STATISTICAL METHODS FOR

BRAIN IMAGING AND GENOMIC DATA ANALYSIS

by

Juemin Yang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2015

Abstract

Modern technologies for imaging brain activity, such as functional MRI, are very useful in studying mechanisms underlying human brain function and structure. In this thesis, statistical methods and computational techniques are developed to investigate brain function through fMRI. In particular, we are interested in the activation of human brain networks and learning patterns. In the first component of the work, we developed a new group ICA approach, Homotopic Group ICA ("H-gICA"), which seeks to identify networks of correlated regions across subjects and measures the degree of synchrony in spontaneous activity between geometrically corresponding interhemispheric brain regions. H-gICA is able to increase the potential for network discovery and facilitate the investigation of functional homotopy via ICA based networks. In the second part of the work, we developed methodology for an investigation of motor learning using activation distributions. We investigated tests of dimension for detecting learning-based changes, particularly motor learning. Our investigation included a large scale simulation study of brain activation maps, motivated by a study of motor learning in healthy adults. In the third part of the work, we devised an approach to phenotype classification from gene expression profiling. We proposed a new high dimensional discriminant analysis method called group Nearest Shrunken Centroids (gNSC), which enables us to use gene pathway information. We also applied our method on a novel context analysis of association between pathways and certain medical words to improve the power of feature selection.

Readers:

Dr. Brian Caffo (Advisor, JHSPH Biostatistics)

Dr. William Eaton (Chair, JHSPH Mental Health)

Dr. Martin Lindquist (JHSPH Biostatistics)

Dr. James Pekar (SOM Radiology)

To my parents, Xiaoying and Jianrong,
and Wenying, with love.

Acknowledgments

I would like to thank my advisor Brian Caffo for his patience, enthusiasm in research and valuable advices in life. His competence and humbleness inspired me over the past five years and will influence me forever. I am deeply grateful for support and friendship. This work would not have been possible without his help.

I am grateful to the members of my thesis committee – Dr. Brian Caffo, Dr. William Eaton, Dr. Martin Lindquist , Dr. James Pekar – for their support throughout my doctoral program, and constructive comments and suggestions during my doctoral dissertation work.

My special thank to the wonderful faculty and staff members of the Department of Biostatistics and elsewhere at Hopkins. A special gratitude I give to Dr. Ciprian Crainiceanu, Dr. Ani Eloyan, Dr. Vadim Zipunnikov, Dr. Han Liu, Dr. Rafael A. Irizarry, Dr. Daniel O. Scharfstein, and Dr Mei-Cheng Wang for their determination in sharing their expertise without restrictions. Thanks to Dr. Stewart Mostofsky, Dr. James J. Pekar, Dr. Marilyn Albert, Dr. Mary Beth Nebel, Dr. Anita Barber, Dr. Lior Shmuelof, Dr. Luo Xiao, and Dr. John W. Krakauer for being a collaborator on a variety of frontier research projects.

ACKNOWLEDGMENTS

Thanks to Mary Joy Argo, Patty Hubbard, Marti Gilbert, Jiong Yang, Marvin Newhouse, Mark Chiveral and Debbie Cooper for their exceptional efforts to help me in every way.

I would like to thank the friends that been with me in the Biostatistics Department. Each of you helped me learn and grown in the past five years: Francis Abreu, Jiawei Bai, Shaojie Chen, Alyssa Frazee, Jonathan Gellar, Fang Han, Bing He, Lei Huang, Jeongyong Kim, Stephen Cristiano, Yi Lu, Parichoy Pal Choudhury, Tianchen Qian, Huitong Qiu, Yifei Sun, Elizabeth Sweeney, Yenny Webb Vargas, Yuting Xu, Chen Yue, Yuxin Zhu.

Finally, I want to express my deepest gratitude to my parents, Jianrong and Xiaoying, and my boyfriend, Wenying, who have always done their best to encourage and support me through the process of seeking my degree. Nothing I say will suffice to describe how much their efforts mean to me.

Contents

Acknowledgments **iv**

List of Tables **x**

List of Figures **xi**

1 Introduction **1**

2 Homotopic Group ICA for Multi-Subject Brain Imaging Data **5**

2.1 Introduction 8

2.2 Background 12

2.2.1 FastICA 13

2.2.2 Measures of Functional Connectivity 15

2.3 Methods 16

2.3.1 Homotopic Group ICA 17

2.3.2 Connection with gICA 18

2.3.3 Functional Homotopy 19

2.4 Simulation Results 20
2.4.1 A Simple Example 20
2.4.2 2D Simulation 21
Case I: Perfect Homotopy 22
Case II: Perfect Lateralization 23
Case III: Mixture of Lateralized and Homotopic Networks 24
2.4.3 Flag Example 24
2.5 Application to the ADHD-200 Dataset 25
2.5.1 Estimated Brain Networks 26
2.5.2 Measure of Network Symmetry and Lateralization 26
2.5.3 Functional Homotopy of the Networks 27
2.6 Discussion 28

3 On tests of activation map dimension for fMRI-based studies of learning 44

3.1 Introduction 47
3.2 Methods 48
3.3 Materials and Simulation 53
3.3.1 Motivating Data Set 53
3.3.2 Simulation Study 56
Simulation Under the Null Hypothesis 57
Simulation Under the Alternative Hypothesis 58
3.3.3 Simulation Results 60

Simulations Under the Null Hypothesis 61

Simulations Under the Alternative Hypothesis 61

3.4 Data Analysis of the Motivating Data Set 62

3.4.1 Motivating Data Results 62

3.5 Discussion 63

3.5.1 Simulation Results 63

3.5.2 General discussion 64

3.6 Appendix: Tables and Figures 66

4 Context Aware Group Nearest Shrunken Centroids in Large-Scale Genomic Studies 76

4.1 Introduction 78

4.2 Background 81

4.2.1 Nearest Shrunken Centroids (NSC) 81

4.2.2 Normal Score Transformation 83

4.3 Method 84

4.3.1 Model 85

4.3.2 Group Nearest Shrunken Centroids 86

4.3.3 Theoretical Properties of gNSC 87

Estimation Consistency. 88

Sparsity Recovery and Misclassification Consistency. 89

4.3.4 Nonparanormal and Normal Score Transformation 92

4.4 Application 93

4.4.1 gNSC for Classifcation 95

Procedures. 96

Results. 97

4.4.2 Context Analysis of Myc pathway 102

Context Analysis using NSC 103

Context Analysis using gNSC 105

4.4.3 Proof of Theorem 2 106

4.4.4 Proof of Theorem 3 107

4.4.5 Proof of Theorem 4 110

4.4.6 Proof of Theorem 5 111

5 Discussion **125**

6 Curriculum Vitae **137**

List of Tables

2.1 The two rows compare the mean of the voxel mean difference of H-gICA and ordinary gICA in the symmetric setting. 34
2.2 Comparison of the H-gICA results versus ordinary gICA results when the true sources are only in one hemisphere. The table provides the mean and standard error of the voxel mean difference with the true sources for the 300 iterations. 34

3.1 Results of the simulation studies. Shown are type I error rates and power across simulation settings. 67
3.2 P-values of the tests of dimensionality for the motivating data set. The first row considers the Session 1 versus Session 2 for the Horizontal task ($H_0 : \beta_{21}(v) = c\beta_{11}(v)$). The second row does the same for the vertical task ($H_0 : \beta_{22}(v) = c\beta_{12}(v)$). The third considers inter-session differences across tasks ($H_0 : \beta_{21}(v) - \beta_{11}(v) = c\{\beta_{22}(v) - \beta_{12}(v)\}$). P-values are given with and without having performed an angle correction. 68

4.1 Leave Fold Out v.s. Leave Experiment Out Cross Validation. 98
4.2 True relations learnt from the GPL96 data. 101
4.3 Leave experiment out cross validation method is used on the GPL96 data. Averaged misclassification errors and the corresponding averaged gene numbers across tissue types are provided with standard deviations included. We highlight the minimum values in bold. 114
4.4 10-fold cross validation method is used on the GPL96 data. Averaged misclassification errors and the corresponding averaged gene numbers across tissue types are provided with standard deviations included. We highlight the minimum values in bold. 115
4.5 The index of all gene pathways IDs. 118
4.6 Document Preparation 120
4.7 NSC(/gNSC) for Context Analysis 124

List of Figures

2.1 Result of the First Example. The top row shows the true sources in three different types:as perfectly symmetric; as only present in one hemisphere; as differing in the two hemispheres. The small blocks in these plots are activated voxels, where the values come from uniform distribution. The first two elements of the second row are the true sources in the left hemisphere and the last element is left minus right. The bottom row consists of the independent components generated by H-gICA. 33

2.2 True Sources in the Perfect Homotopy. The three sources are generated by Gamma distributions with different parameters. 33

2.3 The mean of the voxel mean difference increases with the standard deviation of the noise. The two methods are the same under noise-free settings and when the noise exists, H-gICA works better than ordinary gICA. 34

2.4 The ICs estimated by H-gICA and gICA. The first column shows the true sources. The other columns, from left to right, show the estimated ICs when the noise increases. The 1_{st}, 3_{rd}, and 5_{th} rows are for H-gICA and the 2_{nd}, 4_{th} and 6_{th} are for gICA. 35

2.5 True Sources in Non-homotopic Setting. All three true sources are only in one hemisphere. The values of the activated voxels are generated by Gamma distribution with different parameters. 35

2.6 True Sources in Mixed Setting. The first two sources are homotopic while the third one varies in the two hemispheres. 36

2.7 The homotopic ICs estimated by H-gICA and ordinary gICA. The first column shows the true sources. The other columns, from left to right, contain the estimated ICs with increasing noise. The 1_{st} and 3_{rd} rows are results of H-gICA and the 2_{nd} and 4_{th} are of ordinary gICA. 36

2.8 Actual sources of the flag example. The original flags images were taken from Wikipedia. 37

2.9 The estimated homotopic sources of H-gICA and ordinary gICA. The five columns, from left to right, contain the estimated ICs with increasing noise. The 1_{st}, 3_{rd} and 5_{th} rows shows the estimated sources extracted by H-gICA while the 2_{nd}, 4_{th} and 6_{th} rows shows the estimated sources extracted by ordinary gICA. 37
2.10 QQ plot of the 15 sources extracted by gICA. 38
2.11 The scatter plot of the 15 sources. 39
2.12 Comparison of ICs obtain from H-gICA (left column) and ordinary gICA (right column). 40
2.13 Network symmetry measured by the proportion of variance explained by the first eigenvalue of $\text{cov}(\boldsymbol{A}_{i,1}(\cdot,q)), \boldsymbol{A}_{i,2}(\cdot,q)))$. Network 1, 7, 8 and 11 are the visual, default mode, auditory and motor network respectively. . . . 41
2.14 Network lateralization measured by $\text{var}(\boldsymbol{A}_{i,1}(\cdot,q))/\text{var}(\boldsymbol{A}_{i,2}(\cdot,q))$. Network 1, 7, 8 and 11 are the visual, default mode, auditory and motor network respectively. 42
2.15 Comparison of functional homotopy of ADHD and normal developed children. Each column represents a network (visual, default mode, auditory and motor). Each pair (subjects and controls) are connected via a grey line. The left end points of the grey lines measure the functional homotopy of the control and the right end points are for the ADHD subjects. The black lines represent the group level functional homotopy for the four networks. . 43

3.1 Conceptual diagram for fMRI activation distributions based on the motivating study of motor learning. Shaded areas represent learning based (inter-session differences) between a trained (Y axis) and untrained (X axis) task. Across all panels, Area (A) represents voxels with change in activation across sessions only in the trained task, (B) represents voxels with change in activation across sessions in both the trained and untrained task, (C) voxels with change in activation across sessions only in the untrained task, (D) represents no change in activation for both tasks. The four panels (I-IV) represent different potential shapes of the activation distributions for (B) with Panels I and II showing a two dimensional shape and Panels III and IV showing an approximately one dimensional. In Panels I and III inter-sessions differences are symmetrically represented whereas in II and IV one task had a uniformly greater increase. 69
3.2 Example of the Arc Pointing Task (APT) executed within the fMRI session. Subjects were asked to navigate a cursor lying between the inner and outer concentric circles. Two tasks of similar difficulty were investigated. A horizontal task (Panel A) where subjects were trained in between two scanning sessions and a vertical task (Panel B) where subjects were not trained. . . . 70

3.3 Contrast maps from the horizontal arc pointing task. The X axis for each plot is the first session while the Y axis is the second. Red lines show the direction of the first principal component while a dotted identity line is shown for reference. 71
3.4 Contrast maps of the vertical arc pointing task.The X axis for each plot is the first session while the Y axis is the second. Red lines shows the direction of the first principal while a dotted identity line is shown for reference. 72
3.5 Example simulated data. Panel (a) shows the simulated data. Panel (b) shows the estimate of b using Equation (3.1). Panel (c) shows the estimate of Z, where $Z_k(v) = \text{Var}\{\hat{\delta}_k(v)\}^{-1/2}\hat{\delta}_k(v)$. 73
3.6 Example simulation from the alternative hypothesis. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. In Panel (a) the voxels have Gaussian variation added orthogonally to the major axis. In Panel (b) there is no relationship. 73
3.7 Example simulation for the setting when the principal axis differs across subjects. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. . . . 74
3.8 Example simulation from the alternative with changing activation sets. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. Shown are the true parameter values (leftmost panel), the simulated subject data (middle panel) and the Z values (rightmost panel). 74
3.9 A simulation example highlighting increased power for detecting learning based differences. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. The alternative of the dimensionality test is true and the P-value is 0.03, suggesting that activation extent is unrelated between tasks. However, only 11% of the voxels satisfy a voxel level test of significance (colored in red). 75

4.1 Significant Association Between Pathways and Tissue Types 99
4.2 The True Tissue Types v.s. the Predictive Tissue Types. 100
4.3 gNSC Results of Keywords v.s. Myc pathway. (a) The mean relevance levels of synonym word groups with the 37 genes in Myc pathway; (b) The figure illustrates $\widetilde{\mu}_{mk}$ calculated by Equation (4.3.5). 104
4.4 Non-Gaussian Data. The two rows are QQ-plot and empirical cdf plot of three different genes. 114
4.5 Heatplots for gNSC. The 100 pathways in this figure are randomly chosen. . 116
4.6 NSC Results of Keywords v.s. Genes. 117

Chapter 1

Introduction

CHAPTER 1. INTRODUCTION

Modern biological experiments routinely produce massive amounts of data. The aggregation of this data potentially provides better insight into diseases and their causal factors. Such aggregated data, so called "big data", are often of very large-scale, high-dimensional, processing with complex distributions. Modern statistical technologies are necessary for finding causing relations and meaningful associations in such data. Different methods have been developed to fit the requirements of the biological questions and the characteristics of the data. For example, to deal with data size, methods such as Principal Component Analysis (PCA) and Nearest Shrunken Centroids (NSC) [Tibshirani et al., 2002] are used for dimension reduction or feature selection. Other methods, such as Independent Components Analysis (ICA), help make sense of dimension reduced data.

In the first part of my thesis, we describe a novel modified group ICA method - homotopic group ICA (H-gICA) for brain imaging studies, where data are collected for many subjects and controls along with their demographic information. In functional magnetic resonance imaging (fMRI), such data are represented by a 4D matrix, where the three dimensions correspond to space and the forth dimension is time. The 4D arrays may be transformed into a 2D matrix by vectorizing the 3D image of the brain for each time point and concatenating the vectors over time [Eloyan et al., 2014]. ICA can then be applied on the matrix to uncover putatively underlying brain networks when the sources are believed to behave like independent, non-Gaussian random variables. The use of ICA in neuroimaging was introduced by Calhoun et al. [2001a]. Since then ICA has been used in experiments where subjects are performing tasks, as well as when subjects are at rest. The resting-state

brain functional patterns across individuals provide insights into the baseline activity of the human brain in the absence of deliberate and externally stimulated neuronal activity. The main idea of ICA is that the independent components are common across the group, whereas the mixing matrices are estimated for each subject. The group ICA [Calhoun et al., 2009, 2001b, Li et al., 2012] model can be written as

$$X_i(t,v) = \sum_{q=1}^{Q} A_i(t,q)S(q,v), \tag{1.1}$$

where i is the index for subjects, $X_i(t,v)$ is the 2D matrix transformed from the original fMRI signals, $A_i(t,q)$ is the mixing matrix and $S(q,v)$ is the underlying sources. Our proposed method, H-gICA identifies underlying patterns of brain activity for functionally homotopic networks, where functional homotopy - the high degree of synchrony in spontaneous activity between geometrically corresponding interhemispheric (i.e., homotopic) regions - is a fundamental characteristic of the brain's intrinsic architecture [Zuo et al., 2010]. In addition to improving network estimation, H-gICA allows for the investigation of functional homotopy via ICA based networks.

In the second part of my thesis, we consider task related activation before and after learning. We aim to test the interaction of training on task by investigating changes in activation distribution. The use of the analyses of lower dimensional subspaces of fMRI task-based activation maps [Zarahn, 2002, Worsley et al., 1997] provides our starting framework for such analysis. We investigate dimension tests for the study of learning, particularly mo-

tor learning. A test of one versus two dimensions on the set $\hat{D}$, i.e. rank$(\hat{D})$, investigates the hypothesis

$$H_0 : \beta_{21}(v) - \beta_{11}(v) = c\{\beta_{22}(v) - \beta_{12}(v)\}$$

for unspecified c, for all voxels v, where $\beta_{jk}(v)$ represents the population average of inter-subject voxel-level activation for session- ($j = 1, 2$) and task- ($k = 1, 2$). Comparing longitudinal learning effects with a reference (untrained) task attempts to address non-learning based biases when comparing across sessions. Principal components and root tests of the second eigenvalue [see Mardia et al., 1980] are then employed to investigate the hypothesis of one dimension versus two. The main goal is to investigate the properties of this procedure in the context of a motivating data set.

In the third part of my thesis, we propose a group nearest shrunken centroid method (gNSC) for analyzing large genomics datasets. Compared to the popular nearest shrunken centroid (NSC) method, gNSC improves the classification accuracy by utilizing the gene pathway information. To achieve more modeling flexibility, we exploit the semiparametric nonparanormal (or Gaussian copula) model. Empirically, we apply gNSC on analyzing large genomics datasets: (i) In terms of misclassification error, gNSC outperforms NSC; (ii) In terms of context analysis, gNSC leads to newest biological findings; (iii) In terms of computing, gNSC scales to very large datasets with dimension more than 10^{12} and is suitable for big data analysis. Some theoretical properties of gNSC are also provided.

Chapter 2

Homotopic Group ICA for Multi-Subject Brain Imaging Data

CHAPTER 2. HOMOTOPIC GROUP ICA FOR MULTI-SUBJECT BRAIN IMAGING DATA

Abstract

Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals. It is widely used in a variety of academic fields such as acoustics, electrophysiology and functional neuroimaging. In functional neuroimaging, so called 'group ICA' seeks to identify and quantify networks of correlated regions across subjects. This chapter reports on the development of a new group ICA approach, Homotopic Group ICA ("H-gICA"), for use on blind source separation of resting state functional magnetic resonance imaging (fMRI) data. Brain functional homotopy is the high degree of synchrony in spontaneous activity which exists between geometrically corresponding interhemispheric regions [Zuo et al., 2010]. The approach we proposed allows attainment of improved network estimates via brain functional homotopy. H-gICA increases the potential for network discovery and is theoretically proven to be identical to standard group ICA when the true sources are both perfectly homotopic and noise free. Moreover, compared to commonly applied group ICA algorithms, the structure of the H-gICA input data leads to significant improvement in computational efficiency. A simulation study confirms its effectiveness in homotopic, non-homotopic and mixed settings as well as on the ADHD-200 dataset. From the fifteen components postulated by

H-gICA, several brain networks were found including: visual networks, the default mode network, the auditory network, and others. In addition to improving network estimation, H-gICA facilitates the investigation of functional homotopy via ICA based networks.

2.1 Introduction

The function of the human brain during rest can be investigated using various functional imaging techniques [Biswal et al., 1995, Gusnard et al., 2001]. The resting-state brain functional patterns across individuals provide insights into baseline activity of the human brain in the absence of deliberate and/or externally stimulated neuronal activity.

Resting state functional magnetic resonance imaging (rs-fMRI), obtained using blood oxygen level dependent (BOLD) signals, is a key driving force within the field of brain mapping in neuroimaging studies. Recently, some attention has been focused on identifying the of possible origins of variation in the BOLD signals from fMRI data. Neuroimaging studies have identified associations between resting state brain networks estimated via fMRI data to aging, cognitive function and neurological and psychiatric disorders [Damoiseaux et al., 2008, Rombouts et al., 2005]. One popular approach for locating putative networks is blind source separation, which decomposes neuroimaging data into an outer product of spatial maps multiplied by their respective time courses. Compared with conventional analysis tools, such as the general linear model (GLM), blind source separation does not require a specific fMRI paradigm. The two most popular exploratory data analysis techniques for blind source separation are Principal Component Analysis (PCA) and Independent Component Analysis (ICA). ICA can be distinguished from PCA by its focus on model-level independence and non-Gaussianity. Moreover, ICA as a matrix decomposition does not yield left or right decomposition vectors that are orthonormal. Finally, ICA is usually applied after PCA-based dimensional reduction, and thus can be thought of as a

non-orthonormal reorganization of PCA.

