from the primary group network were expected.
Figure 4.1  Synthetic data results. (a) to (j) Average weight vectors of subjects 1 to 10 over 500 datasets (k) Ground truth subject-specific networks. “P” and “S” indicate if an ROI belongs to the primary or secondary network of each subject. ROIs 1 to 10 constitute the primary group network. “-” implies the ROI does not belong to a network. We zoomed into the weights of ROIs 1 to 40 for clarity. Weights of ROIs 41 to 100 were similar to those of ROIs 31 to 40 since all these ROIs do not belong to any network.
RD (green “x”) also assigned higher weights to ROIs 1 to 10 and the extra subject-dependent ROI for subjects 3 to 10, and ROIs 11 to 20 for subjects 1 and 2 on average. These deviations from the ground truth primary group network were expected since RD applied separately to each subject will obviously not be able to exploit group information. We note that zero weights were assigned to ROIs with random noise as time courses, thus demonstrating the sparse property of RD. Surprisingly, non-zero weights were sporadically assigned to ROIs of the secondary network when we examine results of each synthetic dataset individually. We suspect the amount of noise added might have obscured the true degree of correlations between ROI time courses of the two networks.

ICA (cyan “+”) resulted in higher weights assigned to ROIs 1 to 10 for all subjects, since we knew a priori that the temporal component having highest correlation with \(\sin(0.5t)\) corresponds to the primary group network. Such prior knowledge is rarely available for real experiments, especially for complex tasks, such as the continual motor squeezing task employed in our study (Section 4.4.1). Also, the extra subject-dependent ROI was incorrectly detected, since ICA does not incorporate group information. Moreover, non-zero weights were assigned to ROIs of the secondary group network, though the magnitude was substantially reduced compared to PCA.

Results obtained using Group ICA (blue “”) were very similar to ICA except for a reduction in standard deviation of the weights (not shown in Figure 4.1 to avoid cluttering the plot). Thus, the back-reconstruction step in Group ICA [72] does not preserve the group network structure. If we apply a t-test to the subjects’ weights for each ROI separately as proposed in [72], the subject-dependent ROI would be correctly pruned out, but ROIs in the secondary group network would be mistakenly taken as part of the primary group network since small positive weights of similar magnitude were consistently assigned to these ROIs across all subjects. If we instead threshold the weights to avoid falsely including ROIs from the secondary group network, the subject-dependent ROI would be erroneously treated as part of the primary group network. Thus, no simple automated heuristics would be sufficient to separate out the primary group network.
Using GRD (black “o”), weights were correctly assigned to only those ROIs belonging to the ground truth primary group network for all subjects. Also, zero weights were assigned to all other ROIs, thus no heuristics were required to identify the primary group network. We note that with $\lambda$ set to 0.1, which satisfied the Armijo condition, GRD was found to converge for all synthetic datasets. A plot of (4.3) vs iteration $k$ is shown in Figure 4.2 for an exemplar synthetic dataset.

![Figure 4.2 Convergence of GRD. GRD was found to converge for all synthetic data tested.](image)

Notice that GRD converged approximately within the first ten iterations. In addition to convergence, we also tested the robustness of GRD under random initializations of $w(0)$. Over 500 random initializations for an arbitrary synthetic dataset, GRD correctly assigned weights to only ROIs of the primary group network for 99% of the cases with negligible weights assigned to the secondary group network for the remaining cases. GRD thus appears highly robust to choice of initialization. Also, we tested the case where outlier subjects were present. Specifically, we generated another 500 synthetic datasets, where all subjects’ network structure remained the same as in Figure 4.1(k), except for subjects 9 and 10. For subjects 9 and 10, random noise was used as time courses for ROIs 1 to 20, and ROIs 21 to 30 were made mutually correlated. Weights obtained using GRD for exemplar subjects 1 and 9 are shown in Figure 4.3.
Figure 4.3  Outlier analysis. (a) and (b) Average weight vectors of subjects 1 and 9 over 500 datasets using GRD. Networks of subjects 1 to 8 were the same as in Fig. 1(k). For subjects 9 and 10, ROIs 1 to 20 were random noise, and ROIs 21 to 30 were mutually correlated. Notice negligible weights were assigned to ROIs 1 to 10 for subject 9, thus demonstrating that GRD would not falsely assign high weights to ROIs that do not form a network. Weight vectors of subjects 2 to 8 were similar to subject 1. Weight vector of subject 10 was similar to subject 9. Identifying subjects 9 and 10 as outliers is thus trivial.

For subjects 1 to 8, substantially higher weights were assigned to ROIs of the primary group network than ROIs 21 to 30, and conversely, more weights were assigned to ROIs 21 to 30 than ROIs of the primary group network for subjects 9 and 10. These results thus demonstrate that GRD would not falsely assign high weights to ROIs that do not form a network. Also, based on these differences in weight vectors, subjects 9 and 10 could be trivially identified as outliers.

Furthermore, we tested the robustness of GRD to the choice of $\lambda$ (Figure 4.4). We observed that setting $\lambda$ beyond 0.15 leads to violation of the Armijo condition. In fact, some elements of $W_{RD}(k)+\lambda\delta(k)$ became negative, thus RD was not be able to project the weight vectors back onto the standard simplex. When $\lambda$ is set below 0.05, the sparsity patterns of the weight vectors no longer matched across subjects. Within the range of 0.05 and 0.15, where the Armijo condition is satisfied, GRD was found to be very robust.
Figure 4.4 Effects of varying $\lambda$. $w_i'$s for an exemplar synthetic dataset of scenario 1. Weight values of the first 40 ROIs displayed. Each row of a subfigure corresponds to an ROI and each column corresponds to a subject. (a) Setting $\lambda$ to 0.2 resulted in violation of the Armijo condition and a few negative weights due to too large a step taken during group entropy minimization. (b)-(d) Setting $\lambda$ between 0.05 and 0.15 generated weight vectors with sparsity patterns that exactly match the ground truth (Figure 4.1(k)). Varying $\lambda$ resulted in mild changes to the magnitude of the weights. (e) When $\lambda$ was set to 0.01, the level of group influence appeared too low to steer each subject’s RD process towards the common group network. (f) Setting $\lambda$ to 0 generated weight patterns identical to those obtained by applying RD separately to each subject as expected.