Recently, ICA is popular for analyzing neuroimaging data, being successfully applied to single-subject analysis [Guo and Pagnoni, 2008, Beckmann and Smith, 2004, McKeown et al., 1997]. The extension of ICA to group inferences provides common independent components across subjects and enables identification of putative underlying brain networks for the group. Several multi-subject ICA approaches have been proposed: Calhoun et al. [2001b] presented so-called group ICA and created the Matlab toolbox (GIFT), which provides group estimation. GIFT consists of a first-dimension reduction using PCA for each subject, followed by a temporal concatenation of the reduced data, after which ICA is then applied to the aggregated data. More recently, Beckmann and Smith [2005] proposed a tensor ICA (TICA) by extending the single-session probabilistic ICA (PICA)[Beckmann and Smith, 2004]. TICA factors the multi-subject data as a tri-linear combination of three outer products, which represents the different signals and artefacts present in the data in terms of their temporal, spatial, and subject-dependent variations. Other group methods that have been proposed include the approach by Calhoun et al. [2001c] and by Esposito et al. [2005], both of which perform single-subject ICA and then attempt to combine the output into a group post-hoc by using approaches such as self-organized clustering or spatial correlation of the components.

Among the existing methods, the GIFT is predominantly used for performing group ICA analysis of multi-subject fMRI data. Spatial independence assumed by GIFT is well suited to the sparse distributed nature of the spatial pattern for most cognitive activation

paradigms [McKeown et al., 1997]. Moreover, its empirical performance has been consistently verified [Sorg et al., 2007, Garrity et al., 2007, Sambataro et al., 2010].

This chapter describes a novel modified group ICA method - homotopic group ICA (H-gICA) - to identify the underlying patterns of brain activity for functionally homotopic networks. Although functional and structural connectivity are not equivalent, left-right symmetric patterns of interhemispheric activation are among the most frequent findings in neuroimaging studies [Toro et al., 2008]. Functional homotopy - the high degree of synchrony in spontaneous activity between geometrically corresponding interhemispheric (i.e., homotopic) regions - is a fundamental characteristic of the brain's intrinsic functional architecture [Zuo et al., 2010]. Via H-gICA, the information of homotopic brain function is utilized to improve the identification of underlying brain networks. A spatial independence assumption is made relating to all voxels in each hemisphere. In a simulation study, H-gICA is shown to be preferable to (our implementation of) GIFT when the actual signals are homotopic and is competitive with GIFT under non-homotopic signal conditions. We emphasize that H-gICA does not require all networks to be homotopic. The efficacy of H-gICA methodology is demonstrated by an application to the ADHD-200 dataset [Milham, 2012, Eloyan et al., 2012]. From the fifteen components produced by H-gICA, several common brain networks were found, being clearly represented in smooth, contiguous volumes, despite no smoothing of the underlying data. The main networks found include the visual, the default mode and the auditory. In addition to improving network estimation, H-gICA allows for the investigation of functional homotopy via ICA based networks. Here the

quantification of functional homotopy of such networks are defined using a similar concept to that of Joel et al. [2011].

The remainder of the chapter is organized as follows: Section 4.2 reviews selected background methods for ICA model estimation and for calculating the connectivity of the networks. A fast, fixed-point algorithm known as FastICA proposed by [Hyvärinen, 1999] is reviewed in Section 2.2.1, and subsequently used in the estimation of H-gICA. Section 2.2.2 discusses the measures of functional connectivity proposed by [Joel et al., 2011], which is then the basis for defining the ICA based functional homotopy. Section 2.3 is the theoretical body of the chapter: Section 2.3.1 introduces the homotopic group ICA model; Section 2.3.2 illustrates that H-gICA and GIFT coincide under noise-free settings; and Section 2.3.3 provides a measure of the functional homotopy for ICA based networks. Section 2.4 provides a simulation study to demonstrate the effectiveness of H-gICA under homotopic, non-homotopic and mixed settings. This allows a comparison versus GIFT, demonstrating that by using the information of brain functional homotopy H-gICA improves the power of locating the underlying brain networks. Section 4.4 provides the application of the H-gICA on the ADHD-200 dataset, an open source dataset with 776 children and adolescents, while Section 5 contains a summary discussion.

2.2 Background

A core issue in multivariate biological signal processing is finding a transformation of the data both reduces the dimension as well as captures underlying true signals. The observed data are denoted by $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)^T$, an m-dimensional random vector, and the true underlying signals by $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)^T$, an n-dimensional transform of $\boldsymbol{x}$. The problem is to determine a constant matrix $\boldsymbol{W}$ so that

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{x}. \tag{2.1}$$

A distinguishing feature of ICA components compared with other methods is that the elements of $\boldsymbol{s}$ are assumed independent in an factor analytic model, instead of focusing on data-level uncorrelatedness.. The ICA problem can be formulated using the following generative model for the data:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{2.2}$$

where $\boldsymbol{x}$ is the observed m-dimensional vector, $\boldsymbol{s}$ is the n-dimensional (latent) random vector whose components are assumed mutually independent, and $\boldsymbol{A}$ is a constant $m \times n$ matrix to be estimated. (A noise vector may also be added.) If it is further assumed that the dimensions of $\boldsymbol{x}$ and $\boldsymbol{s}$ are equal, i.e., $m = n$, the estimate of $\boldsymbol{W}$ in (2.1) is then obtained as the inverse of the estimate for matrix $\boldsymbol{A}$.

Because of the requirement for $\boldsymbol{A}$ to be invertible, a first stage dimension reduction is

required. Thus the model can be more accurately stated as

$$\tilde{\boldsymbol{x}} = U^t(\boldsymbol{x} - \bar{\boldsymbol{x}}) = \boldsymbol{A}\boldsymbol{s},$$

where U^t is a dimension reduction matrix of size $m \times n$. Normally, this is reached via a singular value decomposition of the demeaned data.

2.2.1 FastICA

The previously mentioned whitened data input into an ICA algorithm are mean zero and uncorrelated, and thus Gaussian distributional assumptions provide little further insight to linear reorganizations. This leads to the search for an optimized $\boldsymbol{W}$ that maximizes non-Gaussianity. Such non-Gaussianity of the independent components is necessary for the identifiability of the model shown in Equation (2.2) [Comon, 1994]. A number of non-Gaussianity estimates are proposed such as kurtosis [Oja et al., 2001] and negentropy [Pillai et al., 2002].

Hyvärinen [1999] proposed a fast fixed-point algorithm known as FastICA for ICA, which generalizes the higher-order cumulant approximation of the two classic measures of non-Gaussianity such that it uses expectations of general non-quadratic functions. A fixed-point algorithm was developed and proposed that maximizes non-Gaussianity and estimates the independent components one at a time via a deflation scheme.

Following Hyvarinen's treatment of ICA, mutual information [Oja et al., 2001], a natu-

ral measure of the dependence between random variables, can be written as

$$I(y_1, y_2, \cdots, y_n) = J(\boldsymbol{y}) - \sum_i J(y_i),$$

when the variables y_i, $i = 1, \cdots, n$, are uncorrelated. Here $\boldsymbol{y} = (y_1, \cdots, y_i)^T$ and J is negentropy, a measure of non-Gaussianity [Comon, 1994]. Thus, in the fastICA approach to ICA, the independent components the invertible transformation, W, is determined so that the mutual information of the independent components, s_i is minimized. As mentioned by Hyvärinen [1999], this is roughly equivalent to finding directions in which the negentropy is maximized.

For any random variable y_i, with zero mean and unit variance, an approximation of negentropy $J(y_i)$ is of the form:

$$J(y_i) \approx c[E\{G(y_i)\} - E\{G(Z)\}]^2,$$

where G is the contrast function, c is an constant, and Z is a standard Gaussian random variable. If $G(y) = y^4$, J measures distance in the fourth moment (kurtosis), mirroring earlier efforts in ICA. Thus the problem is equivalent to the following optimization problem:

$$\text{argmax}_{w_i} \sum_{i=1}^{n} J_G(\boldsymbol{w}_i^T \boldsymbol{x}) \;\; wrt. \;\; \boldsymbol{w}_i, \; i = 1, \cdots, n,$$

under uncorrelatedness constraint $E\{(\boldsymbol{w}_k^T \boldsymbol{x})(\boldsymbol{w}_j^T \boldsymbol{x})\} = \delta_{jk}$. Note that the maxima of

$J(\boldsymbol{w}^T\boldsymbol{x})$ are obtained at certain constrained optima of the contrast function, $E\{G(\boldsymbol{w}^T\boldsymbol{x})\}$. In practice, adaptive algorithms with fixed-point iterations are used for the purpose of computational simplicity and ordering the estimated components.

The idea behind FastICA, in particular, is to estimates the independent components one-by-one. In the i_{th} iteration Newton's method is used to find $\boldsymbol{w}_i$, which achieves a maxima of $E\{G(\boldsymbol{w}^T\boldsymbol{x})\}$ under the constraints: 1) $E\{(\boldsymbol{w}^T\boldsymbol{x})^2\} = \|\boldsymbol{w}\|^2 = 1$ and 2) $E\{(\boldsymbol{w}^T\boldsymbol{x})(\boldsymbol{w}_j^T\boldsymbol{x})\} = 0$, for $\forall j < i$. The second constraint requires the components to be uncorrelated, which is equivalent to orthogonality in the whitened space.

2.2.2 Measures of Functional Connectivity

Functional magnetic resonance imaging (fMRI) measures proxies for brain activation via the principle of neuro-vascular coupling and the BOLD (blood oxygen level dependent) signal. Based on these measurements, interregional connectivity of brain activity can be estimated. Joel et al. [2011] point out that the functional connectivity between any two brain regions may be due to within network connectivity (WNC) and between network connectivity (BNC). They emphasize the importance of interpreting such connectivity, ostensibly measured by the correlation and variance of the temporal mixing matrices, respectively. Mathematically, Joel et al. [2011] define ICA-based BNC for voxel v_1 and v_2 as:

$$BNC'_{k,l}(v_1, v_2) = \frac{\sum_{t=1}^{m} \boldsymbol{A}^{(k)}(t)\boldsymbol{A}^{(l)}(t)}{\sqrt{\sum_{t=1}^{m} \boldsymbol{X}^2(v_1,t)}\sqrt{\sum_{t=1}^{m} \boldsymbol{X}^2(v_2,t)}},$$

where $\boldsymbol{A}^{(k)}$ is the k_{th} column of $\boldsymbol{A}$ in (2.2), which represents the time course modulating spatial map k. $\boldsymbol{A}^{(k)}(t)$ is its t_{th} element. $\boldsymbol{X}(v,t)$ is the demeaned BOLD fMRI signal from voxel v at time t. Joel et al. [2011] generalize the whole brain BNC as the correlation of the network time courses. A similar approach can be used to measure brain functional homotopy, which is detailed in Section 2.3.3.

2.3 Methods

The extension of ICA to group inferences provides common independent components across subjects, which allows identification of putative common brain networks for the whole group. As in most fMRI studies, spatial independence is assumed in our group ICA model, since it is well-suited to the sparse distributed nature of the spatial pattern for most cognitive activation paradigms [McKeown et al., 1997]. However, instead of the entire brain, the spatial independence assumption in H-gICA is made for voxels within a single hemisphere. To use the information of brain functional homotopy, the fMRI data are registered to symmetric templates and thus the number of voxels are equal in the left and right hemispheres. Similarly to in Calhoun et al. [2001b], a dimension reduction using PCA is applied to each hemisphere of each subject. Another PCA step is then performed on the concatenated data matrix, which is not technically necessary [Eloyan et al., 2013] though is the current standard practice for group ICA.

2.3.1 Homotopic Group ICA

The following presents the Homotopic group ICA (H-gICA) model. Suppose there are N subjects indexed by $i = 1, 2, \cdots, N$ and T PCs obtained from each hemisphere of each subject. Since the brain images are registered to a symmetric template, it can be assumed there are V voxels in each hemisphere. The voxels in the left hemisphere are indexed by $v = 1, 2, \cdots, V$. For each voxel in the right hemisphere the same index as its mirrored voxel in the left hemisphere is used. For each subject, two matrices of dimension $T \times V$ are created to represent the observed data, $\boldsymbol{X}_{i,1}$ and $\boldsymbol{X}_{i,2}$, where the $\boldsymbol{X}_{i,j}$ is data matrix of the left ($j = 1$) or the right ($j = 2$) hemispheres of subject i. The column v of $\boldsymbol{X}_{i,j}$ represents the fMRI series of voxel v in the hemisphere j. The rows of $\boldsymbol{X}_{i,j}$ are the PCs, which are indexed by $t = 1, 2, \cdots, T$.

$\boldsymbol{X}_{i,j}(t, v)$ represents row t, column v of $\boldsymbol{X}_{i,j}$ with the same convention applied to other vectors and matrices. Assuming that a group ICA decomposition implies: $\boldsymbol{X}_{i,j}(t, v) = \sum_{q=1}^{Q} \boldsymbol{A}_{i,j}(t, q)\boldsymbol{S}(q, v)$, for all $i = 1, 2, \cdots, N$ and $j = 1, 2$. This in turn implies that the spatio-temporal process $\boldsymbol{X}_{i,j}(t, v)$ can be decomposed to a hemisphere specific time series $\boldsymbol{A}_{i,j}(t, q)$ and a subject-independent spatial maps, $\boldsymbol{S}(q, v)$. This is equivalent to the following group ICA model:

$$\boldsymbol{X} = \boldsymbol{M}\boldsymbol{S}. \tag{2.3}$$

Here $\boldsymbol{X} = [\boldsymbol{X}_1^T, \boldsymbol{X}_2^T]^T$ is the $2NT \times V$ group data matrix from left and right hemispheres, where $\boldsymbol{X}_j := [\boldsymbol{X}_{1,j}^T, \boldsymbol{X}_{2,j}^T, \cdots, \boldsymbol{X}_{N,j}^T]^T$, $j = 1, 2$, which formed by concatenating N sub-

jects' data in the temporal domain. $\boldsymbol{S}$ is $Q \times V$ matrix containing Q statistically independent spatial maps in its rows. $\boldsymbol{M} = [\boldsymbol{M}_1^T, \boldsymbol{M}_2^T]^T$ is the $2NT \times Q$ group mixing matrix, where $\boldsymbol{M}_j = [\boldsymbol{A}_{1,j}^T, \boldsymbol{A}_{2,j}^T, \cdots, \boldsymbol{A}_{N,j}^T]^T$ is the $NT \times Q$ submatrix corresponding to hemisphere j concatenating the mixing matrix of the N subjects. In the context of fMRI, the $\boldsymbol{S}(q, \cdot)$, $q = 1, 2, \cdots Q$ are spatial maps that are often interpreted as brain networks. It is further assumed that the dimensions of $\boldsymbol{X}_{i,j}$ and $\boldsymbol{S}$ are equal, i.e. $Q = T(\ll V)$. In order for the ICA model to be fully identifiable a further assumption is that the square mixing matrices, $\boldsymbol{A}_{i,j}$, are of full rank. Thus we can define $\boldsymbol{W}_{i,j} = \boldsymbol{A}_{i,j}^{-1}$.

In the H-gICA model, the independent components are assumed to be common across subjects and hemispheres, while how they mix to produce the signal can differ among both subjects and hemispheres. The true mixing matrices are not observed. The FastICA algorithm is used to obtain the model-based estimate. Comparing with group ICA, H-gICA actually double the number of parameters while reducing the number of voxels in the estimated sources by half.

2.3.2 Connection with gICA

In this section H-gICA is linked to the most commonly used group ICA approach [Calhoun et al., 2001b]. By appropriately setting the data structure, the group structure assumed in their approach is equivalent to a setting of $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{M}}\tilde{\boldsymbol{S}}$. Here $\tilde{\boldsymbol{X}} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ is the $NT \times 2V$ group data matrix, where $\boldsymbol{X}_j = [\boldsymbol{X}_{1,j}^T, \boldsymbol{X}_{2,j}^T, \cdots, \boldsymbol{X}_{N,j}^T]^T$ (same as that in the H-gICA model in Section 2.3.1) is the data corresponding to the hemisphere j ($j = 1, 2$)

after dimension reduction. $\tilde{\boldsymbol{S}}$ is a $Q \times 2V$ matrix containing Q statistically independent spatial maps in its rows. And $\tilde{\boldsymbol{M}} = [\tilde{\boldsymbol{A}}_1{}^T, \tilde{\boldsymbol{A}}_2{}^T, \cdots, \tilde{\boldsymbol{A}}_N{}^T]^T$ is the $NT \times Q$ group mixing matrix, where $\tilde{\boldsymbol{A}}_i$ is the $T \times Q$ submatrix corresponding to subject i. Under the assumption that $Q = T$ and that $\tilde{\boldsymbol{A}}_i$, are of full rank, defines $\tilde{\boldsymbol{W}}_i = \tilde{\boldsymbol{A}}_i{}^{-1}$, which can also be estimated by the FastICA algorithm. The following theorem, shows that if the actual sources are truly homotopic and noise free, H-gICA and gICA will have exactly same result when using the estimation method of FastICA. The proof of the theorem is given to the Appendix.

THEOREM 1. *Suppose the actual sources are truly homotopic and noise free. The number of estimated ICs is $Q = T \ll V$. Denote the FastICA estimate of $\boldsymbol{S}$ to be $\hat{\boldsymbol{S}}$ and the estimate of $\tilde{\boldsymbol{S}}$ to be $\hat{\tilde{\boldsymbol{S}}}$. Then for $j \in \{1, 2\}$, we have $\hat{\tilde{\boldsymbol{S}}} = [\hat{\boldsymbol{S}}, \hat{\boldsymbol{S}}]$.*

Theorem 1 shows that when there is no noise, H-gICA and GIFT are the same for the homotopic signals. However, of course, noise exists in most real data sets. Section 2.4 deals with H-gICA's ability to improve locating underlying sources when noise is added to the data.

2.3.3 Functional Homotopy

Similarly to Joel et al. [2011] the following defines a measure of subject- and network-specific functional homotopy for the k_{th} ICA based network of subject i:

$$H_i(k) = \text{Cor}(\boldsymbol{A}_{i,1}^{(k)}, \boldsymbol{A}_{i,2}^{(k)}), \tag{2.4}$$

where $\boldsymbol{A}_{i,j}$ $(i = 1, 2, \cdots, N, j = 1, 2)$ is the mixing matrix of the i_{th} subject corresponding to hemisphere j and $\boldsymbol{A}_{i,j}^{(k)}$ is a vector of the time course modulating spatial map k. Note, typically this requires a back transformation from PCA space. Here $H_i(k)$ measures the spontaneous activity of the k_{th} network between left and right hemispheres. The estimation of $H_i(k)$ in H-gICA is given by replacing $\boldsymbol{A}_{i,1}$ and $\boldsymbol{A}_{i,2}$ with their estimated values in (2.4). Similarly the group level functional homotopy for the k_{th} ICA based network can be define as:

$$H(k) = \text{Cor}(\boldsymbol{M}_1^{(k)}, \boldsymbol{M}_2^{(k)}), \tag{2.5}$$

where $\boldsymbol{M}_j$ is defined in (2.3) as the submatrix corresponding to hemisphere j concatenating the mixing matrix of the N subjects and $\boldsymbol{M}_j^{(k)}(t)$ is the t_{th} element of the time course modulating spatial map k.

2.4 Simulation Results

2.4.1 A Simple Example

The following simple example is given to initially demonstrate the effectiveness of H-gICA. Suppose the number of subject is $N_s = 3$ and the number of underlying sources is $Q = 3$. The data are from the model $\boldsymbol{X}_i = \boldsymbol{A}_i\boldsymbol{S}$ with $T = 3$ and $V = 10^2$. We further

assume that

$$\boldsymbol{A}_1 = \begin{pmatrix} -1 & -5 & 2 \\ 5 & -3 & 2 \\ 5 & 3 & -5 \end{pmatrix}, \boldsymbol{A}_2 = \begin{pmatrix} -1 & -1 & 2 \\ -1 & -2 & -3 \\ -4 & 0 & 5 \end{pmatrix}, \boldsymbol{A}_3 = \begin{pmatrix} -3 & 3 & -5 \\ 5 & -1 & -1 \\ 3 & -2 & -1 \end{pmatrix}$$

(The elements in $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{A}_3$ are randomly picked from $Unif(-10, 10)$.)

As shown in Figure 2.1, three different sources are generated: one is symmetric; another is only in one hemisphere; and the third has two asymmetric blocks of activated voxels. These results are shown in the bottom row of Figure 2.1, highlighted by H-gICA's separation of the three sources and its meaningful prediction for all of them.

2.4.2 2D Simulation

In order to illustrate the performance of the proposed method, simulation studies with $10,000$ voxels in 2D spaces were conducted. Since voxels in each hemisphere are stretched into a vector before ICA analysis. The shape of the sources would not matter for the simulation analysis. We simply use block sources here. Both the homotopic and non-homotopic settings are used in the study. The results are compared with the commonly used group ICA algorithm without consideration of the left and right hemispheres. The simulation shows that the estimation of the networks is improved for homotopic networks regardless of the lateralization. This improvement still exists even when non-homotopic networks are added.

Similar to above, the number of subjects is $N_s = 3$ and the number of underlying sources is $Q = 3$. The data are generated by the ICA model $\boldsymbol{X}_i = \boldsymbol{A}_i\boldsymbol{S}$ with $T = 3$ and $V = 100^2$. $\boldsymbol{A}_i$ has the same value as in the toy example for $i = 1, 2, 3$ is also assumed.