In scenario 2, we tested how well each method scales up to larger size problems by generating 500 synthetic datasets similar to Figure 4.1(k) but with 1000 ROIs (i.e. 1% of the ROIs belong to the primary group network). The results were exactly the same as in Figure 4.1 with GRD correctly detecting only those ROIs belonging to the primary group network, and the other methods prone to subject-specific network structure. GRD could thus be scaled up to whole-brain analysis if we parcellate the brain into thousands of ROIs using automated ROI extraction tools, such as Freesurfer. In fact, GRD could theoretically handle problems of arbitrary size, but memory size limits GRD’s scalability in practice. Results for a smaller scale synthetic test (typical for clinical studies) can be found in [P15].
4.4 Real Data Analysis

4.4.1 Materials

Functional MRI data were provided by Dr. McKeown of the Pacific Parkinson Research Center. 9 healthy subjects, 6 patients with tremor-dominant Parkinson’s disease (PDt) off medication, and 8 patients with akinetic-rigid Parkinson’s disease (PDar) off medication were recruited. Subjects were required to use their right-hands to squeeze a bulb with sufficient pressure such that a horizontal bar shown on a screen was kept within an undulating pathway. The pathway remained straight during rest periods and became sinusoidal at a frequency of 1 Hz during time of stimulus. Increasing level of noise was added to the pathway for studying network changes as the task transitions from being externally-guided (guided by visual stimulus) to being internally-guided (guided by subjects’ memory). We denote the three examined noise levels as A, B, and C with level A having no noise and level C being the noisiest. Each run lasted 600s, alternating between stimulus and rest of 100 s durations.

Functional MRI was performed on a Philips Achieva 3.0 T scanner (Philips, Best, Netherlands). T2*-weighted images with BOLD contrast were acquired using an EPI sequence with an echo time of 37 ms, a repetition time of 1985 ms, a flip angle of 90°, an in plane resolution of 128×128 pixels, and a pixel size of 1.9×1.9 mm. Each volume consisted of 36 axial slices of 3 mm thickness with a 1 mm gap. A 3D T1-weighted image consisting of 170 axial slices was also acquired.

Each subject’s fMRI data were pre-processed using Brain Voyager’s (Brain Innovation B.V.) trilinear interpolation for 3D motion correction and sinc interpolation for slice time correction. Further motion correction was performed using motion corrected ICA [144]. The voxel time courses were then high-pass filtered to account for temporal drifts. No spatial normalization or spatial smoothing was performed. In this study, our main objective was to identify differences in functional networks between the two PD subtypes, namely PDt and PDar. We had thus focused our analysis on ROIs that are pertinent to PD, as chosen by a neurology expert. ROIs included bilateral cerebellum (CER), thalamus (THA), caudate
(CAU), putamen (PUT), pallidum (PAL), hippocampus (HIP), amygdale (AMY), primary motor area (PMA), premotor area (PM), somatosensory area (SSA), caudal anterior cingulate cortex (ACCc), posterior cingulate cortex (PCC), and rostal anterior cingulate cortex (ACCr). All ROIs were manually segmented by a trained technician who was blind to each subject’s diagnosis.

4.4.2 Results and Discussion

The primary, secondary, and tertiary group networks detected using GRD on real data are summarized in Figure 4.5. GRD converged in all cases with all found networks being significant at a p-value of 0.05 with Bonferroni correction. For noiseless case (i.e. “noise level A”), bilateral PMA, bilateral PM, left SSA, and left PCC were detected as the primary group network of the healthy subjects. As we increased the noise (i.e. levels B and C), the right SSA and right PCC were additionally recruited. Since ipsilateral regions are often recruited to handle increasing task difficulty [178], our results are consistent with prior neuroscience knowledge. Removing edges between the ROIs detected as the primary group network and reapplying GRD resulted in bilateral CER, bilateral THA, right PM, and right PCC being detected as the secondary group network, and bilateral CAU, left PM, right SSA, bilateral ACCc and right PCC as the tertiary group network for noise level A. Since the cerebello-thalamo-cortical (CTC) loop is known to be more involved than the striato-thalamo-cortical (STC) loop during externally-guided tasks [179], these results further confirm the validity of GRD. Little changes were observed in the secondary and tertiary networks at noise level B with left CAU becoming part of the secondary network and bilateral PUT recruited into the tertiary network. At noise level C, bilateral CAU, right PM, and bilateral ACCc became the secondary network, and bilateral CER, bilateral THA, left CAU, and left PM became the tertiary network. Since noise level C requires higher memory demands from the subjects, observing this switch in the level of recruitment of the CTC and STC loops again matches expectation.
Figure 4.5 Real fMRI data results. The primary, secondary, and tertiary group networks detected using GRD and their p-values are shown. All networks were statistically significant at p-value of 0.05 with Bonferroni correction.
For PD subjects, similar networks were detected for all noise levels. We thus combined the data of the three noise levels by concatenating the ROI time courses in detecting representative networks for each PD subgroup. Although the primary group network of both PD subgroups comprised the same ROIs, namely bilateral PMA, bilateral PM, and left SSA, the relative weighting of the ROIs were found to be significantly different with a p-value of 0.0017. The secondary and tertiary group networks were also found to be vastly different (p-value = 0 over 100,000 permutations). Specifically, PDt recruited right PM, right SSA, bilateral ACCc and bilateral PCC as the secondary group network, and bilateral CAU, bilateral PUT, left PM, right SSA, and right ACCr as the tertiary group network. In contrast, PDar recruited bilateral cerebellum, right PM, and bilateral PCC as secondary group network, and bilateral THA, bilateral CAU, left PM, and left PCC as the tertiary group network. Our results thus suggest that differences in symptoms between PDt and PDar may originate from brain network differences, in addition to abnormal alterations in brain activity [180, 181]. In particular, excessive activation in the cerebellum has been hypothesized as part of a compensatory mechanism [180], which was found to be more pronounced in akinetic-rigid PD [181]. In support of these findings, our results further suggest that differences between the PD subtypes may be signified by the presence of hyper-active CTC loop in PDar. Since PDar normally portends a worse prognosis, we postulate that an over-active CTC loop may be maladaptive.

We note that we have also applied GRD to real data described in Section 2.7.1 to examine the effects of increasing task difficulties. We found that even at the lowest task frequency, PD subjects off medication were required to recruit a network that was employed by healthy subjects at the highest frequency. Details of this analysis can be found in [P15].
4.5 Summary

In this chapter, we presented a novel group-wise sparse network detection method that we proposed for handling some of the limitations associated with state-of-the-art functional connectivity analysis techniques, as summarized in Table 4.1.

Table 4.1 Summary of functional connectivity analysis methods. GA = group analysis.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multivariate</td>
<td>Extendable to GA</td>
</tr>
<tr>
<td>Seed-based [86]</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Clustering [100, 101]</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Decomposition [34, 73]</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

In brief, the standard seed-based approach is easily extendable to group analysis, but does not model joint voxel interactions and is sensitive to the choice of threshold in isolating the network nodes. Cluster analysis accounts for joint voxel interactions, but is not easily extendable to group analysis and is sensitive to the number of clusters chosen. Decomposition techniques, such as PCA and ICA, are multivariate, but their diffused weighting property renders identification of network nodes difficult. Moreover, drawing a correspondence between spatial component maps that are separately extracted from each subject is non-trivial due to the large inter-subject variability. Group extensions of ICA have been proposed, but assume all subjects share similar spatial components for both signal and noise. Thus ideally, one would want a multivariate method that facilitates simple network identification, while having some mechanisms for handling the large inter-subject variability. To jointly satisfy these criteria, we proposed incorporating sparsity and group priors into PCA for network detection. Specifically, we showed that RD is a solution of the NSPCA optimization problem. Thus, exploiting the sparse property of RD can simplify network identification. We further showed that incorporating a group prior provides an effective means for handling the large inter-subject variability with network correspondence implicitly drawn across subjects. We note that GRD extracts networks in a hierarchical manner, and stops when the extracted network has a correlation value below that of random noise. Hence, GRD does not assume nor require all subjects to have similar spatial noise structures.
5 Generalized Sparse Classification

In fMRI classification, one is often faced with exceedingly more features (e.g. voxels) than samples (e.g. brain volumes). Under such ill-conditioned settings, direct application of standard classifiers will likely result in overfitting. The standard approach to deal with the curse of dimensionality is to perform univariate feature selection, but this strategy ignores the collective information encoded in the voxel signal patterns. A recent approach that has gained major research interest is to incorporate sparsity into classifier learning. Enforcing sparsity implicitly removes classifier weights associated with irrelevant features, i.e. shrinks them to exactly zero. In effect, this approach enables feature selection and classifier learning to be simultaneous and collaboratively performed.

The majority of state-of-the-art sparse linear models [182-188] (Section 5.1) are based on the seminal work by Tibshirani [127], in which he showed that the LASSO penalty induces sparsity. These models provide effective means for imposing structural constraints on the model parameters [184], but lack the flexibility for incorporating potentially advantageous problem-specific properties beyond sparsity [187]. For example, adjacent pixels in an image are typically correlated. Model parameters associated with these pixels should thus be assigned similar magnitudes to reflect the underlying correlations. However, merely enforcing sparsity does not model such associations between model parameters. To encourage smoothness in model parameters, in addition to sparsity, Tibshirani et al. proposed the fused LASSO model [188], which combines the LASSO penalty with a term that penalizes the $l_1$ norm of the differences between successive model parameters. Similarly, Hebiri proposed the smooth LASSO [189], in which a term that penalizes the $l_2$ norm of the differences between successive model parameters is added to the LASSO penalty. To extend beyond smoothness, Tibshirani and Taylor recently proposed the Generalized LASSO model [187], which penalizes the $l_1$ norm of a weighted combination of the model parameters. By varying the choice of weights, Generalized LASSO facilitates numerous applications, such as trend filtering, wavelet smoothing, and outlier detection [187].
To enable greater modeling flexibility without creating overly complex optimization problems, we propose a simple technique, “generalized sparse regularization” (GSR) [P17] that enables incorporation of prior knowledge [P17-P20] into a wide collection of sparse linear models (Section 5.2). In contrast to [187], we show that GSR is applicable to a much broader set of sparse linear models than just the LASSO regression model, and can be easily extended to the classification settings (Section 5.3) for applications, such as brain decoding (Section 5.4), which is our primary interest. The adaptability of GSR to such a wide range of models stems from how any $l_2$ norm penalty can be merged into an $l_2$ data fitting loss through a simple augmentation trick. Thus, adding a generalized ridge penalty to any sparse linear models with an $l_2$ data fitting loss enables problem-specific properties to be easily integrated while preserving the functional form of the original optimization problem that these models entail. This desirable property of GSR facilitates direct deployment of existing sparse optimizers [152, 190-192]. We built various prior-informed sparse classifiers using GSR and applied them to the publicly available StarPlus data (Section 5.5) to demonstrate the benefits of modeling application-specific properties beyond sparsity (Section 5.6).

5.1 Overview of Sparse Linear Models

In this section, we focus on the problem of regression, since most sparse linear models [182-188] are inherently designed for such application. We defer discussion of how the sparse linear models described in this section can be employed for classifier learning in Section 0. Consider the standard least square regression problem:

$$\hat{a} = \min_{a} \| y - Xa \|_2^2,$$  \hspace{1cm} (5.1)

where $y$ is an $N \times 1$ response vector, $X$ is an $N \times M$ predictor matrix, $a$ is an $M \times 1$ coefficient vector, $N$ is the number of samples, and $M$ is the number of predictors. The closed-form solution of (5.1) is given by:
\[
\hat{a} = \left( X^T X \right)^{-1} X^T y. \tag{5.2}
\]

When \( N << M \), \((X^T X)^{-1}\) is ill-conditioned. Thus, direct estimation of \( \hat{a} \) using (5.2) usually results in overfitting. To obtain a more generalizable estimate of \( \hat{a} \), a common strategy is to employ regularization. In particular, Tibshirani proposed enforcing sparsity on \( a \) to achieve the dual objective of reducing overfitting and enhancing interpretability [127]:

\[
\hat{a} = \min_{a} \left\| y - Xa \right\|_2^2 + \alpha \left\| a \right\|_1, \tag{5.3}
\]

where \( \alpha \geq 0 \). The model (5.3) is commonly referred to as the LASSO regression model, where Tibshirani showed that penalizing the \( l_1 \) norm induces sparsity on \( a \) [127].