Case I: Perfect Homotopy

Assume all true sources are perfectly homotopic. For each source, two blocks of voxels are symmetrically activated. In each loop of the simulation, values of these activated voxels are assigned by Gamma distributions. The heatplots of the sources are shown in Figure 2.2. The data are generated by the three sources in Figure 2.2 and the mixing matrices $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, and $\boldsymbol{A}_3$. Gaussian noise is then added to the data. We also tried non-Gaussian noise such as Gamma, which gives similar results. The estimated components are standardized and subsequently compared with the corresponding standardized true sources. Next, the mean difference at each voxel is calculated. The simulation contained 300 iterations in each run. Homotopic gICA results are compared with the commonly used group ICA algorithm in Table 2.1, where the noise is set to be mean zero and with a standard deviation equal to 5.

As seen in Table 2.1, in the setting of symmetric sources, the mean errors of H-gICA are smaller than that of ordinary gICA for all the three sources. Figure 2.3 shows the mean of the voxel mean difference with different settings for the noise. Again, H-gICA works consistently better when noise exists and is the same as ordinary gICA in noise-free settings, which was been proven in Theorem 1.

Figure 2.4 compares the ICs estimated by H-gICA and the ordinary gICA. As we can

see, comparing with the ordinary gICA, H-gICA has a consistently better estimation of the ICs when noise exists.

Case II: Perfect Lateralization

True sources are now presumed to be only present in one hemisphere (perfect lateralization of the brain network), and without loss of generality, the left hemisphere is selected. Similar to Case I, a Gamma distribution was used to generate the value of the activated voxels and, again, 300 iterations were run. Figure 2.5 shows the heatplots of the three true sources in the first iteration. Gassian noise was added to the data generated by the three sources and the results are compared with ordinary gICA. Note that, since the independent components provided by H-gICA contain only one hemisphere by design. Thus, the sole applicable comparison is the true sources in the left hemisphere. Table 2.2 gives these results, and show that the mean error of H-gICA is less than that of gICA for Source 2 while larger than the mean error of gICA for Source 1 and Source 3. In summary, H-gICA is competitive with normal g-ICA when the sources are lateralized with the caveat that H-gICA does not provide lateralization information on which hemisphere the network resides in. We do note, however, that this information is contained in the temporal mixing matrix, just not in a form easily displayed as an image.

Case III: Mixture of Lateralized and Homotopic Networks

For this setting, a mixture of two different types of sources is introduced. As shown in Figure 2.6, the first two sources are homotopic while the third source covers different regions in the two hemispheres. The estimated ICs for the two homotopic sources are show in Figure 2.7, which illustrates that the effectiveness of estimating the homotopic sources is not impacted by adding non-homotopic sources. Since the sources estimated by H-gICA are only in one hemisphere, it is not possible to compare results of the non-homotopic sources.

2.4.3 Flag Example

In this example, the actual sources will be the gray-scaled flags of the USA, Canada, the European Union, China and Russia. As shown in Figure 2.8, three of them are symmetric (Canada, the European Union and Russia) and two are not (USA and China). Similar as in Section 2.4.2, the data are generated by the ICA model with a fixed mixing matrix, Gaussian noise was then added. The results are shown in Figure 2.9, where the 1_{st}, 3_{rd} and 5_{th} rows are the estimated symmetric sources extracted by H-gICA and the 2_{nd}, 4_{th} and 6_{th} rows are the estimated symmetric sources extracted by ordinary gICA. As we can see, for all of the three symmetric sources, H-gICA provides clearer estimation as the noises increased. Moreover, leakage that is apparent in the g-ICA is not apparent in H-gICA.

2.5 Application to the ADHD-200 Dataset

Application of the H-gICA method is illustrated using the ADHD-200 dataset, one of the freely available largest fMRI datasets. The data derive from 776 resting-state fMRI and anatomical datasets aggregated across eight independent imaging sites, 491 of which were obtained from normally developing individuals and 285 in children and adolescents with ADHD (ages: 7-21 years old). We view this analysis as largely a proof of principal in applying the method and defer thorough investigations of ADHD for later work.

This particular analysis focused on 20 subjects randomly picked from the ADHD-200. Data were processed via the NITRC 1,000 Functional Connectome processing scripts [Mennes et al., 2012]. In summary, images were slice-time corrected, deobliqued, skull stripped, smoothed and registered to a 3 mm3 MNI template. The data are then registered to ICBM 2009a nonlinear symmetric templates generated by the McConnell Brain Imaging Centre [Fonov et al., 2009, 2011]. Each fMRI data contain $99 \times 117 \times 95 = 1,100,385$ voxels measured at 176 time points. Figure 2.10 and Figure 2.11 are the QQ plot and scatter plots of the estimated sources extracted by ordinary gICA in left and right hemisphere. As we can see, most of the estimated sources are close to the 45^o line, which suggest that the marginal distribution of left and right hemisphere are quite similar. H-gICA can be benefit from utilizing this information.

2.5.1 Estimated Brain Networks

Our procedure is as follows: Each fMRI image is separated into left and right hemispheres. Thus, each hemisphere contains $49 \times 117 \times 95 = 544,635$ voxels. Similar to standard group ICA [Calhoun et al., 2001b], a dimension reduction using PCA is applied to each hemisphere of each subject. 15 PCs are obtained for each hemisphere. A group data matrix is generated by concatenating the reduced data of both hemispheres of the 20 fMRI imag in the temporal domain. Thus, the aggregated matrix has dimension $2NT \times V$, where $N = 20$, $T = 15$, and $V = 544,635$. Our algorithm of homotopic group ICA is then applied on this matrix. 15 estimated independent components are postulated by H-gICA. As shown in Figure 2.12, out of the 15 components, several brain networks were found including: the visual network 2.12(a), the default mode network 2.12(c), the auditory network 2.12(e), and the motor network 2.12(g). Compared with the ICs obtained ordinary gICA, shown in 2.12(b), 2.12(d), 2.12(f) and 2.12(h), H-gICA improves the estimation of all of these sources by yielding substantially more.

2.5.2 Measure of Network Symmetry and Lateralization

H-gICA allows one to quantify the lateralization of the learnt networks. For each subject, one can measure the symmetry of network q via the proportion of variance explained by the first eigenvector of the $\text{cov}(\boldsymbol{A}_{i,1}(\cdot, q)), \boldsymbol{A}_{i,2}(\cdot, q)))$. A high proportion of variance relates to a homotopic network. A low proportion of variance, on the other hand, relates to a non-homotopic network. Boxplots of the symmetry measure of the 15 networks learnt by

H-gICA are shown in Figure 2.13.

In the Section 2.4.2, we shows that H-gICA improves the estimation of the networks regardless of the lateralization. However one can still measure lateralization of the networks using the estimated mixing matrix in H-gICA. Here we use the variance ratio of the corresponding column for the networks in the mixing matrix for left and right hemisphere, i.e. $\text{var}(\boldsymbol{A}_{i,1}(\cdot, q))/\text{var}(\boldsymbol{A}_{i,2}(\cdot, q))$, as a measure of lateralization. In Figure 2.14, we shows the measure of lateralization of the 15 networks learnt by H-gICA.

2.5.3 Functional Homotopy of the Networks

To compare the functional homotopy of ADHD and typical developed children, choose 20 ADHD subjects and 20 controls. The subjects and controls were matched in gender and age. Via Equations (2.4) and (2.5), the estimated functional homotopy of four networks (visual, default mode, auditory and motor) are shown in Figure 2.15. As one can see, the functional homotopy of ADHD children tends to be lower in both visual networks and the auditory network. These findings represent meaningful leads on the exploration of homotopic network relationships and disease, though a full exploration is reserved for later work.

2.6 Discussion

In this chapter we presented a new group ICA method called homotopic group ICA (H-gICA). Similar with ordinary group ICA methods, H-gICA can analyze data of multi-subjects and estimate common underlying IC's across individuals. By concatenating the fMRI data of the two hemispheres, H-gICA effectively doubles the sample size. It improves the power of finding the underlying brain networks by rearranging the data structure and utilizing the information of spontaneous brain activity between geometrically corresponding inter-hemispheric regions. Both the simulation study and the application on ADHD 200 data show that H-gICA works better than the ordinary gICA when the data are homotopic and it is competitive with ordinary gICA at estimating homotopic sources even in the presence of non-homotopic sources. Effectiveness was demonstrated by application on the ADHD-200 dataset: several brain networks were found and clearly represented in smoother (compared with ordinary gICA), contiguous volumes despite no smoothing of the underlying data. The main networks found included visual networks, the default mode network, the auditory network, as well as others. Moreover H-gICA enables certain measurement of the functional homotopy of the underlying functional networks. This potentially offers the opportunity to analyze the relation of the brain functional homotopy between the left and right hemisphere with diseases.

Appendix

Proof of Theorem 1

Proof. Since the true sources are truly homotopic and noise free, we have:

$$\boldsymbol{X}_{i,1} = \boldsymbol{X}_{i,2}$$

, where $\boldsymbol{X}_{i,j}$, $j = 1, 2$, are the data of left and right hemispheres after dimension reduction by PCA. This is equivalent to

$$\boldsymbol{X}_1 = \boldsymbol{X}_2.$$

Without losing of generality, we assume $\boldsymbol{X}_j$ ($j = 1, 2$) are demeaned i.e. the row means of $\boldsymbol{X}_j$ are all 0. Assume the singular value decomposition of the matrix $\boldsymbol{X}_1/\sqrt{V}$ is

$$\boldsymbol{X}_1/\sqrt{V} = \boldsymbol{U}\Sigma\boldsymbol{V}^T,$$

where $\boldsymbol{U}$ is of dimension $NT \times NT$ which consists of the left singular vectors of $\boldsymbol{X}_1$, Σ is a diagonal matrix of dimension $NT \times NT$ with the singular values, and $\boldsymbol{V}$ is of dimension $V \times NT$ which consists of the right singular vectors. Both $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal. Then

we have

$$\begin{aligned} \boldsymbol{X}/\sqrt{V} &= [(\boldsymbol{X}_1/\sqrt{V})^T, (\boldsymbol{X}_2/\sqrt{V})^T]^T \\ &= [(\boldsymbol{U}\Sigma\boldsymbol{V}^T)^T, (\boldsymbol{U}\Sigma\boldsymbol{V}^T)^T]^T \\ &= [\boldsymbol{U}^T, \boldsymbol{U}^T]^T\Sigma\boldsymbol{V}^T \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} \tilde{\boldsymbol{X}}/\sqrt{V} &= [(\boldsymbol{X}_1/\sqrt{V}), (\boldsymbol{X}_2/\sqrt{V})] \\ &= [(\boldsymbol{U}\Sigma\boldsymbol{V}^T), (\boldsymbol{U}\Sigma\boldsymbol{V}^T)] \\ &= \boldsymbol{U}\Sigma[\boldsymbol{V}^T, \boldsymbol{V}^T] \end{aligned}$$

Note that in (2.6), since $\boldsymbol{X}$ is not full ranked, only NT singular value are non-zero and only NT singular vectors are estimated. Thus we can define

$$\boldsymbol{K} := \frac{1}{2}\Sigma^{-1}[\boldsymbol{U}, \boldsymbol{U}] \;\; and \;\; \tilde{\boldsymbol{K}} := \Sigma^{-1}\boldsymbol{U} \tag{2.7}$$

to be the pre-whitening matrix that projects data onto the principal components:

$$\boldsymbol{Z} = \boldsymbol{K}\boldsymbol{X} \;\; and \;\; \tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{K}}\tilde{\boldsymbol{X}}$$

, where $\boldsymbol{Z} = \sqrt{V}\boldsymbol{V}^T$ and $\tilde{\boldsymbol{Z}} = [\sqrt{V}\boldsymbol{V}^T, \sqrt{V}\boldsymbol{V}^T]$ are the whiten data for H-gICA and GIFT respectively. Clearly, if we assume the random variables $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$ are taking values from the columns of $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Z}}$, then we will have:

$$E[\boldsymbol{z}\boldsymbol{z}^T] = \boldsymbol{I} \;\; and \;\; E[\tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}^T] = \boldsymbol{I}$$

Suppose we have the same initial value for the vectors $\boldsymbol{w}_p$ for all $p \in \{1, 2, \cdots, Q\}$. No matter how contrast function G is defined for each fixed $\boldsymbol{w}$ we have:

$$\begin{aligned} E\{\boldsymbol{z}g(\boldsymbol{w}^T\boldsymbol{z})\} &= \frac{1}{n}\sum_{j=1}^{n} g(\boldsymbol{w}^T\boldsymbol{Z}(\cdot, j)) \\ &= \frac{\sqrt{V}}{n}\sum_{j=1}^{n} g(\boldsymbol{w}^T\boldsymbol{V}^T(\cdot, j)) \\ &= \frac{\sqrt{V}}{2n}\left(\sum_{j=1}^{n} g(\boldsymbol{w}^T\boldsymbol{V}^T(\cdot, j)) + \sum_{j=1}^{n} g(\boldsymbol{w}^T\boldsymbol{V}^T(\cdot, j))\right) \\ &= \frac{1}{2n}\sum_{j=1}^{2n} g(\boldsymbol{w}^T\tilde{\boldsymbol{Z}}(\cdot, j)) \\ &= E\{\tilde{\boldsymbol{z}}g(\boldsymbol{w}^T\tilde{\boldsymbol{z}})\} \end{aligned}$$

and similarly,

$$E\{g'(\boldsymbol{w}^T\boldsymbol{z})\}\boldsymbol{w} = E\{g'(\boldsymbol{w}^T\tilde{\boldsymbol{z}})\}\boldsymbol{w}$$

, where g is the derivative of G. So we have:

$$E\{\boldsymbol{z}g(\boldsymbol{w}^T\boldsymbol{z})\} - E\{g'(\boldsymbol{w}^T\boldsymbol{z})\}\boldsymbol{w} = E\{\tilde{\boldsymbol{z}}g(\boldsymbol{w}^T\tilde{\boldsymbol{z}})\} - E\{g'(\boldsymbol{w}^T\tilde{\boldsymbol{z}})\}\boldsymbol{w}$$

Thus, following the FastICA algorithm, it is easy to see that the estimates of $\boldsymbol{w}_p$, $p = 1, 2, \cdots, m$, will be same for these two approaches. By the definition of $\boldsymbol{K}$ and $\tilde{\boldsymbol{K}}$ in (2.7),

we will have:

$$\begin{aligned}\hat{\boldsymbol{W}} &= [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_m]^T \boldsymbol{K} \\ &= \frac{1}{2}[\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_m]^T [\tilde{\boldsymbol{K}}, \tilde{\boldsymbol{K}}] \\ &= \frac{1}{2}[\hat{\tilde{\boldsymbol{W}}}, \hat{\tilde{\boldsymbol{W}}}]\end{aligned}$$

And thus,

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{W}}\boldsymbol{X} = \frac{1}{2}[\hat{\tilde{\boldsymbol{W}}}, \hat{\tilde{\boldsymbol{W}}}][\boldsymbol{X}_1^T, \boldsymbol{X}_2^T]^T = \hat{\tilde{\boldsymbol{W}}}\boldsymbol{X}_1$$

$$\hat{\tilde{\boldsymbol{S}}} = \hat{\tilde{\boldsymbol{W}}}\tilde{\boldsymbol{X}} = \hat{\tilde{\boldsymbol{W}}}[\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_2] = \hat{\tilde{\boldsymbol{W}}}[\tilde{\boldsymbol{X}}_1, \tilde{\boldsymbol{X}}_1]$$

This lead to the result of Theorem 1. □

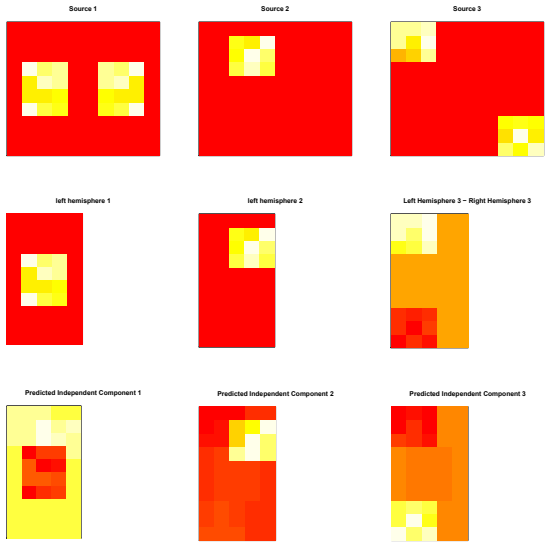


Figure 2.1: Result of the First Example. The top row shows the true sources in three different types:as perfectly symmetric; as only present in one hemisphere; as differing in the two hemispheres. The small blocks in these plots are activated voxels, where the values come from uniform distribution. The first two elements of the second row are the true sources in the left hemisphere and the last element is left minus right. The bottom row consists of the independent components generated by H-gICA.

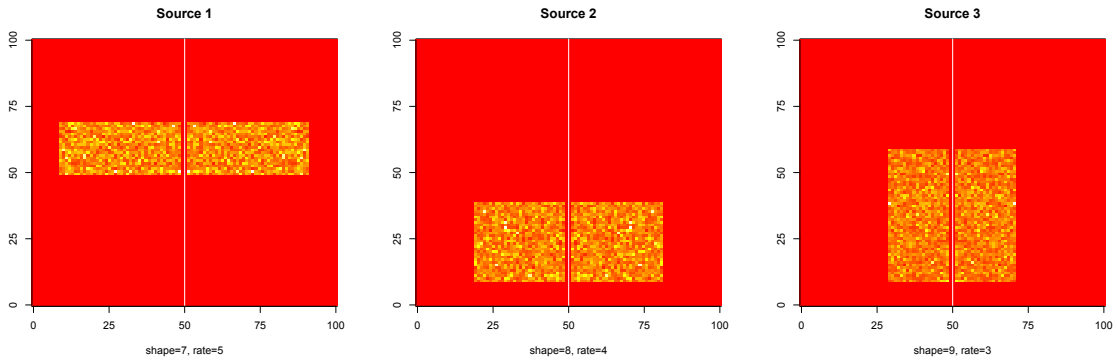


Figure 2.2: True Sources in the Perfect Homotopy. The three sources are generated by Gamma distributions with different parameters.

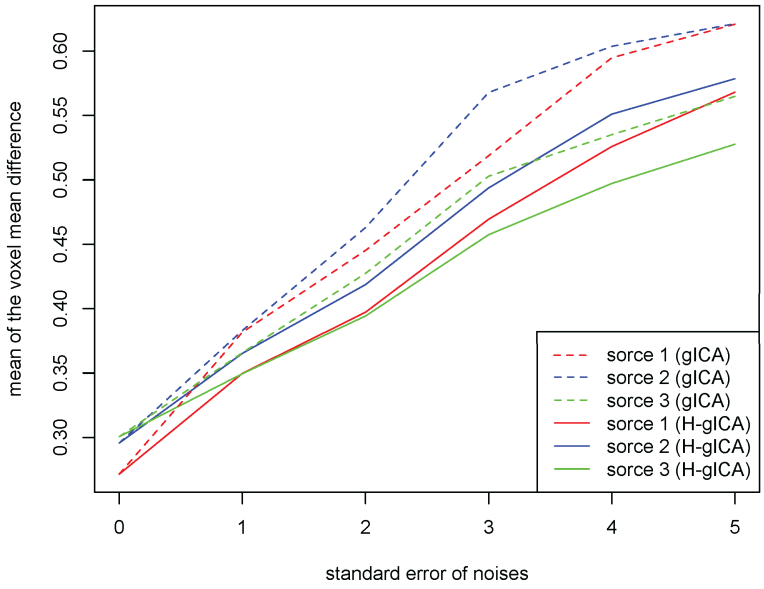


Figure 2.3: The mean of the voxel mean difference increases with the standard deviation of the noise. The two methods are the same under noise-free settings and when the noise exists, H-gICA works better than ordinary gICA.

Table 2.1: The two rows compare the mean of the voxel mean difference of H-gICA and ordinary gICA in the symmetric setting.

	Source 1	Source 2	Source 3
H-gICA	0.414 (s.e. 0.001)	0.412 (s.e. 0.001)	0.423 (s.e. 0.002)
gICA	0.507 (s.e. 0.001)	0.496 (s.e. 0.001)	0.433 (s.e. 0.001)

Table 2.2: Comparison of the H-gICA results versus ordinary gICA results when the true sources are only in one hemisphere. The table provides the mean and standard error of the voxel mean difference with the true sources for the 300 iterations.

	Source 1	Source 2	Source 3
H-gICA	0.508 (s.e. 0.002)	0.494 (s.e. 0.001)	0.431 (s.e. 0.002)
gICA	0.502 (s.e. 0.003)	0.533 (s.e. 0.002)	0.312 (s.e. 0.003)

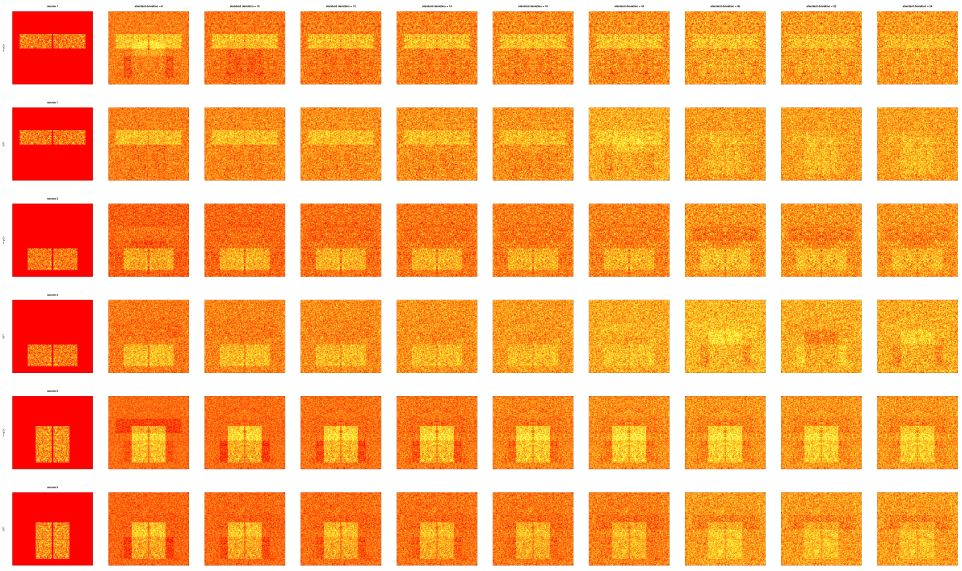


Figure 2.4: The ICs estimated by H-gICA and gICA. The first column shows the true sources. The other columns, from left to right, show the estimated ICs when the noise increases. The 1_{st}, 3_{rd}, and 5_{th} rows are for H-gICA and the 2_{nd}, 4_{th} and 6_{th} are for gICA.