The success of the LASSO regression model in numerous applications resulted in an explosion of research on sparse linear models [192]. Although many LASSO variants involve only a simple addition of other penalty terms to (5.3), the modeling power that these extensions facilitate proved substantial. For example, Zou and Hastie proposed adding a ridge penalty to (5.3), which is known as the elastic net model [182]:

\[
\hat{a} = \min_{a} \left\| y - Xa \right\|_2^2 + \beta \| a \|_2^2 + \alpha \| a \|_1, \tag{5.4}
\]

where \( \beta \geq 0 \). This model has two key advantages over LASSO. First, in contrast to LASSO, the number of non-zero elements in \( a \) is no longer limited by the number of samples [182]. This property is especially important in medical imaging applications, since the number of samples is often much smaller than the number of predictors. Second, in cases where the predictors are correlated, elastic net tends to jointly select the correlated predictors, whereas LASSO would arbitrarily select only one predictor among each correlated set [182].

Another widely-used extension of LASSO is the group LASSO, proposed by Yuan and Lin for applications where a natural grouping of the predictors is present [183]:

\[
\]
where \( a_{gh} \) are the coefficients associated with predictors belonging to the predefined group \( g_h \), and \( h \in \{1, \ldots, H\} \) where \( H \) is the number of predefined groups \[183\]. As evident from (5.6), \( \|a_g\|_2,1 \) is exactly the \( l_1 \) norm of \( \|a_{gh}\|_2 \). Thus, minimizing \( \|a_g\|_2,1 \) encourages sparse subsets of groups to be selected with non-zero coefficients assigned to all predictors within each selected group. However, since not all predictors within a group are necessarily relevant, Sprechmann et al. \[184\] proposed the hierarchical LASSO model, also known as sparse group LASSO \[185\]:

\[
\hat{a} = \min_a \|y - Xa\|_2^2 + \alpha \|a_g\|_{2,1} + \beta \|a\|_1,
\]

which is essentially group LASSO combined with a LASSO penalty. Reintroducing the LASSO penalty encourages internal sparsity such that only a sparse subset of predictors within each selected group is chosen \[184\]. This property of (5.7) is particularly useful in cases where one is uncertain about the exact group assignment.

To extend (5.7) to scenarios where multiple sets of samples, \( y^s \), are generated from different processes, e.g. when observations are collected from multiple subjects \[193\], but associated with the same set of predictors, e.g. using the same set of semantic features to predict brain activation patterns of different subjects \[193\], Sprechmann et al. \[184\] proposed the collaborative hierarchical LASSO model:

\[
\hat{a} = \min_a \|Y - XA\|_F^2 + \alpha \sum_{k=1}^K \|a_{g_k}\|_F + \beta \sum_{s=1}^S \|a^s\|_1
\]
where $S$ is the number of processes, $Y = (y^1, \ldots, y^S)$, $A = (a^1, \ldots, a^S)$, $A_{gh}$ are the rows of $A$ belonging to group $g_h$, and $a^s$ is the coefficient vector of process $s$. $\| \cdot \|_F$ denotes the Frobenius norm. Minimizing (5.8) promotes the same groups to be selected across processes with the chosen sparse subset of predictors within each selected group being potentially different [184]. This model thus enables information across processes to be shared while providing flexibility in handling inter-process variability by allowing sparsity patterns within the selected groups to vary across processes.

All of the LASSO extensions described above provide effective means for modeling data structure through promoting groups of predictors to be jointly selected. However, the associations between predictors are ignored in these models, i.e. associated predictors may be jointly selected but the coefficients assigned to these predictors can substantially vary. Since the associations between predictors can be an important source of information for further constraining the ill-conditioned problem (5.1), we thus propose a simple approach to incorporate predictor associations into all of the models above, as discussed next.

### 5.2 Generalized Sparse Regularization

To integrate properties beyond sparsity into sparse linear models as those described in Section 5.1, we propose complementing these models with the following penalty:

$$J_{GSR}(a) = \| \Gamma d \|_2^2,$$

where $\Gamma = (\gamma_1, \ldots, \gamma_R)^T$ is an $R \times M$ penalty matrix for modeling the associations between predictors, $\gamma_r$ is an $M \times 1$ vector encoding the penalty for each association $r$, and $R$ is the number of associations being modeled. For instance, if predictors $x_p$ and $x_q$ are associated with each other, say by virtue of being spatially adjacent voxels, then $a_p$ and $a_q$ should presumably be assigned similar values to reflect this intrinsic association. This can be accomplished by, e.g. setting $\gamma_{rp}$ to 1 and $\gamma_{rq}$ to $-1$ such that differences in $a_p$ and $a_q$ are
penalized. Many other context-specific properties can be analogously modeled by varying $\Gamma$ as discussed in Section 5.4 and [187].

One may envisage other penalties to facilitate similar modeling flexibility. However, not all penalties added to the models in Section 5.1 will result in a practical optimization problem. The critical question is thus whether the optimization problem resulting from integrating (5.9) into models in Section 5.1 can be efficiently minimized. To address this question, we draw upon a basic property of the Euclidean norm:

$$\sum_{\omega=1}^{\Omega} \|z_{\omega}\|_2^2 = \|z\|_2^2, \quad z = \begin{pmatrix} z_1 \\ \vdots \\ z_\Omega \end{pmatrix}, \quad (5.10)$$

where $z_{\omega}$ is a vector of arbitrary length. Since models (5.3), (5.4), (5.5) and (5.7) in combination with (5.10), can all be written in the form:

$$\hat{a} = \min_a \|y - Xa\|_2^2 + \lambda \|\Gamma a\|_2^2 + J(a), \quad (5.11)$$

where $\lambda \geq 0$ and $J(a)$ is the sparse penalty term of the respective models, invoking (5.10) on (5.11) results in the following optimization problem:

$$\hat{a} = \sqrt{1 + \lambda} \min_{\tilde{a}} \|\tilde{y} - \tilde{X}\tilde{a}\|_2^2 + J(\tilde{a}), \quad (5.12)$$

$$\tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \tilde{X} = (1 + \lambda)^{-\frac{1}{2}} \begin{pmatrix} X \\ \sqrt{\lambda} \Gamma \end{pmatrix}, \quad (5.13)$$

which has the exact same functional form as the optimization problems of the original sparse linear models. Thus, existing solvers of the respective sparse linear models [152, 190-192] can be directly employed to efficiently minimize (5.11).
For the multi-process model (5.8), if the associations between predictors are similar across processes, then the same matrix augmentation trick can be applied with the zero vector in (5.13) replaced by an $R \times S$ zero matrix. If associations between predictors vary across processes, a minor modification is needed:

$$
\tilde{y} = \left( \text{vector}(Y) \right), \quad \tilde{X} = (1 + \lambda)^{-\frac{1}{2}} \begin{pmatrix} X & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & X \end{pmatrix},
$$

where vector(·) is an operator that stacks the columns of its argument into a vector, and $a$ in the first term of (5.11) is now vector($A$). We highlight that, in addition to enabling different predictor associations to be modeled for each process, (5.14) permits modeling of predictor associations across processes. This modification of the matrix augmentation trick thus builds even greater flexibility into (5.8).

### 5.3 GSR for Classification

The models described in Section 5.1 are inherently designed for regression problems with $y$ being a continuous variable. To extend the GSR-augmented sparse regression models to the classification setting without altering the functional form of the associated optimization problems, we employ the spectral regression technique [194], which comprises two steps:

**Step 1.** Learn the constraint-free optimal projection of the training data $X$, e.g. using graph embedding (GE) [195]:

$$
Wy = \lambda Dy,
$$

where $y$ is the projection of $X$ on the subspace defined by $W$ [195]. $W_{ij}$ models the intrinsic relationships between samples $i$ and $j$ of $X$, and $D$ is a diagonal matrix with $D_{ii} = \sum W_{ij}$. We
note that the key advantage of GE is that it enables various subspace learning algorithms to be used as classifiers by simply varying $W$ [195].

**Step 2.** Determine the classifier weights, $a$, such that $y$ and $Xa$ are as similar as possible under the desired constraints, $Q(a)$:

$$
\hat{a} = \min_a \| y - Xa \|_2^2 + Q(a).
$$

(5.16)

Clearly, setting $Q(a)$ to $\| y \|_2^2 + J(a)$ results in our proposed GSR model (5.11). Hence, exploiting spectral regression in combination with GSR facilitates sparsity and predictor associations to be jointly integrated into classifier learning.

### 5.4 Application to fMRI Spatiotemporal Classification

Given $N M \times 1$ feature vectors, $x_i$, forming the rows of an $N \times M$ matrix, $X$, the general classification problem involves finding the corresponding $N \times 1$ label vector, $l$, containing the class label $l_i$ of $x_i$. We assume here that each feature $x_{ip}$ can be naturally assigned to a group $g_h \in G = \{ g_1, \ldots, g_H \}$. In the context of spatiotemporal fMRI classification, we treat the signal intensity of each brain voxel $p$ at time $t_k$ within a trial as a feature, and all brain volumes within a trial of the same experimental condition as a sample $x_i$, as illustrated in Figure 5.1. Our goal is to determine to which condition, $l_i$, does each concatenated set of brain volumes $x_i$ belongs. Since the brain is known to be functionally organized into specialized neural regions [196], this provides a natural grouping of the voxels.
Figure 5.1  Predictor matrix. Brain volumes within the same trial are concatenated and treated as a single sample $x_i$. $N_i$ is the number of volumes within a trial and $N_T$ is the number of trials for a particular experimental condition.

To model this modular structure of the human brain, as well as the intrinsic spatiotemporal correlations in brain activity, we build an anatomically-informed spatiotemporally smooth sparse LDA (ASTSLDA) classifier by first solving for $y$ in (5.15) with:

$$W_{ij} = \begin{cases} 1/m_c, & l_i = l_j = c \\ 0, & \text{otherwise} \end{cases}, \quad (5.17)$$

where $m_c$ is the number of samples in class $c$ [194]. We then optimize (5.11) with $J(a)$ set to the group LASSO penalty (6), and the signal intensity of the voxels within each ROI (Section 5.5) at each time point $t_k$ treated as a group. $\Gamma$ is set as the spatiotemporal Laplacian operator:

$$\Gamma_{p_k q_s} = \begin{cases} -1, q_s \in N_{p_k} \\ 0, \text{otherwise} \end{cases}, \quad \Gamma_{p_k p_k} = -\sum_{q_s \neq p_k} \Gamma_{p_k q_s}, \quad (5.18)$$

where $N_{p_k}$ is the spatiotemporal neighborhood of voxel $p$ at time $t_k$, as shown in Figure 5.2. Specifically, 6-connected spatial neighbors and signal intensity of voxel $p$ itself at adjacent time points are defined as the spatiotemporal neighbors.
We can easily build other sparse LDA variants in an analogous manner. To build an anatomically-informed spatially smooth sparse LDA (ASSLDA) classifier, we employ (5.18) but with $N_{pk}$ being the 6-connected spatial neighborhood of voxel $p$. To build a sparse LDA classifier that is only anatomically-informed (ASLDA), we set $\lambda$ in (5.11) to 0 with $J(a)$ being the group LASSO penalty (5.6). Voxel-level sparse LDA classifiers that incorporates a spatiotemporal prior (STSLDA), a spatial prior (SSLDA), and no prior (SLDA) can be built in a similar manner as their anatomically-informed counterparts, but with $J(a)$ in (5.11) being the LASSO penalty (5.3). LDA classifier with elastic net penalty (ENLDA) can also be built from (5.11) by setting $\Gamma$ to identity and $J(a)$ as the LASSO penalty (5.3). Our proposed model (5.11) thus encompasses many of the widely-used sparse linear models.