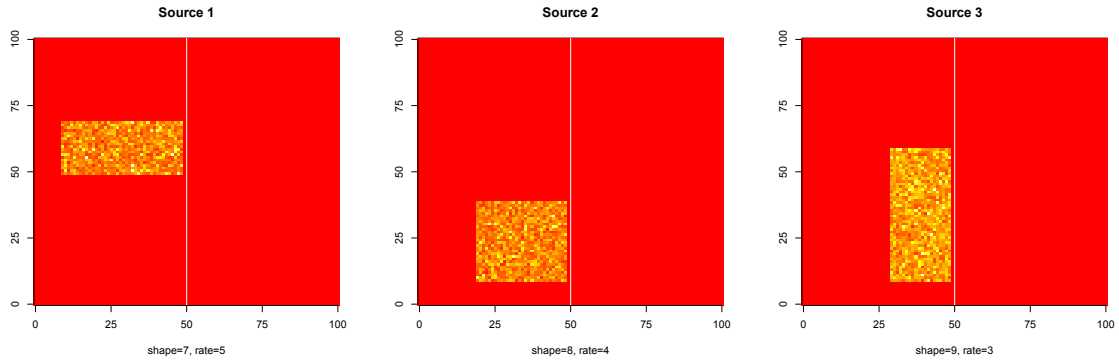


Figure 2.5: True Sources in Non-homotopic Setting. All three true sources are only in one hemisphere. The values of the activated voxels are generated by Gamma distribution with different parameters.

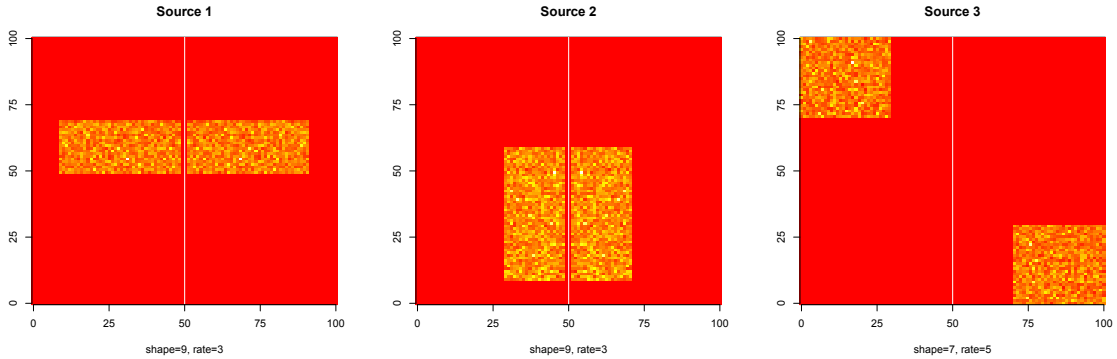


Figure 2.6: True Sources in Mixed Setting. The first two sources are homotopic while the third one varies in the two hemispheres.

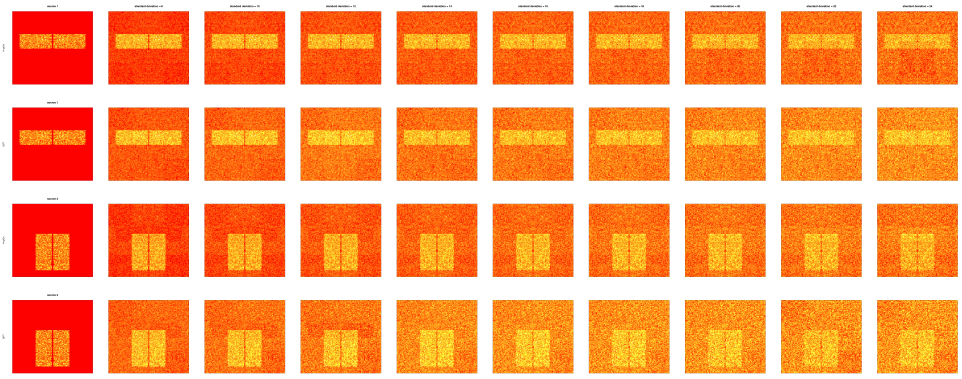


Figure 2.7: The homotopic ICs estimated by H-gICA and ordinary gICA. The first column shows the true sources. The other columns, from left to right, contain the estimated ICs with increasing noise. The 1_{st} and 3_{rd} rows are results of H-gICA and the 2_{nd} and 4_{th} are of ordinary gICA.

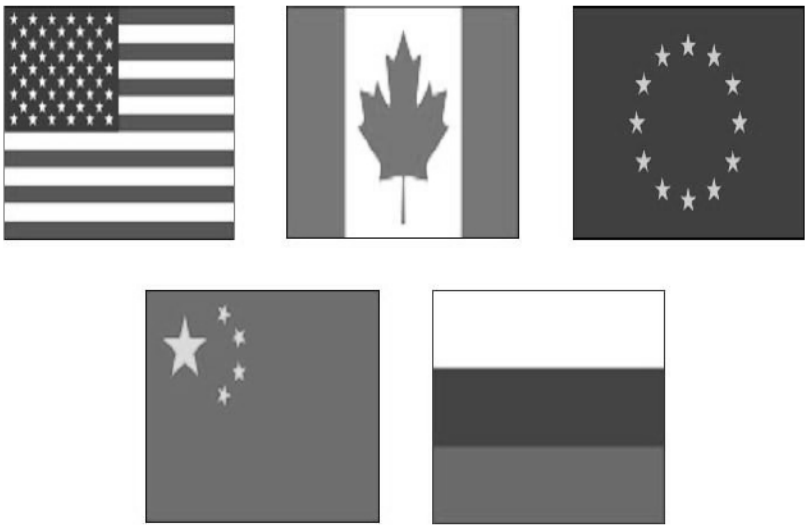


Figure 2.8: Actual sources of the flag example. The original flags images were taken from Wikipedia.

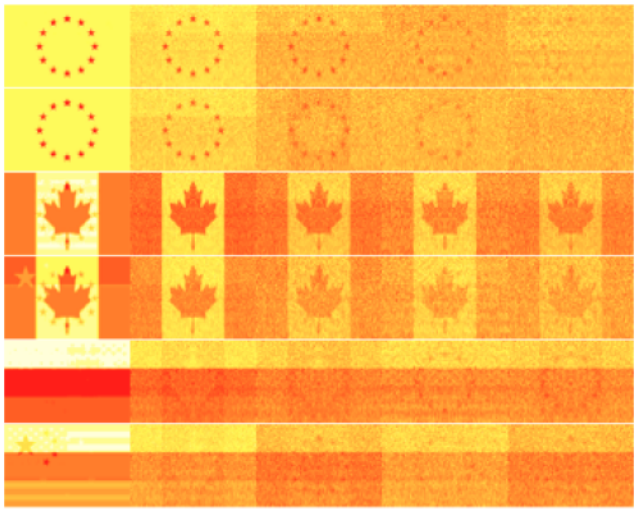


Figure 2.9: The estimated homotopic sources of H-gICA and ordinary gICA. The five columns, from left to right, contain the estimated ICs with increasing noise. The 1_{st}, 3_{rd} and 5_{th} rows shows the estimated sources extracted by H-gICA while the 2_{nd}, 4_{th} and 6_{th} rows shows the estimated sources extracted by ordinary gICA.

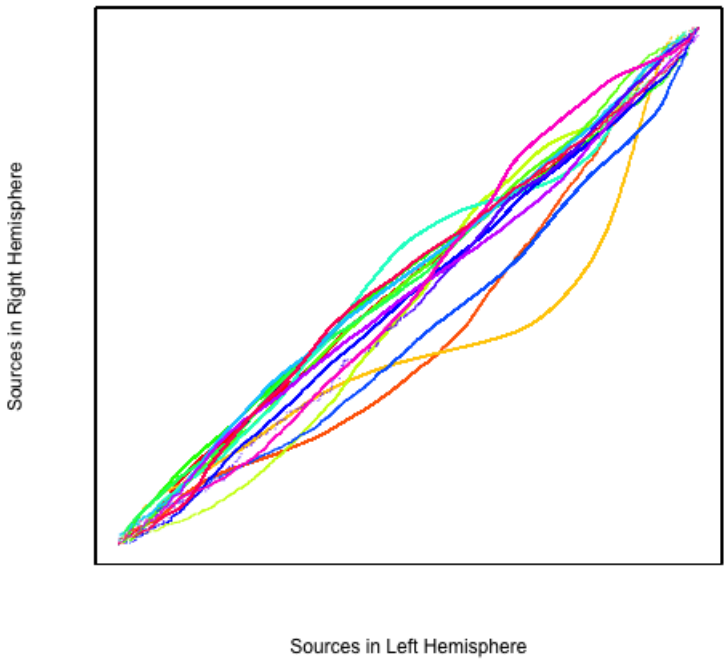


Figure 2.10: QQ plot of the 15 sources extracted by gICA.

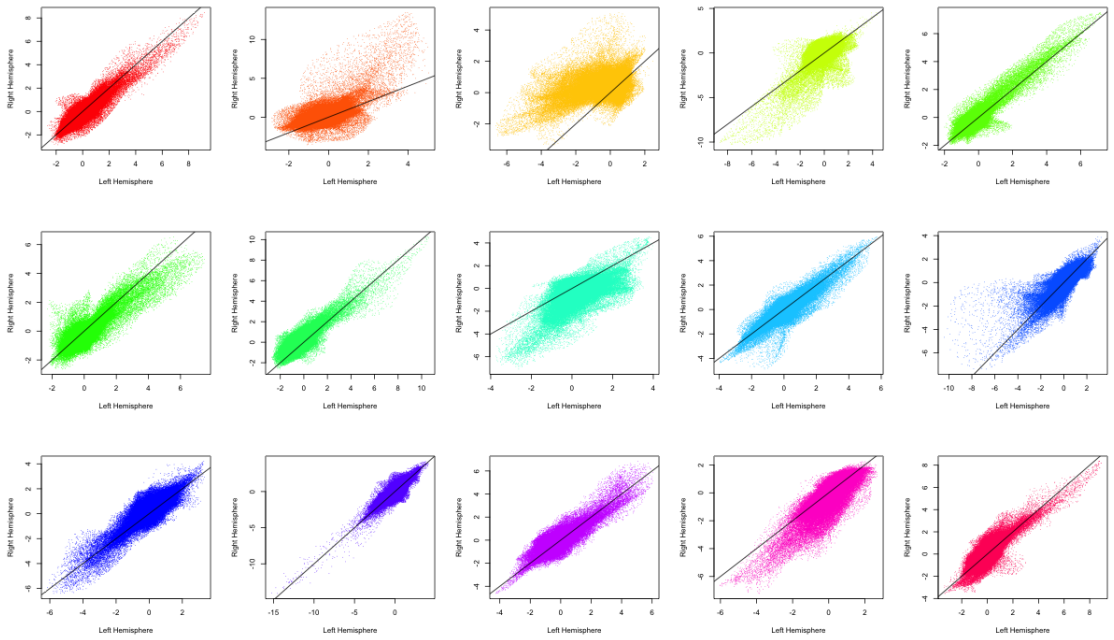


Figure 2.11: The scatter plot of the 15 sources.

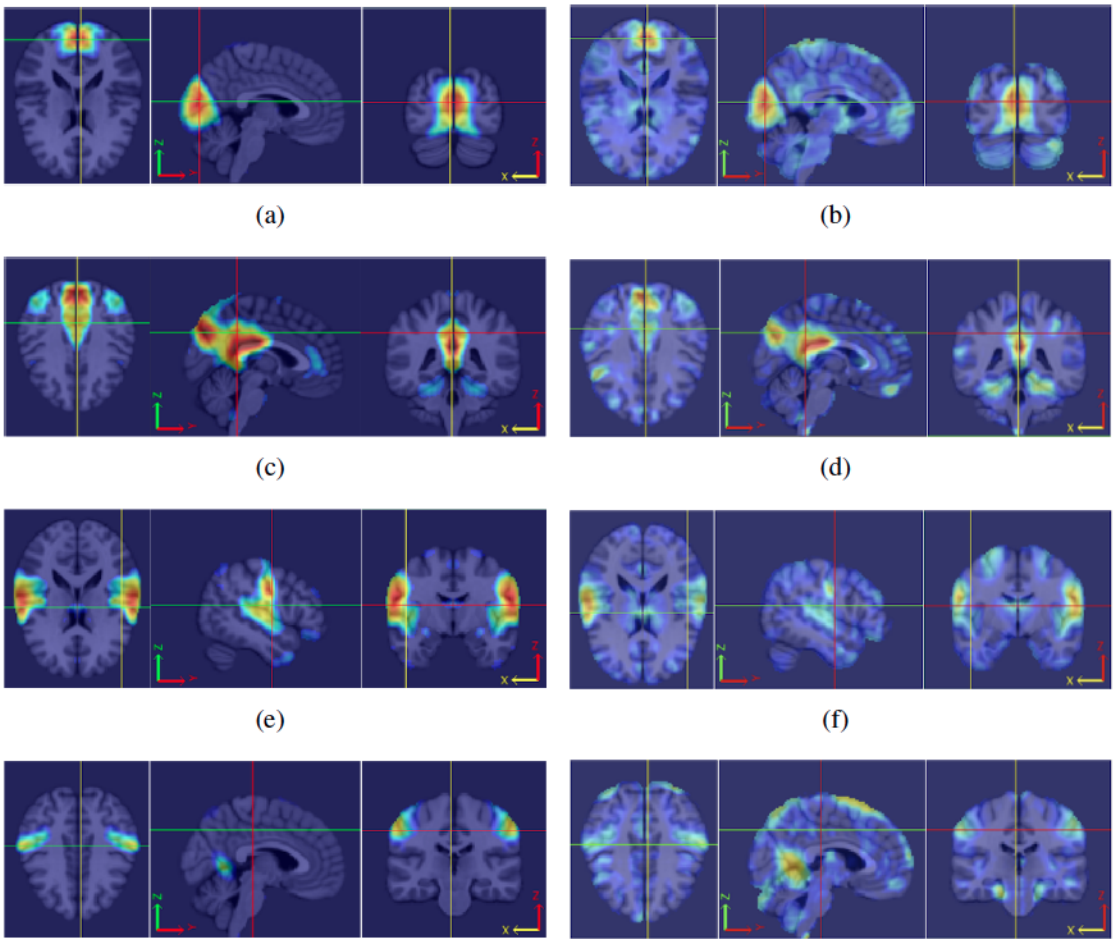


Figure 2.12: Comparison of ICs obtain from H-gICA (left column) and ordinary gICA (right column).

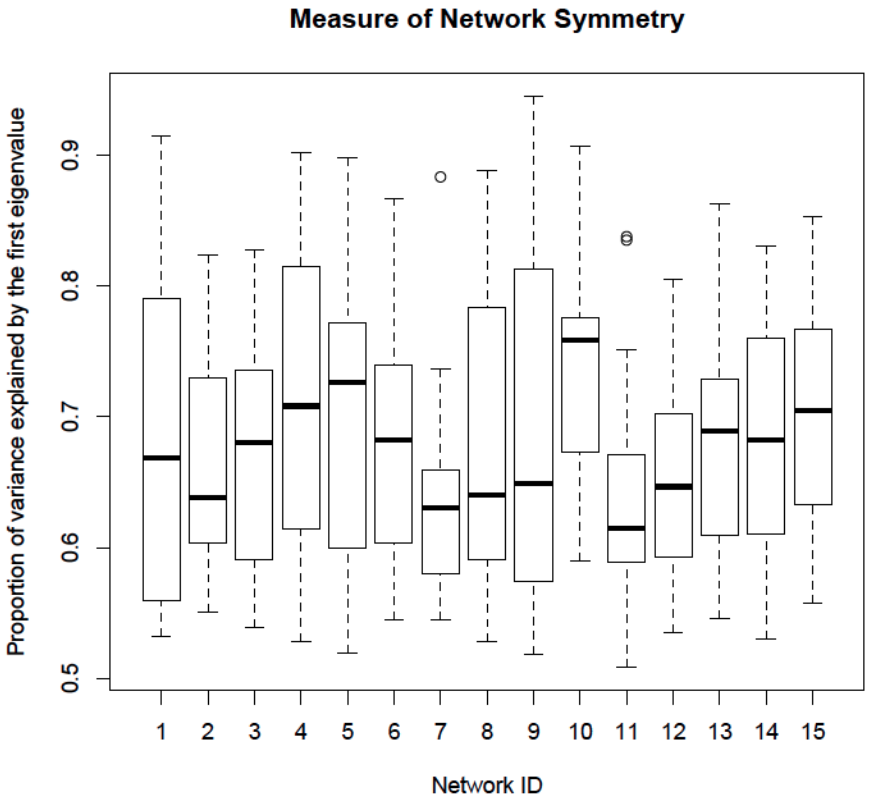


Figure 2.13: Network symmetry measured by the proportion of variance explained by the first eigenvalue of $\mathrm{cov}(\boldsymbol{A}_{i,1}(\cdot, q)), \boldsymbol{A}_{i,2}(\cdot, q)))$. Network 1, 7, 8 and 11 are the visual, default mode, auditory and motor network respectively.

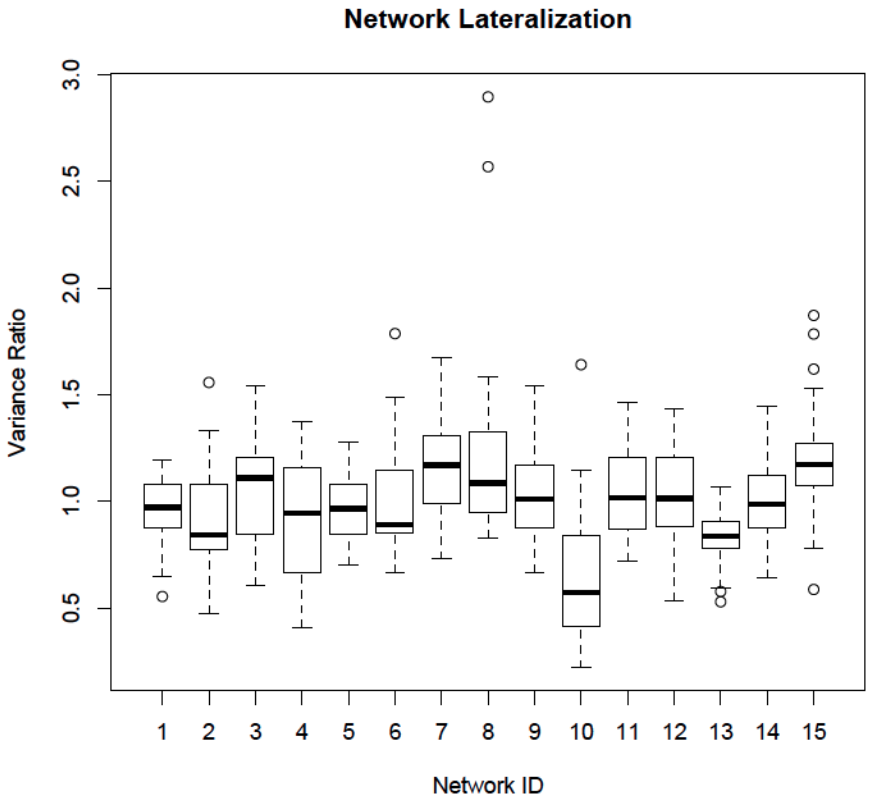


Figure 2.14: Network lateralization measured by $\text{var}(\boldsymbol{A}_{i,1}(\cdot, q))/\text{var}(\boldsymbol{A}_{i,2}(\cdot, q))$. Network 1, 7, 8 and 11 are the visual, default mode, auditory and motor network respectively.

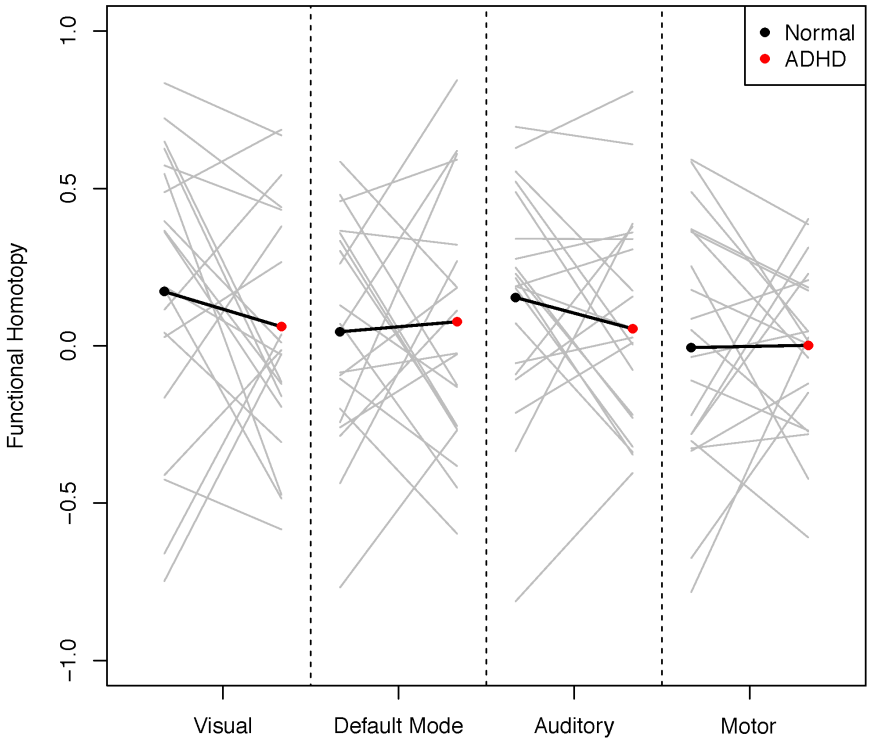


Figure 2.15: Comparison of functional homotopy of ADHD and normal developed children. Each column represents a network (visual, default mode, auditory and motor). Each pair (subjects and controls) are connected via a grey line. The left end points of the grey lines measure the functional homotopy of the control and the right end points are for the ADHD subjects. The black lines represent the group level functional homotopy for the four networks.

Chapter 3

On tests of activation map dimension for fMRI-based studies of learning

CHAPTER 3. ON TESTS OF ACTIVATION MAP DIMENSION FOR FMRI-BASED STUDIES OF LEARNING

Abstract

A methodology for investigating learning is developed using activation distributions, as opposed to standard voxel-level interaction tests. The approach uses tests of dimensionality to consider the ensemble of paired changes in voxel activation. The developed method allows for the investigation of non-focal and non-localized changes due to learning. In exchange for increased power to detect learning-based changes, this procedure sacrifices the localization information gained via voxel-level interaction testing. The test is demonstrated on an arc-pointing motor task for the study of motor learning, which served as the motivation for this methodological development.

The proposed framework considers activation distribution, while the specific proposed test investigates linear tests of dimensionality. This paper includes: the development of the framework, a large scale simulation study, and the subsequent application to the motivating study of motor learning in healthy adults. While the performance of the method was excellent when model assumptions held, complications arose in instances of massive numbers of null voxels or varying angles of principal dimension across subjects. Further analysis found that careful masking addressed the former concern, while an angle correction successfully resolved the latter. The simulation results demonstrated that the study

of linear dimensionality is able to capture learning effects. The motivating data set used to illustrate the method evaluates two similar arc-pointing tasks, each over two sessions, with training on only one of the tasks in between sessions. The results suggests distinctions in training based learning via different activation distribution dimensionality when considering the trained and untrained tasks separately. Specifically, the untrained task evidences greater activation distribution dimensionality than the trained task. However, the direct comparison between the two tasks did not yield a significant result. The nature of the indication for greater dimensionality in the untrained task is explored and found to be non-linear variation in the data.

3.1 Introduction

This manuscript considers settings where task-related activation may be present before and after learning, yet the distribution of activated voxels changes. For context, consider the motivating study for the work, where two motor tasks of equal difficulty were performed in a scanner over two sessions. Training for one of the tasks occurred in between the sessions, while the other task served as a control. Current methodology would use random effects statistical parametric mapping [SPM Friston et al., 2011] to test for a differential effect of training between tasks to study learning. However, this approach suffers from considering only voxel-level activation, or change in activation, in isolation. In contrast, learning may induce changes in *activation distribution*, i.e. the distribution of intensities of BOLD responses to the paradigm. Moreover, activation distribution offers many potential benefits over voxel-level testing, including: the elimination of multiplicity concerns, robustness to registration, and sensitivity to hypotheses of particular interest in the study of learning.

Analysis of dimensionalality of fMRI task-based activation maps [Zarahn, 2002, Worsley et al., 1997] provides a starting framework. The proposed procedure considers the distribution of activation maps and tests the dimensionality using eigenvalue decompositions. To illustrate the goals of the test, consider our motivating example. Learning could manifest itself in many ways in the collection of voxels that are activated. For example, BOLD contrast estimates in the activated voxels could be identical across sessions, increased or decreased, change from activated to not (and vice versa) or uncorrelated. The test of dimensionality should be considered one of several possible probes to interrogate

such hypotheses.

Our investigation includes a large scale simulation study of brain activation maps. The simulation results demonstrate that the study of dimensionality in a framework similar to Zarahn [2002] is able to capture learning effects. The motivating data set is used to illustrate the method, which is applied to the trained and untrained tasks separately and then jointly.

3.2 Methods

We will use the motivating study to develop the methods within a context. Recall that subjects performed an fMRI motor task on two scanning sessions, with training between them. A second, similarly difficult, fMRI motor task was performed at the two sessions, but had no training in between. We focus on activation maps within an appropriately selected spatial mask, such as one encapsulating the primary motor cortex. Let $\hat{\gamma}_{ijk}(v)$ be the subject- (represented by index $i = 1, \ldots, N$), session- ($j = 1, 2$), task- ($k = 1, 2$) and voxel- ($v = 1, \ldots, V$) specific estimates of task activation. These are obtained by voxel-wise regression of a HRF-convolved task paradigm in registered space [see Lindquist et al., 2009, Lindquist, 2008, for descriptions and discussion], conducted separately for each subject's visit.

This paper is concerned with the statistical analysis of, and hypotheses associated with, the collection of subject-specific activation maps, represented by the $V \times 2$ matrix $\hat{\Gamma}_{ik} = \{\hat{\gamma}_{i1k}(v), \hat{\gamma}_{i2k}(v)\}_{v=1}^{V}$.

A conceptual model is considered where the activation maps are estimates of assumed true activation maps, $\Gamma_{ik} = \{\gamma_{i1k}(v), \gamma_{i2k}(v)\}_{v=1}^{V}$. Thus, variation in the elements of Γ_{ik} is (intra-subject) biological variation in the hemodynamic BOLD response to the paradigm. In contrast, variation in $\hat{\Gamma}_{ik}$ includes this biological variation, as well as all of the variation and biases that occur in the practical process of computing the BOLD paradigm response.

Both $\hat{\Gamma}_{ik}$ and Γ_{ik} also vary across subjects. Consider the $V \times 2$ matrix,

$$A_k = \{\beta_{1k}(v), \beta_{2k}(v)\}_{v=1}^{V}$$

as representing the population average of voxel-level activation. Here $\beta_{jk}(v) = E(\gamma_{ijk}(v))$, $j = 1, 2$. A non-zero $\beta_{jk}(v)$ indicates that, on average, subjects activated at that particular location. Treating v as being meaningfully consistent across subjects requires that appropriate template-based (or equivalent) registration has been performed. The matrix, $\hat{A}_K$, is thus a data-level estimate of A_k, obtained by taking empirical means across subjects at each voxel.

A straightforward investigation of learning for the first (trained) task arises from a sharp null hypothesis test of:

$$H_0 : \beta_{21}(v) - \beta_{11}(v) = \beta_{22}(v) - \beta_{12}(v),$$

conducted separately, voxel-by-voxel. This tests the difference in the longitudinal change in the BOLD response between the trained and untrained tasks. Comparing longitudi-

nal learning effects with a reference (untrained) task addresses non-learning based biases across sessions. The test in question is normally conducted with standard interaction tests - perhaps accounting for subject-level correlation [see Diggle et al., 2002, for a general treatment of correlated data]. Typically, the test is performed separately at each voxel, via so-called Statistical Parametric Mapping (SPM). Significance is usually ascertained with super-threshold voxel level statistics using random field theory [see Friston et al., 2011, and the references therein] or via resampling statistics [Nichols and Holmes, 2001].

This SPM approach has several benefits for the study of learning. However, it also has limitations. Notably, the approach suffers from multiplicity issues and concentrates only on focal and localized interaction hypotheses, one voxel at a time. Moreover, it is highly dependent on accurate co-registration across subjects. Little information is gained from the ensemble of voxels, except through smoothing during preprocessing.

As an alternative, examine the activation distribution. Let $D = A_2 - A_1 = \{\beta_{21}(v) - \beta_{11}(v), \beta_{22}(v) - \beta_{12}(v)\}$ be the $V \times 2$ matrix of longitudinal changes in the contrasts of interest, with its associated estimate, $\hat{D}$. The SPM approach tests whether the two entries of each row of D are the same. Suppose one instead assumes that elements of D arise from a bivariate distribution and interest lies on the ensemble of voxel-specific pairs, instead of individual voxels.

Figure 3.1 gives a conceptual diagram showing possible shapes associated with the distribution of voxel pairs. The conceptual model is informed by the idea of Gaussian mixture models [see McLachlan and Peel, 2000, for an introduction] governed by four

major areas: (A) voxels that were ”activated” (had a change across sessions) only in the trained task, (B) voxels that were activated in both tasks, (C) voxels that were activated only in the untrained task and (D) voxels that were not activated in either task.

It is the shape of (B) that is of primary interest. For instance, any shift in B above the diagonal line represents training based learning. If the shape is spherical, there is no correlation between training status and change in activation across sessions. In contrast, the more ellipsoidal the shape, the greater the correlations in activation extent across sessions.

While acknowledging that SPM operates voxel-by-voxel, and that Figure 3.1 displays voxel groups, the SPM approach would investigate each point's distance from the diagonal line, assessing significance relative to inter-subject variability. Therefore, given enough data, the SPM approach would conceptually reject for voxels in groups (A) and (C) in the cases represented by all panels. However, it would reject most of the voxels in group B in panels II and IV only. The approach would reject few of the voxels in (B) for panels I and III. Contrast this with the shape and dimensionality of (B) being constant for panels I and II together and III and IV together. Thus, to the extent that learning represents itself as changes in the shape of the activation distribution, the voxel-wise approach would not tell the complete story.

Instead, we view the shape of the bivariate distributions of points in group (B) as informative for studying changes in task activation. One key attribute to study is its intrinsic dimensionality (1 versus 2 dimensional). Ignoring groups (A), (C) and (D), one would conclude that (B) is two dimensional in panels I and II and intrinsically one dimensional in III

and IV. The dimensionality of (B) is useful for differentiating whether changes in intensity or distribution account for activation changes following learning.

The use of principal components to investigate the dimensionality in learning studies builds upon an existing literature on the use dimensionality testing in the study of activation maps [Worsley et al., 1997]. Specifically, Zarahn [2002] promoted this work in activation studies, while Moeller and Habeck [2006] considered it within the context of functional imaging. The aim of this work is to study the goals, limitations and hypotheses of tests of dimensionality of the paired fMRI activation maps. A test of one versus two dimensions on the set $\hat{D}$, that is rank$(\hat{D})$, investigates the null hypothesis

$$H_0 : \beta_{21}(v) - \beta_{11}(v) = c\{\beta_{22}(v) - \beta_{12}(v)\}$$

for unspecified c and for all voxels v.

Let $\hat{A}_k = \frac{1}{N}\sum \hat{A}_{ik}$ and recall that $\hat{D} = \hat{A}_2 - \hat{A}_1$. Following the existing work on tests of dimensionality in fMRI, we use root tests of the second eigenvalue [see Mardia et al., 1980] to investigate the hypotheses of one dimension versus two. A simulation-based investigation of this test follows. The simulation study includes: the strength of the effect, the intrinsic dimensionality (considering power and error rates), and the impact of biological and measurement variation, including variation in the angle of the subject-specific principal direction.

3.3 Materials and Simulation

3.3.1 Motivating Data Set

The motor learning study served as motivation for this work, though we emphasize that the methodology generally applies to any study of change in activation. The goal of the underlying study centered on investigating skilled motor learning via an Arc Pointing Task [Shmuelof et al., 2012], where the task was designed to better understand neural correlates of motor skill acquisition. The subjects completed two similarly demanding motor tasks of drawing an arc within reference lines by moving their (non-dominant in all cases) left wrist. The interior circles in Figure 3.2 represent the starting and end points of the path. Subjects were directed to stay within the lines of the outer circles while tracing the arc. Subjects were scanned while performing the tasks as a baseline and again 5 days later, with training on just one of the two tasks in the interim. Comparison of fMRI activation (or any measurement of motor function) from baseline to follow-up considers both effects related to motor learning and those related to changes between sessions. Comparison with the, otherwise similar, untrained task as a reference eliminates additive inter-session biases unrelated to learning.

The specifics of the trial are as follows. Thirteen right-handed subjects (8 females, 18-27 years of age) engaged in the above described motor tasks, none having performed these tasks previously. Subjects participated in a five day protocol consisting of daily behavioral sessions in the lab and two fMRI scans on the baseline and final days (1 and 5

respectively). During scanning, subjects performed the APT whereby they performed the required arc movements per training in the behavioral sessions as well as control APT movements that were not taught. Horizontal (trained) and vertical (untrained, control) APT movements were performed in separate block design experiments before and after training for the horizontal task. Six movements were performed in 18 blocks (repeated 6 times), at a slow speed (1.5 seconds per movement). During block movements, subjects received online feedback regarding the position of the cursor, but no further information about their success or failure, or about their movement speed. In the trained task, targets were presented on the horizontal line (similarly to their configuration during the behavioral pattern in the lab) and in the untrained task, targets were aligned vertically. Movements were always in the clockwise direction. Subjects performed the movements with their (non-dominant) left wrist, while lying on their back, and receiving visual feedback of their movements through goggles (resonance technology, Los Angeles, CA). Further details on the experimental paradigm can be found in Shmuelof et al. [2014].

Data was acquired on a Philips Intera 3T scanner using a Philips SENSE head coil. The functional scans were collected using a gradient echo EPI, with voxel size of $3 \times 3 \times 3\ mm$ ($240 \times 240 \times 240\ mm$ matrix). TR = $2\ s$, flip angle = 77^o, axial slices, TE = $25\ ms$. 40 slices were gathered in an interleaved sequence at a thickness of $3\ mm$ (no gap). 96 volumes were accumulated in each experimental run. The first 2 volumes were discarded to allow magnetization to reach equilibrium. A single T1-weighted anatomical scan was also obtained for each subject (MPRAGE, $1\ mm^3$).

Functional data was preprocessed using SPM5 (Friston et al. 1999). Before statistical analysis, this data was also corrected for slice timing acquisition and head motions, re-sliced to $2 \times 2 \times 2\ mm$ voxels using a fourth degree B-spline interpolation, and transformed into a Talairach standard space [Talairach and Tournoux, 1988]. A general linear model was used for data analysis, followed by calculation of beta maps. Scatter plots of beta before training and after training are shown in Figure 3.3 and 3.4.

Interest lays in comparing the impact of training on activation related to the task. By comparing the trained and untrained tasks, the population impact of learning in specific is estimated by considering differences in the change in activation maps over sessions. Using the developed notation, the collections compared are, $\{\beta_{21}(v) - \beta_{11}(v)\}_{v=1,...V}$ to $\{\beta_{22}(v) - \beta_{12}(v)\}_{v=1,...,V}$, where, as previously noted, the first index indicates session (baseline and fifth day) and the second indicates task (horizontal and vertical). The test of dimensionality then considers whether the changes in activated voxels after training is uncorrrelated from the change in the untrained (but otherwise similar) task. Under Gaussian assumptions, absence of correlation among activated voxels implies that the extent of activation is unrelated between sessions.

All subjects gave written, informed consent and received a small compensation for participating in the Study, which was approved by the Columbia University Institutional Review Board.

3.3.2 Simulation Study

Assume there are $V = V_1 + V_2$ voxels in total: V_1 that are significantly different across sessions (group B in Figure 3.1) and referred to as "activated", and V_2 that are not (group D in Figure 3.1). Under this working example, the term activated implies a non-zero change in the contrast values across sessions. Thus, $\pi = \frac{V_2}{V}$ is the percentage of non-activated voxels. The simulation model is:

$$b_{iv} \overset{iid}{\sim} N\left\{ \begin{pmatrix} \beta_{21}(v) - \beta_{11}(v) \\ \beta_{22}(v) - \beta_{12}(v) \end{pmatrix}, I\sigma^2 \right\} = N(\delta(v), I\sigma), \tag{3.1}$$

where $\delta(v) = \{\delta_1(v), \delta_2(v)\} = \{\beta_{21}(v) - \beta_{11}(v), \beta_{22}(v) - \beta_{12}(v)\}$ and $b_{iv} = \{b_{1iv}, b_{2iv}\}$ is a subject-specific realization plus noise. The generation of the $\delta(v)$ parameters varied across simulation settings, and is described separately for each case below.

In all simulation settings, the estimate of the $V \times 2$ matrix of the $\delta(v)$, labeled $\hat{D}$, was obtained via the voxel-specific mean across subjects. Following Worsley et al. [1997], the $V \times 2$ matrix, Z, denotes $\hat{\delta}$ divided by its standard error. That is, $Z_k(v) = \text{Var}\{\hat{\delta}_k(v)\}^{-1/2}\hat{\delta}_k(v)$ make up row v and column k of Z. Here the variance was calculated across subjects separately for each voxel. The cross-product matrix is then

$$S = \sum_{v=1}^{V} Z(v)'Z(v)/V.$$

The Lawley/Hotelling trace statistic is:

$$S_q = \sum_{j=q+1}^{h} \lambda_j/(h-q),$$

where λ_j, $j = 1, 2, \cdots, h$ are the eigenvalues of S, h is the total number of eigenvectors and q is the testing rank. Under independence and Gaussian assumptions, S_q follows an F distribution under the null hypothesis, where the first q principal components capture all of the signal. In our case, $h = 2$, $q = 1$ and the test statistic is simply the second eigenvalue of S.

Simulation Under the Null Hypothesis

The first simulation setting considers the hypothesis of unidimensionality; that is, whether $\delta_1(v) = c\delta_2(v)$, where c is constant across subjects. The parameter $\delta_1(v)$ for the activated voxels was simulated as uniformly distributed in $[min, max]$, with this range computed from values of $[0, 1]$ to $[10, 15]$. Note that for voxels inactive in both time points, $\delta_1(v) = 0$. Thus $\delta_1(1), \ldots, \delta_1(V_1) \neq 0$ while $\delta_1(V_1 + 1), \ldots, \delta_1(V) = 0$. Note that, $\delta_2(v) = c\delta_1(v)$ regardless of null status.

Figure 3.5 shows example data for a simulated subjects as well as the estimated statistics. The null simulation varied according to the following: $i)$ distance of the activated voxels from the inactivated ones, as well as the range of activation, (controlled by min and max); $ii)$ the percentage of inactivated voxels (π); and $iii)$ the number of subjects (N). For all of the null hypothesis scenarios, $c = 1$. The type I error rates correspond to the

percentage of rejections of the Lawley/Hotelling trace statistic for each simulation setting. The specifics of each scenario are described below while the results are shown in Table 3.1.

Simulation under variation in the distance: In this scenario, $N = 12$, $V_1 = 40$, $V_2 = 200$, and $\sigma = 1$. Five scenarios for each pair of min and max were considered. The results suggest that the type I error is not significantly affected by the distance of the activated voxels from the inactivated ones.

Changing the percentage of inactivated voxels: In this case, $N = 12$, $\sigma = 1$, $min = 0.5$ and $max = 1.5$. The total number of voxels was set at $V = 240$. These results suggest that the test is not significantly affected by the percentage of inactivated voxels.

Varying the number of subjects: In this case, $V_1 = 40$, $V_2 = 200$, $\sigma = 1$, and $[min, max] = [0.5, 1.5]$. The results imply that the type I error does not change significantly as N varies.

Simulation Under the Alternative Hypothesis

There are a variety of ways in which the null hypothesis can fail to be true; herein, several key departures were analyzed. First, consider a straightforward departure, where Figure 3.1 holds, with sets (A) and (C) both empty. The extent of spherical and elliptical variation around the principal axis are evaluated. However, other departures could also be present. Most importantly, the null could be true for each subject, but with a varying angle

along the principal axis. In addition, a non-trivial percentage of voxels changing activation status (i.e. sets (A) and (C) from Figure 3.1 being non-empty) would similarly represent a departure from the null hypothesis. The simulation scenarios for these parameters are described below.

The number of subjects remains $N = 12$ while $min = 0.5$, $max = 1.5$, $V_1 = 40$ and $V_2 = 200$.

Simulation under a basic alternatives: Two basic alternative settings were considered. In the first, the $\delta(v)$ were simulated as two dimensional, yet one dimension dominates the other. This method of simulation added orthogonal variation around the line used in the simulation under the null hypothesis. Specifically, the activated voxels have Gaussian variation orthogonal to the major axis (see Figure 3.6(a)). This was done in lieu of simulating a bivariate Gaussian with a non-zero correlation to consider an even, non-concentrated spread along the major axis. Simulations using a bivariate normal yielded similar results. In the second setting the correlation was assumed to be zero (see Figure 3.6(b)).