5.5 Materials

The publicly available StarPlus database [117] was used for validation. We provide a brief description of the data and the experimental setup for convenience. Details can be found in [117]. In the StarPlus experiment, all subjects performed 40 trials of a sentence/picture matching task. In each trial, subjects were required to look at a picture (or sentence) followed by a sentence (or picture), and then decide whether the sentence (picture) correctly described the picture (sentence). The first stimulus was presented for 4 s followed by a blank screen for 4 s. The second stimulus was then presented for up to 4 s followed by a 15 s rest period. In half of the trials, the picture preceded the sentence, and vice versa.
Functional MRI brain volumes were acquired from 13 normal subjects at a TR of 500 ms, but only 6 of the subjects’ data are available online. Each subject’s dataset comprised voxel time courses within 25 ROIs, resulting in approximately 5000 voxels per subject. Inter-subject differences in the number of voxels were due to anatomical variability. Motion-correction and temporal detrending were applied on the voxel time courses to account for head motions and low frequency signal drifts. No spatial normalization was performed.

5.6 Results and Discussion

To examine the benefits of incorporating a spatiotemporal prior in addition to enforcing sparsity, we treated the signal intensity of each voxel at each time point within a trial as a feature, and all brain volumes within the same trial of each experimental condition as a sample. The classification task was thus to discriminate sets of brain volumes associated with a picture from those associated with a sentence. To account for the delay in the HRF [24], we only used the 8 brain volumes collected 4 s after stimulus onset within each trial. Each sample thus consisted of approximately 40,000 features (i.e. roughly 5000 voxels × 8 volumes, see Section 3) with 40 samples per class (i.e. 40 trials per class) for each subject. For comparison, we contrasted ASTSLDA against LDA, linear SVM, SLDA, ENLDA, SSLDA, ASLDA, and ASSLDA. We employed the spectral projected gradient technique (SPG) [191] to minimize the optimization problem of the respective classifier models. SPG was chosen for its ability to efficiently handle large-scale problems (classifiers with ~40,000 coefficients in this study). All contrasted classifiers could be learned within a few seconds using SPG for a fixed set of parameter values $\lambda$ and $\alpha$ ($\beta$ was not involved in the classifier models contrasted in this work). Ten-fold nested cross validation was used [117] to select $\lambda$ and $\alpha$, and estimate the prediction accuracy. The average prediction accuracies over subjects are shown in Figure 5.3. An axial slice of the classifier weight patterns corresponding to 5.5 s and 6 s after stimulus onset (i.e. when HRF typically peaks) for 3 exemplar subjects are shown in Figure 5.4 for result interpretations.
Figure 5.3  Prediction accuracy comparisons. The p-values obtained by statistically comparing ASTSLDA against the other methods are indicated. Significance was declared at a p-value threshold of 0.05. ASTSLDA significantly outperformed the standard classifiers, LDA and SVM. The improvement over the other existing methods, namely SLDA, ENLDA, and ASLDA, were close to significant.

LDA resulted in the worst average prediction accuracy, which was likely due to overfitting. The classifier weight patterns also appeared spatially-random. Using linear SVM led to similar predictive performance and randomly-distributed weight patterns. Reducing overfitting using SLDA increased accuracy, but the weight patterns seemed overly-sparse. The reason was likely due to LASSO’s constraint on the number of non-zero elements in the classifier weight vector $\alpha$, i.e. restricted by the sample size [2]. In fact, since we treated the signal intensity of the brain volumes within a trial as a single feature vector, weights would be spread across time, resulting in even sparser spatial patterns than what would be obtained in the conventional case where each brain volume is taken as a sample. Also, we observed little consistency in the SLDA weight patterns across subjects and time.
Figure 5.4 Classifier weight patterns. Weight patterns 5.5 s and 6 s after stimulus onset for 3 exemplar subjects. “*” = dorsolateral prefrontal cortex. “+” = temporal lobe. “^” = visual cortex. (a) LDA and (b) SVM resulted in randomly-distributed patterns. (c) SLDA generated overly-sparse patterns, which was mildly improved with (d) EN-LDA. (e) SSLDA and (f) STSLDA produced spatially smoother patterns. (g) ASLDA and (h) ASSLDA weight patterns displayed little consistency across time and subjects. (i) ASTSLDA patterns were more consistent across subjects and time compared to the contrasted classifiers.
Alleviating LASSO’s constraint on the number of non-zero elements in $a$ using ENLDA [182] improved prediction accuracy over SLDA. However, the weight patterns remained overly-sparse, which demonstrate the highly-overlooked fact that higher prediction accuracy does not necessarily translate to more neurologically-sensible weight patterns. In contrast, modeling spatial correlations using SSLDA resulted in higher prediction accuracy and spatially smoother weight patterns. Our results thus illustrate the added-value of incorporating prior knowledge, both in terms of predictive performance and result interpretability. These benefits of exploiting prior knowledge were further exemplified using STSLDA, which obtained similar spatially smooth patterns but a further increase in accuracy. This increase likely arose from how STSLDA pooled information across brain volumes through penalizing discrepancies in classifier weights between spatiotemporal neighbors.

Accounting for the modular structure of the human brain using ASLDA improved prediction accuracy over the non-anatomically-informed classifiers except STSLDA, which again demonstrate the importance of constraining the model learning problem with prior knowledge (i.e. through grouping of associated features) to handle the curse of dimensionality. However, the weight patterns displayed little consistency between subjects and across time points. Modeling spatial correlations using ASSLDA resulted in higher accuracy than ASLDA, but it too obtained (slightly) lower accuracy than STSLDA. These results further support that exploiting the commonality between adjacent brain volumes, in addition to modeling spatial correlations, can greatly improve predictive performance.