Variability of the angle of the principal axis Consider a null setting, as in Section 3.3.2. However, assume that the constant, c, varies across subjects. Let c_i denote this constant for subject i. To simulate the data, first the null simulation from Section 3.3.2 was performed then the observed bivariate points $\{b_{1iv}, b_{2iv}\}$ were multiplied by the rotation matrix $\begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}$, where θ_i is a subject-specific rotation angle

from the 45^o line, generated from a Gaussian distribution with mean 0 and standard deviation σ_a, which varied from 0.01 to 0.5. Before the rotation, $c = 1$, while afterwards, $c_i = \tan(45^o - \theta_i)$. Examples of the simulated data are shown in Figure 3.7.

Changing Activation Sets In this setting, the impact of a non-trivial percentage of voxels, or change in voxels that switch activation status, i.e. corresponding to a large collection of voxels in sets (A) and (C) in Figure 3.1. An example simulation is shown in Figure 3.8. Here, the b_{kiv} were either 0 or uniform, where a $min = 0.5$ and $max = 1.5$. The specific values were: 200 voxels set to be inactive for both the trained and untrained tasks, 40 voxels set to be activated for the trained and untrained groups, V_a voxels were activated with training, but inactivated without training, while another V_a voxels were inactivated with training but activated without training. Here V_a was varied between 10 and 400. Note that, in this setting, the Z matrix (see Figure 3.8) is substantially different from the direction of its first eigenvector.

3.3.3 Simulation Results

Table 3.1 displays the results across the simulation settings. All tests were performed at a nominal 5% error rate.

Simulations Under the Null Hypothesis

Adherence to the specified nominal error rate was remarkably consistent as parameter settings varied. When varying the distance, the test showed only slight liberalism (Type I error rate larger than the nominal) across settings. Only for unrealistically small activation sets did the test demonstrate liberalism when altering the activation set size. In addition, varying the number of subjects had little impact. Adherence to the nominal error rate was acceptable, even at very low numbers of subjects.

Simulations Under the Alternative Hypothesis

Under the basic alternative, where the true voxel states possessed a strong (but not perfectly linear) correlation, power varied as expected. Under a strong correlation (σ_b close to 0), power trended to the nominal type I error rate. Encouragingly, power quickly trended to one as the true relationship moved away from a dominant dimension. As expected, the power tended to 1 as the sample size increased (confirming the relevant asymptotics). However, the sample size needed to be relatively large to have adequate power at the modest value of $\sigma_b = 0.2$.

In the case where no dimension dominated under the basic alternative of absence of correlation, power changed significantly with the spread of activation, σ_b. When the angle of principal direction varied, power suffered dramatically. To address this, a first stage subject-specific principal components rotation was investigated. This appeared to improve power in settings where the null and non-null voxels were more clearly delineated, but

continued to exhibit low power (11%) when the distance was large ($\min = 10$, $\max = 15$). A non-trivial fraction of voxels changing activation status had a negative impact on power.

3.4 Data Analysis of the Motivating Data Set

This section investigates the impact of training on activation using the motivating APT data described in Section 3.3.1 and represented in Figures 3.3 and 3.4, which show estimated beta maps. A null hypotheses suggests that the data points are close to the principal line. Notably, the relationship is difficult to ascertain between the null and alternative graphically. However, it is apparent that the axis of principal direction varies by subject. Next, dimensionality is tested via three methods: first considering only the (trained) horizontal task, then only the (untrained) vertical task, and then comparing both. When considering the untrained task in isolation we are testing $H_0 : \beta_{21}(v) = c\beta_{11}(v)$, then $H_0 : \beta_{22}(v) = c\beta_{12}(v)$ for the trained and $H_0 : \beta_{21}(v) - \beta_{11}(v) = c\{\beta_{22}(v) - \beta_{12}(v)\}$ when comparing trained and untrained. (The paper used the latter as the primary motivating example.) In Table 3.2, the results before and after angle correction are shown.

3.4.1 Motivating Data Results

The axis of principal direction varied by subject (see Figure 3.3 and 3.4). Before correcting for the principal angle, the tests of dimensionality were insignificant, for both the horizontal and the vertical tasks. However, after correcting the principal angle by subject,

the p-values of the tests were highly reduced. Focusing only on the tasks separately, the test of dimensionality yielded a p-value of 0.05 for the vertical task and 0.16 for the horizontal one. When comparing across tasks, the p-value was 0.36. Thus, the untrained task has a significant second dimension that does not appear to be present in the trained. Inspecting the data, excess variability in the trained task appears to be due to biomodal changes in activation. It is not surprising that the comparison across tasks was non-significant, given the increased variability obtained from taking differences and the issues of power for the test.

3.5 Discussion

3.5.1 Simulation Results

The simulation results suggest that tests of dimensionality are a reasonable exploratory testing procedure for investigating the distribution of paired activation maps. However, their confirmatory performance was hindered by instances with low power in situations that are realistic. The adherence to the nominal type I error rate, on the other hand, was uniformly acceptable across simulation settings. Thus a rejection from this test is likely informative, while an acceptance less so.

The low power cases occurred where there is substantial variability in the principal axis, or where activation status changed. This latter condition created confusion between noise and signal, with the test attributing signal variability as noise. Of the two cases, careful

masking could eliminate concern over changing activation status. However, variability in the principal axis is likely the norm and could arise from a number of plausible biological, technological and processing causes. The straightforward refinement of a first stage subject-level principal component rotation improves the power.

3.5.2 General discussion

This manuscript posited a different paradigm for statistically evaluating learning using task-related BOLD fMRI activation maps. At its core, the primary advancement is the supposition of using the bivariate distribution of the activation maps, or changes in activation maps, when comparing tasks over sessions. Under this framework, changes in the *distribution* of activated voxels are key, not voxel level changes in activation extent, as would be evaluated in voxel-level parametric mapping interaction tests. An unintended benefit of this distributional approach in this setting is avoiding the classical issue of having to determine interactions where main effects are not present.

The intended benefit of increasing power over voxel-level interaction tests was found to be true, provided assumptions hold. For example, Figure 3.9 provides a simulation example where the alternative test of dimensionality is both true and detected (P-value of 0.03). However, only 11% of the voxels would satisfy a voxel level test of significance. However, we emphasize the different nature of the hypotheses interrogated by these approaches so that comparisons of power should be taken with a grain of salt.

Evaluating distributional differences of learning-based activation offers a different sci-

entific hypothesis than that of voxel level testing. In our motivating example, interest lies in how BOLD activation, or changes in activation, relate between trained and untrained tasks. Investigating activation distributions is less sensitive to the requirements of focal localization of effects required by interaction testing. For example, two small spatially separated significant interaction regions may have different voxel-level interaction significance than a single contiguous region of the same aggregate size. In contrast, the distribution of interest in this manuscript may not change. Conversely, evaluating contrast map distributions does not enjoy the benefits of localization to inform results.

It is worth emphasizing that the investigation of activation distribution represents a complementary procedure to voxel-level testing and does not represent a form of omnibus test to be performed prior to it.] Thus, it is perhaps not useful to generate a single analytic pipeline, whereby omnibus distributional tests are followed by voxel level contrasts of interest.

An interesting next direction in this line of research would consider full models of the joint distribution of $\{\beta_{11}(v), \beta_{12}(v), \beta_{21}(v), \beta_{22}(v)\}$. This could be accomplished using a Bayesian random effects approach via mixtures of Gaussian random variables. However, the feasibility, applicability and gain of such an approach over simpler solutions remains unknown. A tantalizing possible benefit would be robustness to inter-subject registration to a template. In contrast, interaction tests focus on localization and as such, place a heavy burden on accurate inter-subject registration. A full random effect mixture model could possibly remove the need for inter-subject registration, or at least remove the need for non-

affine registration.

The far simpler approach discussed in this manuscript addresses dimensionality. The results show that the operating characteristics of the approach are viable, if modeling assumptions are met. Particularly encouraging was the robustness to variation in the distance of the bolus of activation from null voxels. However, its sensitivity to the angle of the principal axis is a core issue, as such variation is clear from the data.

In the real data analysis it is noteworthy that the vertical and horizontal tasks differed in their respective tests of dimensionality. Particularly, the null hypothesis was not rejected in the trained task (horizontal) while it was in the untrained task (vertical). However, there does appear to be more apparent non-Gaussianity in the vertical task, suggesting a component of the rejection is related to a form of dimensionality not well covered by the model. The contrast test comparing vertical versus horizontal was not significant. Therefore, it cannot be concluded that the activation distribution given by the inter-session differences across tasks is not linear. However, for all three cases, the data analysis suggests large variability in the subject-specific principal axes, a setting where low power was evidenced in the simulation study. Thus the null results are perhaps indicative of low power.

3.6 Appendix: Tables and Figures

Table 3.1: Results of the simulation studies. Shown are type I error rates and power across simulation settings.

		V_1	V_2	N	min	max	σ		Type I error
H_0	Variation in the distance	40	200	12	0	1	1		0.051
		40	200	12	0.5	1.5	1		0S.053
		40	200	12	1.5	2.5	1		0.051
		40	200	12	3	5	1		0.059
		40	200	12	10	15	1		0.053
	Changing the percentage of inactivated voxels	20	220	12	0.5	1.5	1		0.069
		40	200	12	0.5	1.5	1		0.052
		80	160	12	0.5	1.5	1		0.062
		120	120	12	0.5	1.5	1		0.060
		200	40	12	0.5	1.5	1		0.056
	Varying the number of subjects	40	200	4	0.5	1.5	1		0.048
		40	200	8	0.5	1.5	1		0.052
		40	200	12	0.5	1.5	1		0.051
		40	200	20	0.5	1.5	1		0.058
		40	200	100	0.5	1.5	1		0.052
H_a	Basic alternatives - correlated	V_1	V_2	N	min	max	σ	σ_b	Power
		40	200	12	0.5	1.5	1	0.05	0.042
		40	200	12	0.5	1.5	1	0.1	0.059
		40	200	12	0.5	1.5	1	0.2	0.184
		40	200	12	0.5	1.5	1	0.5	0.972
		40	200	12	0.5	1.5	1	1	1.000
		40	200	12	0	1	1	0.2	0.195
		40	200	12	0.5	1.5	1	0.2	0.186
		40	200	12	1.5	2.5	1	0.2	0.196
		40	200	12	3	5	1	0.2	0.188
		40	200	12	10	15	1	0.2	0.195
		40	200	4	0.5	1.5	1	0.2	0.046
		40	200	8	0.5	1.5	1	0.2	0.101
		40	200	12	0.5	1.5	1	0.2	0.171
		40	200	20	0.5	1.5	1	0.2	0.342
		40	200	100	0.5	1.5	1	0.2	0.998
	Basic alternatives - uncorrelated	40	200	12	1	1	1	0.05	0.045
		40	200	12	1	1	1	0.1	0.065
		40	200	12	1	1	1	0.2	0.171
		40	200	12	1	1	1	0.5	0.973
		40	200	12	1	1	1	1	1.000
	Variability of the angle of the principal axis - without angle correction	V_1	V_2	N	min	max	σ	σ_a	Power
		40	200	12	0.5	1.5	0.5	0.01	0.041
		40	200	12	0.5	1.5	0.5	0.02	0.051
		40	200	12	0.5	1.5	0.5	0.05	0.056
		40	200	12	0.5	1.5	0.5	0.1	0.025
		40	200	12	0.5	1.5	0.5	0.5	0.007
		40	200	12	0	1	0.5	0.01	0.042
		40	200	12	0.5	1.5	0.5	0.01	0.045
		40	200	12	1.5	2.5	0.5	0.01	0.054
		40	200	12	3	5	0.5	0.01	0.051
		40	200	12	10	15	0.5	0.01	0.030
	Variability of the angle of the principal axis - with angle correction	40	200	12	0	1	0.5	0.01	0.005
		40	200	12	0.5	1.5	0.5	0.01	0.027
		40	200	12	1.5	2.5	0.5	0.01	0.064
		40	200	12	3	5	0.5	0.01	0.078
		40	200	12	10	15	0.5	0.01	0.109
	Changing activation sets	V_1	V_2	N	min	max	σ	V_a	Power
		40	200	12	0.5	1.5	0.5	10	0.057
		40	200	12	0.5	1.5	0.5	20	0.048
		40	200	12	0.5	1.5	0.5	40	0.057
		40	200	12	0.5	1.5	0.5	100	0.097
		40	200	12	0.5	1.5	0.5	400	0.205

Table 3.2: P-values of the tests of dimensionality for the motivating data set. The first row considers the Session 1 versus Session 2 for the Horizontal task ($H_0 : \beta_{21}(v) = c\beta_{11}(v)$). The second row does the same for the vertical task ($H_0 : \beta_{22}(v) = c\beta_{12}(v)$). The third considers inter-session differences across tasks ($H_0 : \beta_{21}(v) - \beta_{11}(v) = c\{\beta_{22}(v) - \beta_{12}(v)\}$). P-values are given with and without having performed an angle correction.

Tasks	Without angle correction	With angle correction
Horizontal Session 1 versus Session 2	0.520	0.163
Vertical Session 1 versus Session 2	0.598	0.050
Horizontal versus Vertical		0.3620

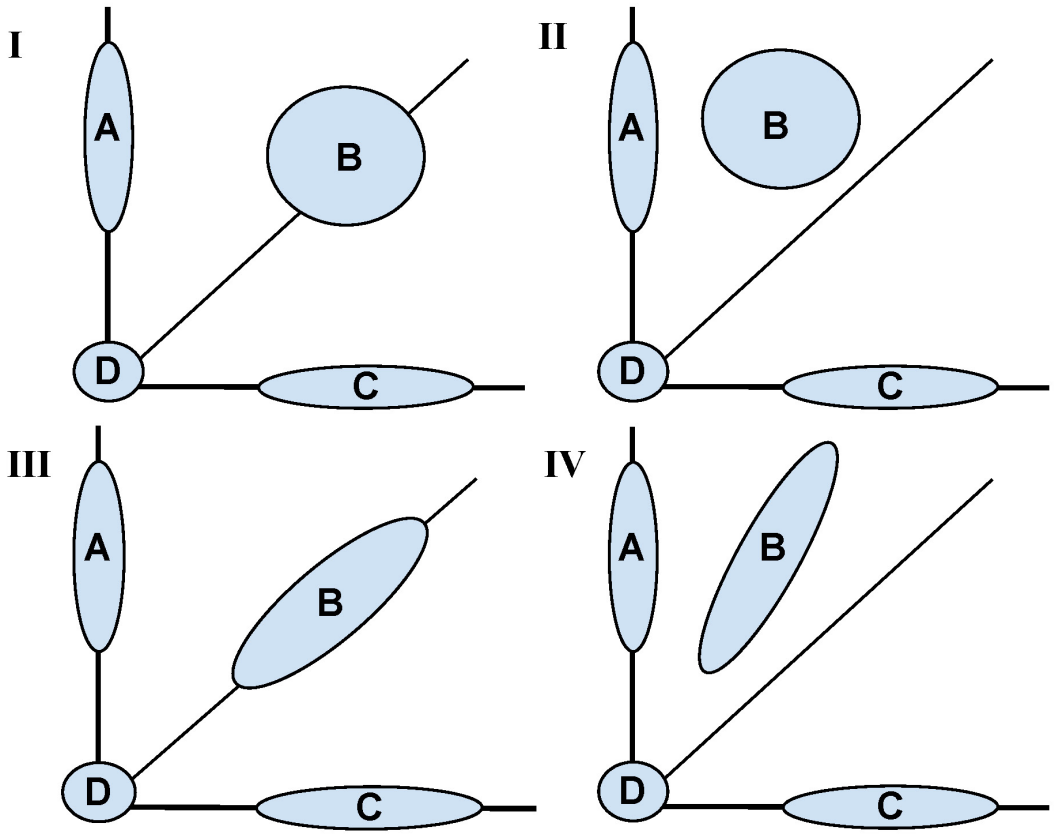


Figure 3.1: Conceptual diagram for fMRI activation distributions based on the motivating study of motor learning. Shaded areas represent learning based (inter-session differences) between a trained (Y axis) and untrained (X axis) task. Across all panels, Area (A) represents voxels with change in activation across sessions only in the trained task, (B) represents voxels with change in activation across sessions in both the trained and untrained task, (C) voxels with change in activation across sessions only in the untrained task, (D) represents no change in activation for both tasks. The four panels (I-IV) represent different potential shapes of the activation distributions for (B) with Panels I and II showing a two dimensional shape and Panels III and IV showing an approximately one dimensional. In Panels I and III inter-sessions differences are symmetrically represented whereas in II and IV one task had a uniformly greater increase.

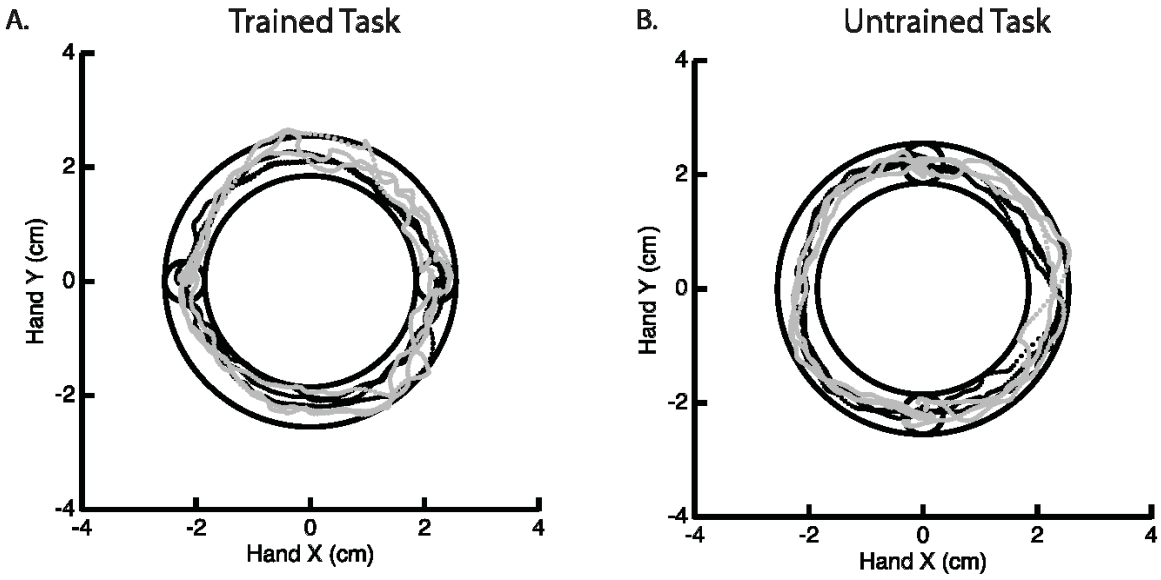


Figure 3.2: Example of the Arc Pointing Task (APT) executed within the fMRI session. Subjects were asked to navigate a cursor lying between the inner and outer concentric circles. Two tasks of similar difficulty were investigated. A horizontal task (Panel A) where subjects were trained in between two scanning sessions and a vertical task (Panel B) where subjects were not trained.

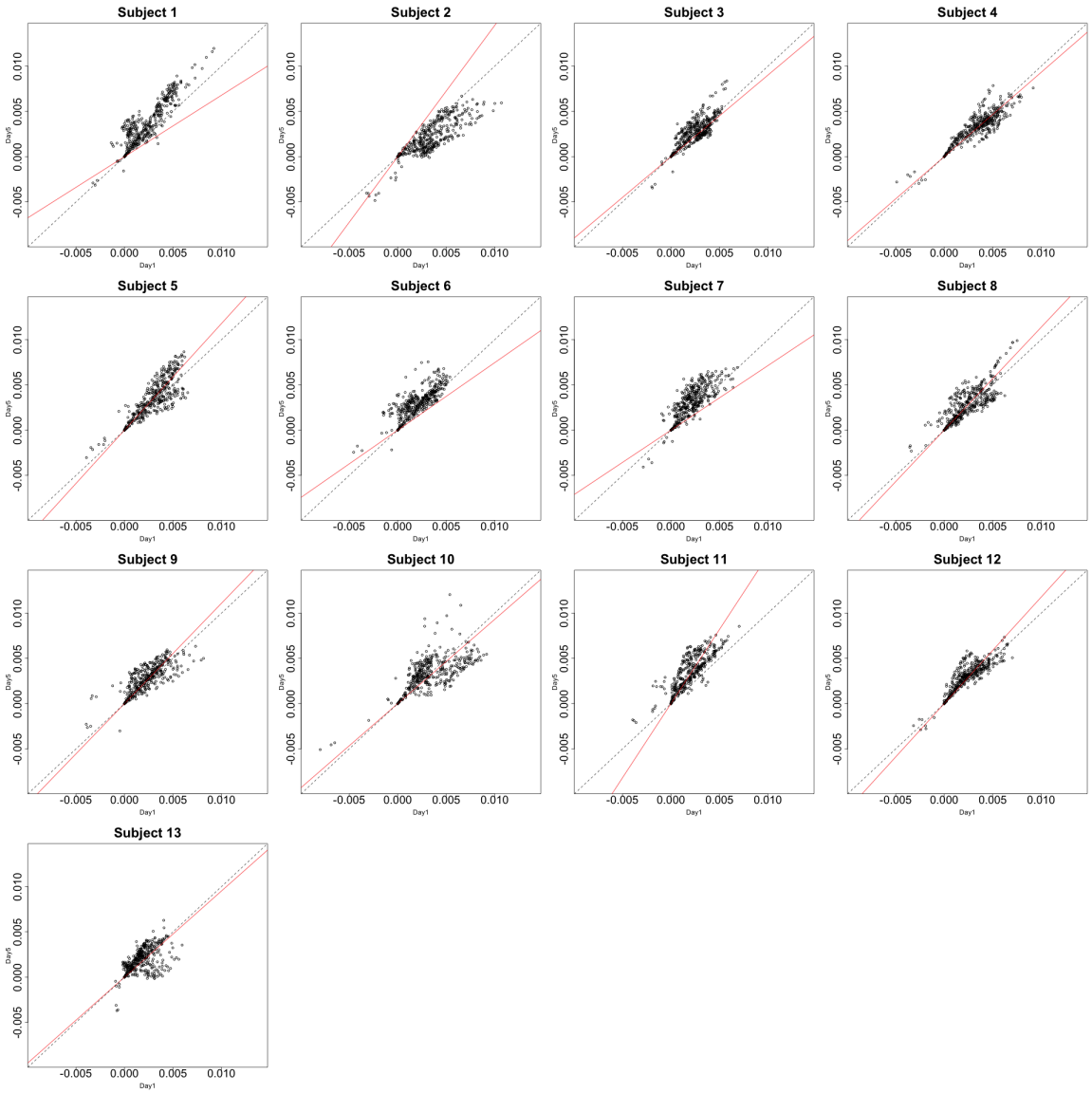


Figure 3.3: Contrast maps from the horizontal arc pointing task. The X axis for each plot is the first session while the Y axis is the second. Red lines show the direction of the first principal component while a dotted identity line is shown for reference.

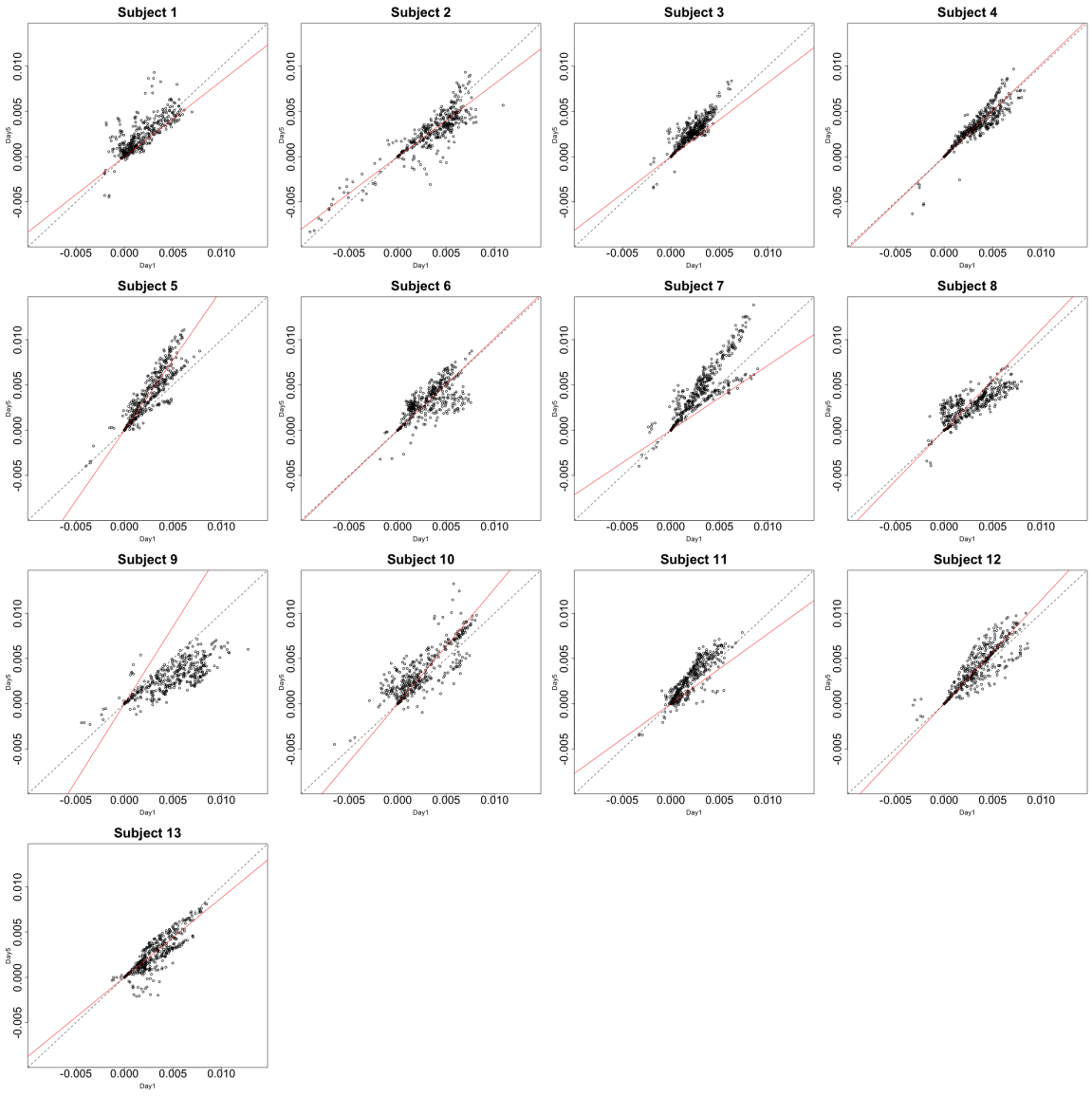
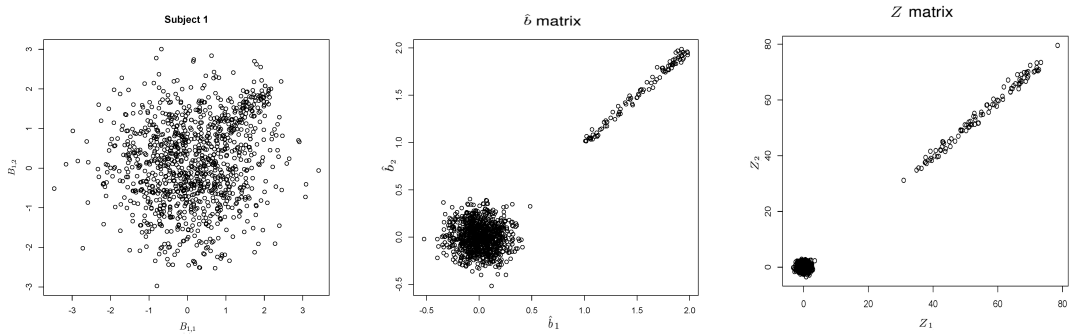
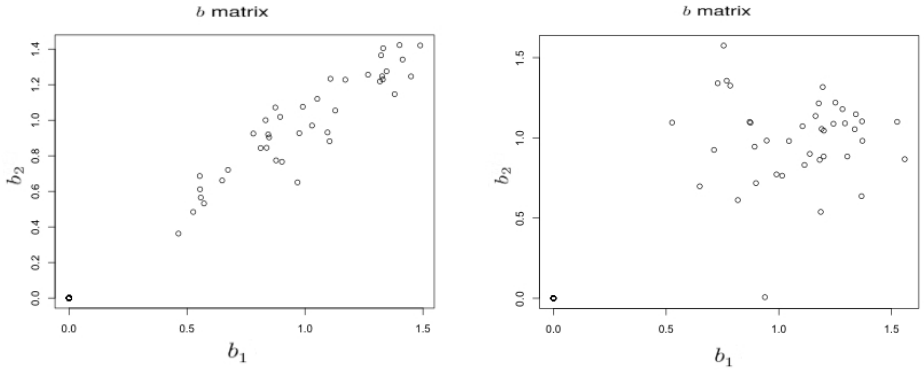


Figure 3.4: Contrast maps of the vertical arc pointing task.The X axis for each plot is the first session while the Y axis is the second. Red lines shows the direction of the first principal while a dotted identity line is shown for reference.



(a) Simulated subject data (b) Estimate of $\hat{b}$ (c) Estimate of Z

Figure 3.5: Example simulated data. Panel (a) shows the simulated data. Panel (b) shows the estimate of b using Equation (3.1). Panel (c) shows the estimate of Z, where $Z_k(v) = \text{Var}\{\hat{\delta}_k(v)\}^{-1/2}\hat{\delta}_k(v)$.



(a) Gaussian around a line (b) Gaussian around a point

Figure 3.6: Example simulation from the alternative hypothesis. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. In Panel (a) the voxels have Gaussian variation added orthogonally to the major axis. In Panel (b) there is no relationship.

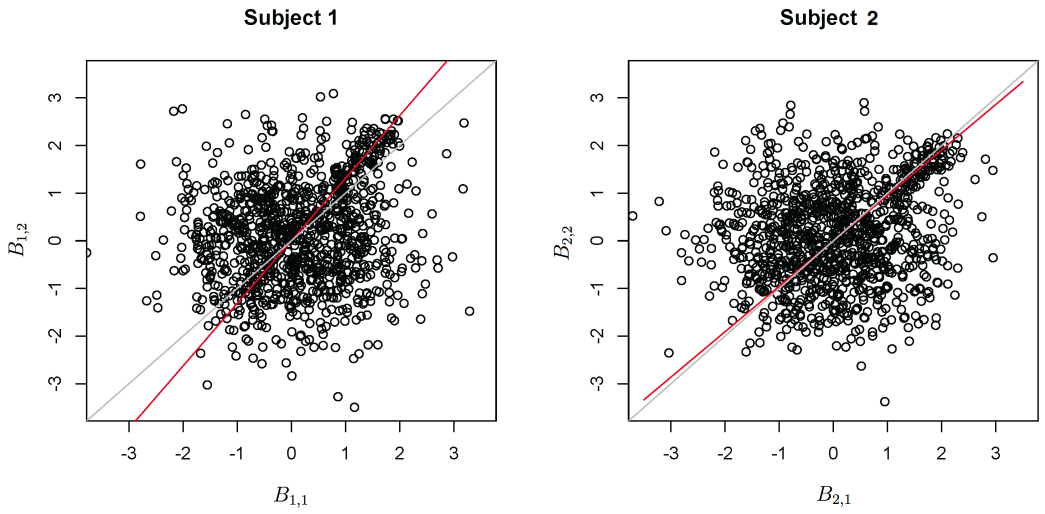


Figure 3.7: Example simulation for the setting when the principal axis differs across subjects. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study.

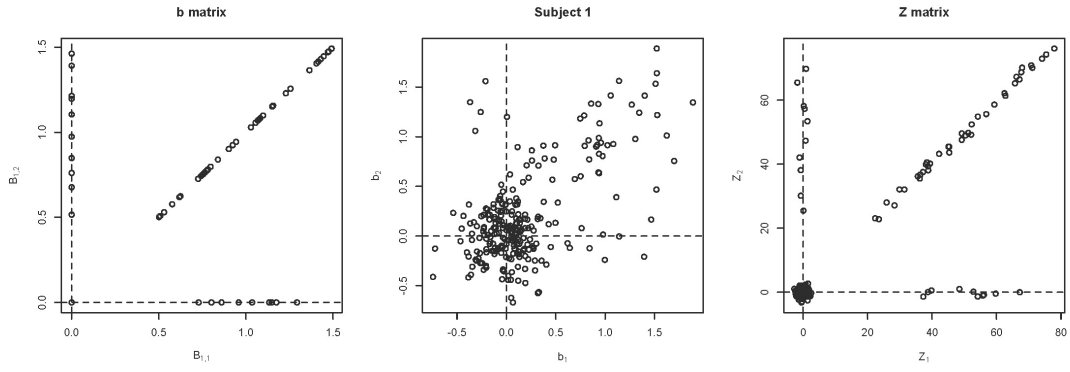


Figure 3.8: Example simulation from the alternative with changing activation sets. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. Shown are the true parameter values (leftmost panel), the simulated subject data (middle panel) and the Z values (rightmost panel).

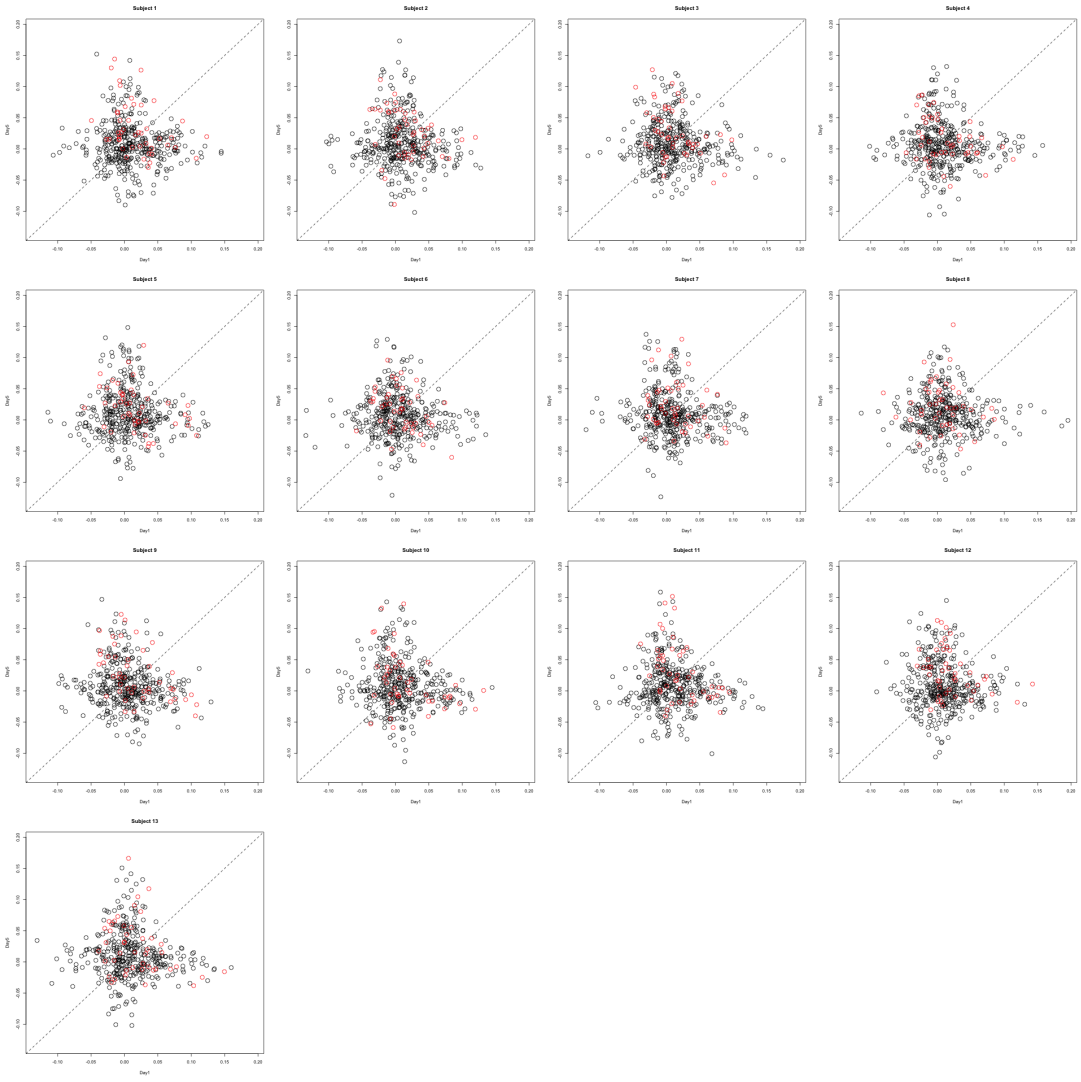


Figure 3.9: A simulation example highlighting increased power for detecting learning based differences. The axes are the two dimensional bivariate simulated data representing inter-session differences for each task in the motivating study. The alternative of the dimensionality test is true and the P-value is 0.03, suggesting that activation extent is unrelated between tasks. However, only 11% of the voxels satisfy a voxel level test of significance (colored in red).

Chapter 4

Context Aware Group Nearest Shrunken Centroids in Large-Scale Genomic Studies

CHAPTER 4. CONTEXT AWARE GROUP NEAREST SHRUNKEN CENTROIDS IN LARGE-SCALE GENOMIC STUDIES

Abstract

Recent genomic studies have identified genes related to specific phenotypes. In addition to marginal association analysis for individual genes, analyzing gene pathways (functionally related sets of genes) may yield additional valuable insights. We have devised an approach to phenotype classification from gene expression profiling. Our method named "group Nearest Shrunken Centroids (gNSC)" is an enhancement of the Nearest Shrunken Centroids (NSC) [Tibshirani et al., 2002] which is a popular and scalable method to analyze big data. While fully utilizing the variable structure of gene pathways, gNSC shares comparable computational speed as NSC if the group size is small. Comparing with NSC, gNSC improves the power of classification by utilizing the gene pathway information. In practice, we investigate the performance of gNSC on one of the largest microarray datasets aggregated from the internet. We show the effectiveness of our method by comparing the misclassification rate of gNSC with that of NSC. Additionally, we present a novel application of NSC/gNSC on context analysis of association between pathways and certain medical words. Some newest biological findings are rediscovered.

4.1 Introduction

Recent advances in DNA microarray experiment are generating data sets of the expression levels of large number of genes simultaneously. The aggregation of these data sets across experiments provides better representation of the overall population and contains more information which allows better insights into certain diseases and their causing genes. The aggregated data, however, is often of large scale, high-dimensional (*number of variables* $>$ *number of observations*), with non-Gaussian structure, and thus beyond the ability of typical analysis. This kind of data is so called "big data" [Manyika et al., 2011]. It was only recently that people have begun to develop methods of analyzing big data deriving from microarray experiments. One of the aims of such data is to identify a small subset of functional genes which discriminate between certain phenotypes such as the tumor and the normal tissues. Traditional discriminant analysis methods such as linear discriminant analysis (LDA), support vector machine (SVM), and logistic regression are either restricted to relatively small data set or not consistent under the high-dimensional situation. Take the standard LDA as an example. The standard LDA, which uses a linear combination of features as the criterion for classification, has been shown to perform well and enjoy certain optimality as the sample size tends to infinity while the dimension is fixed. In the high-dimensional settings, however, Bickel and Levina [2004] show that the classical LDA is asymptotically equivalent to random guessing when $p/(n_1 + n_2) \to \infty$, even if a Gaussian assumption is made. To handle this problem, known as "the curse of dimensionality," a sparsity condition has to be added, which leads to a variety of works: Cai

and Liu [2011] made a sparsity assumption on the precision matrix and proposed a direct estimation method for sparse LDA by estimating $\Omega\delta$ (the product of the precision matrix and the difference of the means) through a constrained l_1 minimization method; Ravikumar et al. [2010] presented a sparse logistic regression method which involves performing l_1-regularized logistic regression of each variable on the remaining variables and then using the sparsity pattern of the regression vector to infer the underlying neighborhood structure; Zhu et al. [2004] considered the l_1 norm SVM to accomplish the goal of automatic feature selection in the SVM and Friedman et al. [2004] shows that the l_1 norm is preferred if the underlying true model is sparse.

Although significant process has been made in this direction, the Nearest Shrunken Centroids (NSC) proposed by Tibshirani et al. [2002] is still one of the most scalable methods in the field of large-scale data analysis. Comparing with sparse LDA, l_1 norm SVM and sparse logistic regression, the NSC is appealing in the sense that the algorithm is much faster, able to deal with big data, and easy to implement. Moreover good empirical performances have been constantly verified in recent years [Kobayashi et al., 2011]. Particularly useful is the NSC's ability to simultaneously conduct feature selection and classification via shrinking the marginal centroids. Theoretically speaking, the NSC works the best in a Naive Bayes situation [Fan and Fan, 2008], where variables are supposed to be independent of each other; its robustness is such that a high-degree of efficiency is maintained even under more complicated high dimensional models [Fan and Fan, 2008].

In this paper, we propose a new high dimensional discriminant analysis method of group

Nearest Shrunken Centroids (gNSC). Our new method is one of the earliest attempts to deal with big data of microarray expressions with context information. Also, this is the first paper provide theoretical justification for NSC like methods. Similar with the NSC, gNSC can simultaneously perform sample classification and feature selection. The non-gaussianity of the data is overcome by conducting a normal score transformation in data prepocessing. This has been discovered to work well when the true data are coming from the Nonparanormal [Liu et al., 2012] and lose little when the true data are indeed Gaussian. Moreover, in addition to marginal association analysis for individual genes, gNSC enables us to use gene pathway information. Genes work independently and interactively to perform various biological functions. A gene pathway refers to a set of genes that work together to finish a specific biological function. Utilizing the variable structure information of gene pathways may lead to valuable insight into the disease etiology or treatment effect and could inform clinical decisions concerning disease prevention or therapeutic maneuvers. Furthermore, when multiple genes from a same pathway show concerted signals, there may be enhanced the power of sample classification, which is convinced in our experiment. We test the effectiveness of gNSC in analyzing big data against GPL96, a large-scale microarray dataset aggregated from the internet [McCall et al., 2010]. Pathway information for the genes is extracted from Molecular Signature Database (MSigDB) [Subramanian et al., 2005]. We compare our gNSC with NSC to show that gNSC improve the power of sample classification by utilizing the pathway information. We also apply gNSC to a context analysis where we combine the sample text information into the GPL96 data [McCall et al., 2010].

Our results are consistent with the newest biological finding: the expression of MYC target genes is correlated with B cell lymphomas and Wilms tumor [Ji et al., 2011].

We arrange the rest of the paper as follows. In Section 4.2, sees the introduction of the Nearest Shrunken Centroids (NSC) proposed by Tibshirani et al. [2002] and normal score transformation [Liu et al., 2009]. In Section 4.3, the theoretical body, sees our group Nearest Shrunken Centroids method. We prove some theoretical properties of gNSC. Notably, we prove that (i) under certain regularity conditions, the sparsity pattern can be recovered in an exponential rate; (ii) under certain conditions, we prove that $\mathcal{C}(g) - \mathcal{C}(g^*) = O_P(n^{-1})$, where we denote by $\mathcal{C}(g^*)$ and $\mathcal{C}(g)$ the Bayes risk and the gNSC misclassification rate. We also show the semiparametric efficiency of performing normal score transformation in data preprocessing. In Section 4.4, we apply both our gNSC method and our context analysis algorithm to the GPL96 microarray dataset.