Modeling both the modular property and spatiotemporal characteristics of brain activity using ASTSLDA resulted in the best overall predictive performance with an average accuracy of 91%, which is an 11% increase over LDA and SVM. ASTSLDA also achieved the lowest variability in prediction accuracy, thus demonstrating stable performance despite the considerable inter-subject variability seen in fMRI studies. Applying paired t-tests to statistically compare ASTSLDA against the other methods (Figure 5.3) showed that the increases in accuracy over the standard classifiers, LDA and SVM, were significant at a p-value threshold of $0.05$. The accuracy increases over the other existing methods, namely SLDA, ENLDA, and ASLDA, were also close to significant. Provided that ASTSLDA
maintains the observed stability in performance, hence achieves similar level of accuracy for new subjects, we expect that the improvements over SLDA, ENLDA, and ASLDA would become statistically significant if more subjects are available [132]. We note that if we pool all samples together and ignore the factor of subject, we can assess whether two methods’ accuracy levels are significantly different using the McNemar’s test, which accounts for how accuracy levels are not normally-distributed. Qualitatively, higher consistency in weight patterns was observed across both subjects and time, with weights correctly assigned to the dorsolateral prefrontal cortex, which is responsible for verification task (e.g. decide if a sentence matches a picture) [197], as well as the visual cortex and the temporal lobe, which pertain to picture/sentence discrimination [198].
5.7 Summary

The main challenge in fMRI classification is the need to learn very high dimensional classifiers given few noisy samples. The state-of-the-art feature selection methods that are commonly-used to handle this curse-of-dimensionality problem are summarized in Table 5.1.

Table 5.1 Summary of feature selection approaches for fMRI classification.

<table>
<thead>
<tr>
<th>Feature Selection Approaches</th>
<th>Multivariate</th>
<th>Data Structure</th>
<th>Feature Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate [117]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multivariate [125, 126]</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sparsity [120]</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structured Sparsity [119, 121, 129]</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Selecting features using univariate techniques ignores the collective information encoded by the voxel signal patterns. Using multivariate feature selection techniques can alleviate this limitation, but these techniques do not model the structure within the data, e.g. adjacent voxels might not be jointly selected. Enforcing sparsity enables classifier learning and feature selection to be simultaneously and collaboratively performed, but again does not model data structure. Imposing structured sparsity penalties, such as group LASSO and elastic nets, account for data structure, but not feature associations, e.g. adjacent voxels might be jointly selected but might not be assigned similar weights. In fact, negligence of feature associations in classifier learning appears to be a common limitation to most existing methods for fMRI classification. Therefore, we proposed a simple, effective regularization technique that facilitates modeling of feature associations in addition to structured sparsity during classifier learning. We showed that adding sparsity, anatomical, and spatiotemporal priors to constrain the ill-conditioned fMRI classifier learning problem can substantially improve predictive accuracy over standard classifiers. Moreover, we demonstrated that incorporating relevant priors results in more neurologically-plausible classifier weight patterns.
6 Conclusions

We presented in this thesis our technical contributions in the area of fMRI analysis. Our methods were based on novel incorporation of various pertinent priors to address key challenges facing current state-of-the-art approaches. We conclude here with a discussion of the strengths and weaknesses of our methods and some promising directions to explore.

6.1 Activation Detection

The standard univariate approach completely ignores voxel interactions despite that such interactions may encode additional information for identifying the activated voxels. Current techniques that impose spatial priors account for local voxel interactions, but not the longer range interactions. We thus proposed modeling both types of interactions using RW [P1] and ANSPCA [P2], and showed that such information can improve detection sensitivity at the expense of potentially including noise-induced false correlations. This limitation was especially evident during synthetic testing at low SNR, hence suggesting that the extra information provided by voxel correlations estimated from the same dataset as that used for generating activation statistics might be limited. We thus proposed GMRF [P3, P5] to incorporate group information into intra-subject activation detection with improved detection sensitivity and stronger false positives control over RW shown. Also, GMRF detected a spatial focusing effect of L-dopa on the brain activation within the left M1 that was consistent across all PD subjects examined.

In addition to GMRF, we proposed GRW [P4] to integrate functional connectivity and group information. The activation detection results were similar to those obtained with GMRF, but as opposed to generating only binary output (i.e. active or non-active), GRW additionally provides posterior activation probabilities. From an interpretation perspective, having probability maps enables the experimenters to assess the uncertainty in the results. From a spatial pattern analysis point of view, since the probability maps are essentially “refined”
versions of the corresponding activation statistics maps, the spatial distribution of the activation probabilities would also encode important task-pertinent information, which can be captured using our proposed spatial characterization approach [P1]. A drawback to both GMRF and GRW is scalability. Having to store and analyze all subjects’ data within a group simultaneously would not scale well to databases with hundreds of subjects. Hence, GMRF and GRW are more suitable for clinical studies, where only ten to twenty subjects are typically recruited. For larger studies, some methods for sub-grouping the subjects based on demographic data or functional similarities will be needed.

Another complication to using GMRF and GRW, as well as RW and ANSPCA, is the need to “properly” balance the degree of influence between the activation statistics and the connectivity and group priors in delineating the activated voxels. Currently, we set all functional attributes on equal footings, but learning the relative weighting from data would be advantageous. One rigorous way to determine this weighting is to define a generative model and choose the weighting based on maximum model evidence. We thus proposed a connectivity-informed generative activation model [P6] that permits exact closed-form solutions for both the posterior activation effects estimates and the model evidence. We validated our model on a large cohort of 65 subjects and demonstrated that integrating a RS-connectivity prior improves sensitivity in detecting group activation. Our results thus support that the correlation structure of task activation effects is pertinent to RS connectivity, and this relationship is common across subjects. The flexibility of this model permits integration of other priors, such as group priors as used in GMRF and GRW and anatomical connectivity priors learned from diffusion MRI data, which would be promising directions to explore. Also, it would be interesting to add RS-connectivity prior to RW, ANSPCA, GMRF, and GRW, and compare results with our connectivity-informed activation model in the future.


6.2 ROI Characterization

Conventional ROI analysis employs features that only capture amplitude changes in activation statistics without accounting for the spatial distribution of regional activation. During my Master studies, we proposed using invariant spatial features to model information encoded in the spatial patterns [81, 82]. In this thesis, we focused on further validating this analysis approach. In particular, we showed over a large cohort of 65 subjects undergoing ten different experimental conditions that brain activation can generally be inferred from the spatial modulations of regional BOLD distributions, as measured with our proposed invariant spatial features [P11].

We also explored the benefits of our spatial characterization approach for analyzing ROI activation statistics [81]. Specifically, we showed over more subjects that using our proposed spatial features provides greater task discriminability than traditional amplitude-based features [P7]. In fact, we detected the same spatial focusing effect of L-dopa on PD subjects’ regional activation [P9] as found with GMRF and GRW, as well as task-difficulty-related changes in the spread of regional activation [P10]. Also, we investigated into the temporal dynamics of spatial response, and found that its peak time antecedes that of BOLD amplitude changes [P12]. Furthermore, we showed that applying our spatial characterization approach [81] to warp-free activation statistics patterns outperformed the same approach applied to spatially normalized data [P8], at least with the registration technique employed [64]. Our results thus suggest that certain information might be lost during spatial normalization. Nevertheless, high dimensional registration techniques, such as HAMMER [62], might better preserve the commonalities in spatial patterns across subjects than the registration technique used in this work. Comparing these high dimensional techniques against warp-free spatial analysis would thus be beneficial in providing further insights into the pro’s and con’s of spatial normalization.

In most of our studies, we concentrated on examining the spatial extent of activation, since this feature is easy to interpret. Nevertheless, there are other spatial phenomena that have
simple physical meanings. In particular, we proposed a spatial feature that measures the skewness in the spatial distribution of regional BOLD distributions [P13]. Our results on real data (Section 2.7.1) suggest that this skewness feature can play a complementary role to more fully describe task-induced spatial changes. An interesting question is which feature is most dominant. We currently examined each of our spatial features separately to infer brain activation. Determining the optimal weightings to combine these spatial features in detecting activation may further improve sensitivity, and provide us an intuition as to which spatial feature is most task-relevant.

A limitation to our spatial characterization approach, in fact ROI analysis in general, is the need to define ROIs. Currently, we define the ROIs based on neuroanatomy, but some of the anatomical ROIs may consist of multiple functional subregions. Thus, extending our prior-informed activation model to divide the anatomical ROIs into functional subregions would be beneficial. Other applications of our spatial characterization approach include inferring functional connectivity from the spatial feature time courses [P14]. Our results demonstrate such feasibility, but the exact benefits and physical interpretation of the results are currently unclear, and thus, in need for further investigations.
6.3 Functional Connectivity Analysis

The standard seed-based approach is limited in the type of networks that may be detected. Using decomposition techniques, such as PCA and ICA, enables functional networks to be identified in an exploratory manner. However, drawing a correspondence between the intra-subject spatial maps derived from PCA and ICA to infer group networks is nontrivial due to the large inter-subject variability seen in fMRI studies. Also, the diffused weighting problem of PCA and ICA renders delineation of network nodes from the spatial maps difficult. We thus proposed GRD [P15, P16] to incorporate group information in identifying common sparse networks across a group of subjects. Enforcing sparsity simplifies the delineation of network nodes, while incorporating group information eases the correspondence problem. Applying GRD to synthetic data correctly detected networks that perfectly matched with ground truth. In contrast, PCA, RD, ICA, and Group ICA were found to be prone to inter-subject variability, demonstrating the effectiveness of GRD in handling subject variability. When applied to real fMRI data, GRD detected task-specific group networks that matched well with prior neuroscience knowledge. Moreover, our results provide new explanations for the observed differences in symptoms between PDt and PDar, suggesting that altered brain networks may be implicative to the PD subtype differences.

A limitation of GRD is the need for storing large correlation matrices. Hence, GRD does not easily extend to voxel level analysis. Also, the memory requirement scales up with the number subjects. Nevertheless, if we instead divide the brain into ~1,000 regions, GRD can easily handle up to ~100 subjects. For future work, it will be interesting to examine whether differences in subject-specific regional weightings are correlated with e.g. diseased severity or genetic factors. Also, applying GRD to diffusion MRI data and comparing the resulting networks with that derived from RS-fMRI data would enable us to further our understanding of the relationships between anatomical and functional connectivity.
6.4 Brain Decoding

In fMRI classification, the number of features typically far exceeds the number of samples. Naively applying standard classifiers will thus result in overfitting. Implicitly selecting features by enforcing sparsity can lead to spurious classifier weight patterns. We thus proposed GSR [P17] to model feature associations, in addition to sparsity. We showed that incorporating spatial [P20], spatiotemporal [P19], and anatomical priors [P18] can enhance predictive performance and generate more neurologically plausible classifier weight patterns. However, drawing a balance between mis-classification errors and the amount of regularization usually involves internal cross validation, which is very computationally demanding. Also, the anatomical prior that we used jointly assigns weights to all voxels within the selected ROIs. However, some larger anatomical ROIs may comprise multiple functional sub-regions, thus dividing these larger ROIs into functional parcels [130] might further improve results. Alternatively, using hierarchical LASSO [184] to prune out the non-relevant voxels within the selected anatomical ROIs would be a promising direction to investigate. Furthermore, it would be interesting to explore other priors, such as RS-connectivity prior and group prior, analogous to our approach taken for activation detection.
References


Assessment with diffusion-tensor MR imaging and parallel imaging techniques,"

synchrony in Parkinson's disease," Parkinsonism Relat. Disord., vol. 13, pp. S440-

[14] S. Thobois, S. Guillouet, and E. Broussolle, "Contributions of PET and SPECT to the
understanding of the pathophysiology of Parkinson's disease," Neurophysiol. Clin.,

oxidative stress in preclinical Alzheimer's disease," Ann. N. Y. Acad. Sci., vol. 1147,


9868-9872, 1990.


[22] V. Calhoun. Preprocessing III & IV: Anatomic transformation to standard brain