4.2 Background

4.2.1 Nearest Shrunken Centroids (NSC)

Tibshirani et al. [2002] proposed the Nearest Shrunken Centroid method for sample classification in DNA microarray studies. They use shrunken centroids as prototypes for each class and identify subsets of genes that best characterize each class. The NSC shrinks each of the class centroids toward the overall centroid for all classes by a threshold and makes the classifier more accurate by eliminating the effect of noisy genes. As a result it

also has an internal gene selection facility [Zou and Hastie, 2005]. In detail, given x_{ij} for variable j and sample i where $j = 1, .., d$ and $i = 1, .., n$, we have M classes, each with n_m samples. $i \in C_m$ means that the i-th sample is in class m. The NSC utilizes the simple two sample t-test statistic between $\{x_{ij}, i \in C_m\}$ and $\{x_{ij}, i = 1, .., n\}$, and define the classification score d_{mj} as:

$$d_{mj} := \frac{\bar{x}_{mj} - \bar{x}_{\cdot j}}{\eta_m \cdot (s_j + s_0)}, \tag{4.2.1}$$

where $s_j^2 = \frac{1}{n-M} \sum_m \sum_{i \in C_m} (x_{ij} - \bar{x}_{mj})^2$, $\bar{x}_{mj} = \frac{\sum_{i \in C_m} x_{ij}}{n_m}$, $\bar{x}_{\cdot j} = \frac{\sum_{i=1}^{n} x_{ij}}{n}$, where $\eta_m := (1/n_m + 1/n)^{-1/2}$ and s_0 is chosen as a global constant to control the variance term. In practice, Tibshirani et al. [2002] suggest setting s_0 equal to the median of the s_j over all genes, i.e., $s_0 := \text{median}\{s_1, \ldots, s_d\}$. Tibshirani et al. [2002] recommend using a soft thresholding function to balance the estimation bias and the model complexity: $\widehat{d}_{mj} = \text{sign}(d_{mj})(|d_{mj}| - \lambda)_+$, where for any $x \in \mathbb{R}$,

$$(x)_+ := \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

and consider j-th variable to be nonfunctional to the m-th class if $\widehat{d}_{mj} = 0$. With regard to classification, given a new sample $\boldsymbol{x}' = (x'_1, \ldots, x'_d)^T$, the discriminant score for class m

is defined as:

$$\delta_m(\boldsymbol{x}') = \sum_{j=1}^{d} \frac{(x'_j - \widehat{x}_{mj})^2}{s_j^2} - 2\log(\frac{n_m}{n}),$$

with $\widehat{x}_{mj} = \bar{x}_{\cdot j} + (\eta_m(s_j + s_0))\widehat{d}_{mj}$ and $\widehat{d}_{mj}$ defined in Equation (4.2.1). λ is accordingly chosen by 10-fold cross validation procedure.

It is further shown in Wang and Zhu [2007] and Hastie et al. [2009] that the Nearest Shrunken Centroids can be explained as a solution to an optimization problem, provided that the data has a certain type of structure. In detail, suppose that $x_{ij} \sim N(\mu_j + \mu_{mj}, \sigma_j^2)$ for $i \in C_m$ with $\sum_{m=1}^{M} \mu_{mj} = 0$ to make the model identifiable, then

$$\begin{aligned}(\bar{x}_{\cdot j}, \widehat{d}_{mj}) = \underset{\mu_j, \mu_{mj}}{\arg\min}\, & \frac{1}{2} \sum_{j=1}^{d} \sum_{m=1}^{M} \sum_{i \in C_m} \frac{(x_{ij} - \mu_j - \mu_{mj})^2}{s_j^2} \\ & + \lambda \sum_{m=1}^{M} \sqrt{n_m} \sum_{j=1}^{d} \frac{|\mu_{mj}|}{s_j}, \end{aligned} \tag{4.2.2}$$

where $\widehat{d}_{mj}$ is shown in Equation (4.2.1) with $s_0 = 0$ and $\eta_m = \sqrt{1/n_m}$.

4.2.2 Normal Score Transformation

More recently, the Gaussian assumption commonly adopted by almost all high dimensional discriminant analysis methods is weakened by Liu et al. [2009]. They generalize the Gaussian distribution family to a strictly larger Nonparanormal (Gaussian Copula) family. A random variable $X := (X_1, ..., X_d)^T \in \mathbb{R}^d$ belongs to a nonparanormal family if and only if there exist a set of univariate monotone functions $\{f_j\}_{j=1}^{d}$ such that

$(f_1(X_1), ..., f_d(X_d))^T$ is multivariate Gaussian. Liu et al. [2009] utilize the normal score transformation to infer the variable structure, which is proved to be semiparametric efficiency in low dimensional settings by Klaassen and Wellner [1997]. Moreover, they analyze its theoretical performance in high dimensional settings [Liu et al., 2012]. We refer to their papers for further discussions.

In detail, given n data points $x_1, \ldots, x_n \in \mathbb{R}$, we define

$$\widetilde{F}(t; x_1, \ldots, x_n) := \frac{1}{n+1} \sum_{i=1}^{n} I(x_i \leq t), \tag{4.2.3}$$

to be the skewed empirical cumulative distribution function. Let $\Phi^{-1}(\cdot)$ be the quantile function of standard Gaussian, we define $\widetilde{f}(t) = \Phi^{-1}\big(\widetilde{F}(t; x_1, \ldots, x_n)\big)$. The normal score transformed data points $\{z_1, \ldots, z_n\}$ are then defined to be:

$$z_i := \widehat{\mu} + \widehat{\sigma} \cdot \widetilde{f}(x_i), \quad i = 1, \ldots, n,$$

where $\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\widehat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widehat{\mu})^2}$, are the sample mean and standard deviation.

4.3 Method

We begin by establishing some notations. Let $\boldsymbol{M} = [M_{jk}] \in \mathbb{R}^{d \times d}$ and $v = (v_1, ..., v_d)^T \in \mathbb{R}^d$. Let v's subvector with entries indexed by I be denoted by v_I, $\boldsymbol{M}$'s submatrix with

rows indexed by I and columns indexed by J be denoted by M_{IJ}, $\boldsymbol{M}$'s submatrix with all rows and columns indexed by J is denoted by $\boldsymbol{M}_{\cdot J}$. We define $\|v\|_2 = (\sum_{i=1}^{d} |v_i|^2)^{1/2}$ and $\|v\|_\infty = \max_{1\le i\le d} |v_i|$. We define the matrix $\ell_{\max}$ norm as the elementwise maximum value: $\|\boldsymbol{M}\|_{\max} = \max\{|M_{ij}|\}$ and the ℓ_∞ norm as $\|\boldsymbol{M}\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |M_{ij}|$. $\Lambda_{\min}(\boldsymbol{M})$ and $\Lambda_{\max}(\boldsymbol{M})$ are the smallest and largest eigenvalues of M. We further define the matrix operator norm as $\|M\| = \lambda_{\max}(\boldsymbol{M})$.

4.3.1 Model

Let $\boldsymbol{X} = [x_{ij}]$ be the dataset we are interested in, with $i = 1, \ldots, n$ and $j = 1, \ldots, d$ representing the n samples and d variables. We assume that there are d variables belonging to K groups, and collect the set of indices of the d_k variables in the k-th group in the set G_k, $k = 1, \ldots, K$. Similarly, we assume that there are n samples belonging to M classes, and that C_m is equal to the set of indices of the n_m samples in the m-th class, $m = 1, \ldots, M$. For simplicity, we rearrange the variables such that $\boldsymbol{X}_{i\cdot} = (\boldsymbol{X}_{iG_1}^T, \ldots, \boldsymbol{X}_{iG_K}^T)^T$.

We consider the data matrix $\boldsymbol{X}$ and denote by: $\boldsymbol{x}_{ik} = (x_{ij}, j \in G_k)^T \in \mathbb{R}^{d_k}$, $\boldsymbol{x}_{ik}^* = (x_{ij} - \bar{x}_{\cdot j}, j \in G_k)^T$. We suppose that for $k = 1, \ldots, K$,

$$\boldsymbol{x}_{ik} \sim^{i.i.d} N(\boldsymbol{\mu}_k + \boldsymbol{\mu}_{mk}, \Sigma_k), \quad \forall\, i \in C_m, \tag{4.3.1}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}_{mk}$ are both unknown vectors with $\sum_{m=1}^{M} \boldsymbol{\mu}_{mk} = \mathbf{0}$, to make the model identifiable.

4.3.2 Group Nearest Shrunken Centroids

Let $\widetilde{\Sigma}_k$ be an arbitrary estimator of Σ_k. We propose a loss function with a similar version as the NSC's shown in Equation (4.2.2), but with a group penalty:

$$\begin{aligned} L := \tfrac{1}{2}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i\in C_m}\|\boldsymbol{x}_{ik}-\boldsymbol{\mu}_k-\boldsymbol{\mu}_{mk}\|_k^2 \\ +\lambda\sum_{k=1}^{K}\sum_{m=1}^{M}(n_m\omega_{mk})\|\boldsymbol{\mu}_{mk}\|_k, \end{aligned} \tag{4.3.2}$$

where for any $v\in\mathbb{R}^{d_k}$, $\|v\|_k$ is defined as:

$$\|\boldsymbol{v}\|_k := (\boldsymbol{v}^T\widetilde{\Sigma}_k^{-1}\boldsymbol{v})^{1/2}. \tag{4.3.3}$$

The following theorem, whose proof we defer to Appendix 4.4.2, provides the closed form of the minimizers to Equation (4.3.2):

THEOREM 2. *We denote by* $\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K$ *and* $\{\widetilde{\boldsymbol{\mu}}_{mk}\}_{m=1,k=1}^{m=M,k=K}$ *the optima to Equation* (4.3.2)*:*

$$\{\{\widetilde{\boldsymbol{\mu}}_k\}_{k=1}^K,\{\widetilde{\boldsymbol{\mu}}_{mk}\}_{m=1,k=1}^{m=M,k=K}\} := \arg\min L. \tag{4.3.4}$$

Then for all $k\in\{1,\ldots,K\}$ *and* $m\in\{1,\ldots,M\}$*,*

$$\widetilde{\boldsymbol{\mu}}_k=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_{ik}\ \text{ and }\ \widetilde{\boldsymbol{\mu}}_{mk}=(1-\frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+\widehat{\boldsymbol{\mu}}_{mk}, \tag{4.3.5}$$

where $\widehat{\boldsymbol{\mu}}_{mk} := \frac{1}{n_m}\sum_{i\in C_m}\boldsymbol{x}_{ik}^*$.

Remark 4.3.1. *In practice, while defining* $\|\cdot\|_k$ *in Equation* (4.3.3)*, we adopt a similar idea as Tibshirani et al. [2002] and choose*

$$\widetilde{\Sigma}_k = \widehat{\Sigma}_k + s_0^2 I_{d_k \times d_k}, \tag{4.3.6}$$

where $\widehat{\Sigma}_k$ *is the sample covariance matrix of the* k*-th group of variables using the whole samples,* $I_{d_k \times d_k}$ *is the* $d_k \times d_k$ *identity matrix and*

$$s_0^2 = \text{median}(((\text{diag}(\widehat{\Sigma}_1))^T, \ldots, (\text{diag}(\widehat{\Sigma}_K))^T)^T),$$

is the median of all marginal sample variances.

Given a new data point $x \in \mathbb{R}^d$, the discriminant score for class m is defined as:

$$\delta_m(x) = \sum_{k=1}^{K} \|x_{G_k} - \widetilde{\boldsymbol{\mu}}_k - \widetilde{\boldsymbol{\mu}}_{mk}\|_k^2 - 2\log(\frac{n_m}{n}). \tag{4.3.7}$$

4.3.3 Theoretical Properties of gNSC

For simplicity, we analyze the theoretical performance of a slightly simpler version of the model proposed in Section 4.3.1, where for any $k \in \{1, \ldots, K\}$, $\boldsymbol{\mu}_k := \mathbf{0}$. In this way, the estimators in Equation (4.3.5) can be reduced to:

$$\widetilde{\boldsymbol{\mu}}_k = \mathbf{0} \quad \text{and} \quad \widetilde{\boldsymbol{\mu}}_{mk} = (1 - \frac{\lambda \omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+ \widehat{\boldsymbol{\mu}}_{mk}, \tag{4.3.8}$$

where $\widehat{\boldsymbol{\mu}}_{mk} = \frac{1}{n_m}\sum_{i\in C_m}\boldsymbol{x}_{ik}$ and $\|v\|_k := v^T\widehat{\Sigma}_k^{-1}v$. Remind that $\widehat{\Sigma}_k$ is the sample covariance matrix of the k-th group of variables using the whole samples. Furthermore, to achieve a better theoretical performance, we define the sparse set of $\{\boldsymbol{\mu}_{m1},\ldots,\boldsymbol{\mu}_{mK}\}$ with respect to the sample class C_m to be $S_m := \{k\in\{1,\ldots,K\} : \boldsymbol{\mu}_{mk}\neq\mathbf{0}\}$, and the corresponding estimated sparse set with respect to the m-th sample class to be $\widehat{S}_m := \{k\in\{1,\ldots,K\} : \widetilde{\boldsymbol{\mu}}_{mk}\neq\mathbf{0}\}$.

Estimation Consistency.

To achieve the estimation consistency result, we need the following three "boundedness" assumptions:

(A1) There exist two finite constants $c_1, c_2 \in (0,\infty)$ such that

$$c_1 < \min_{1\le m\le M}\min_{k\in S_m}\left\{(\boldsymbol{\mu}_{mk}^T\Sigma_k^{-1}\boldsymbol{\mu}_{mk})^{\frac{1}{2}}, \|\boldsymbol{\mu}_{mk}\|\right\} \le \max_{1\le m\le M}\max_{k\in S_m}\left\{(\boldsymbol{\mu}_{mk}^T\Sigma_k^{-1}\boldsymbol{\mu}_{mk})^{\frac{1}{2}}, \|\boldsymbol{\mu}_{mk}\|\right\} < c_2;$$

(A2) There exists $0 < c_3 = \min\{\Lambda_{\min}(\Sigma_k^{-1}), k = 1, ..., K\} < \infty$.

(A3) $\omega_{mk} \propto \sqrt{d_k}$ is upper bounded by $\omega_0 = O((\min_{1\le m\le M} n_m)^{\gamma_0/2})$ for some $0 \le \gamma_0 < 1$.

THEOREM 3. *(Estimation Consistency) Under assumption (A1)-(A3), $\forall m \in \{1,\ldots,M\}$, for large enough n_m, if we further suppose that $\lambda \to 0$, $n_m^{1/2}\lambda \to \infty$, and $\epsilon = \gamma\lambda^2$ with $\gamma > \omega_0^2c_2^2/c_1^2$, we have $\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon) = O(\exp(-C_k n_m^2\lambda^4)) \to 0$, where $C_k = \min(\frac{3\gamma^2c_3^2}{16d_k}, \frac{3\omega_k^4}{64d_k})$.*

Sparsity Recovery and Misclassification Consistency.

For sparsity recovery and misclassification consistency, we only consider the situation when there are only two groups of samples, indexed by C_1 and C_2. We denote by y_i the label of i-th sample: $y_i = 0$ if $i \in C_1$ and $y_i = 1$ if $i \in C_2$. Suppose that $(\boldsymbol{X}_{i\cdot}, y_i)$ are i.i.d drawn from the joint distribution of (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$. The target of the classification is to determine the value of Y given a new data point $\boldsymbol{x} \in \mathbb{R}^d$. Here we further suppose that

$$(X|Y = 0) \sim N(\boldsymbol{\mu}_1, \Sigma) \quad \text{and} \quad (X|Y = 1) \sim N(\boldsymbol{\mu}_2, \Sigma),$$

where $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1d})^T = (\boldsymbol{\mu}_{11}^T, \ldots, \boldsymbol{\mu}_{1K}^T)^T$, $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2d})^T = (\boldsymbol{\mu}_{21}^T, \ldots, \boldsymbol{\mu}_{2K}^T)^T$, and $\Sigma_{G_k G_k} = \Sigma_k$, $\forall k \in \{1, \ldots, K\}$. Define the prior probabilities $\pi_1 = \mathbb{P}(Y = 0)$, $\pi_2 = \mathbb{P}(Y = 1)$, and we assume that $\pi_1 = \pi_2$. It is easy to extend it to the case where $\pi_1 \neq \pi_2$ [Hastie et al., 2009]. In this way, the Bayes rule is given by:

$$g^*(x) = \begin{cases} 1, & \text{if } \langle \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1),\ x - \boldsymbol{\mu}_a \rangle > 0 \\ 0, & \text{otherwise} \end{cases} \tag{4.3.9}$$

where $\langle a, b \rangle$ is the inner product of a and b and $\boldsymbol{\mu}_a := (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Define

$$\begin{aligned} \mathcal{S} &:= \{k : \|\boldsymbol{\mu}_{1k} - \boldsymbol{\mu}_{2k}\|_2 > 0\}, \\ \widehat{\mathcal{S}} &:= \{k : \|\widetilde{\boldsymbol{\mu}}_{1k} - \widetilde{\boldsymbol{\mu}}_{2k}\|_2 > 0\}, \end{aligned} \tag{4.3.10}$$

where $\widetilde{\boldsymbol{\mu}}_1 = (\widetilde{\boldsymbol{\mu}}_{11}^T, \ldots, \widetilde{\boldsymbol{\mu}}_{1K}^T)^T$ and $\widetilde{\boldsymbol{\mu}}_2 = (\widetilde{\boldsymbol{\mu}}_{21}^T, \ldots, \widetilde{\boldsymbol{\mu}}_{2K}^T)^T$. Remind that $\widetilde{\boldsymbol{\mu}}_{mk}$ is defined in Equation (4.3.8). Define $s = \mathrm{card}(\mathcal{S})$. It is easy to see that

$$\mathcal{S} \subset S_1 \cup S_2 \quad \text{and} \quad \widehat{\mathcal{S}} = \widehat{S}_1 \cup \widehat{S}_2, \ a.s.. \tag{4.3.11}$$

Given $\pi_1 = \pi_2$, the gNSC classification procedure g proposed in Equation (4.3.7):

$$g(x) = \begin{cases} 1, & \text{if} \quad \delta_2(x) - \delta_1(x) > 0, \\ \\ 0, & \text{otherwise}, \end{cases} \tag{4.3.12}$$

can be further written as:

$$g(x) = \begin{cases} 1, & \text{if} \left\langle \widehat{\Sigma}^{-1}(\widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1), \ x - \widetilde{\boldsymbol{\mu}}_a \right\rangle > 0, \\ \\ 0, & \text{otherwise}, \end{cases} \tag{4.3.13}$$

where $\widetilde{\boldsymbol{\mu}}_a = \frac{\widetilde{\boldsymbol{\mu}}_1 + \widetilde{\boldsymbol{\mu}}_2}{2}$, $\widehat{\Sigma}$ and $\widehat{\Sigma}^{-1}$ are defined to be:

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \widehat{\Sigma}_2 & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \cdot & \cdot & \ldots & \cdot \\ \mathbf{0} & \mathbf{0} & \ldots & \widehat{\Sigma}_K \end{pmatrix}. \tag{4.3.14}$$

The corresponding misclassification errors are

$$\mathcal{C}(g^*) = \bar{\Phi}(\Psi_\Sigma(\Delta, \xi)) \quad \text{and} \quad \mathcal{C}(g) = \bar{\Phi}(\Psi_\Sigma(\widetilde{\Delta}, \widehat{\xi})),$$

where $\bar{\Phi}$ is the survival probability of the standard Gaussian distribution, i.e. $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$, and $\Delta := \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, $\widetilde{\Delta} := \widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1$, $\xi := \Sigma^{-1}\Delta$, $\widehat{\xi} := \widehat{\Sigma}^{-1}\widetilde{\Delta}$ $\Psi_\Sigma(\boldsymbol{a}, \boldsymbol{b}) := \frac{\boldsymbol{a}^T\boldsymbol{b}}{2\sqrt{\boldsymbol{b}^T\Sigma\boldsymbol{b}}}$.

To obtain a fast rate on the misclassification consistency, we need the following assumption:

(A4) For any $k \in S_1 \cap S_2$, we have $k \in \mathcal{S}$. In other words, $S_1 \cup S_2 = \mathcal{S}$.

The next theorem states that the sparsity pattern can be recovered consistently:

THEOREM 4. *(Sparsity Recovery) Under assumptions (A1)-(A4), if we further suppose that $\epsilon = \gamma\lambda^2$, $\gamma > c_2^2\omega_0^2/c_1^2$, $\lambda \to 0$ and $\lambda n^{1/2} \to \infty$. Then if $K = o(e^{Cn^2\lambda^4})$, $C > 0$ is a sufficient small constant, we have:*

$$\mathbb{P}(\widehat{S} \neq S) \to 0.$$

Finally, we define $\mathcal{M} := \{j \in \{1, \ldots, d\} : j \in G_k \text{ for some } k \in \mathcal{S}\}$.

THEOREM 5. *(Misclassification Consistency) Under assumptions (A1)-(A4), if we define $\|(\Sigma_{\mathcal{M}\mathcal{M}})^{-1} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\| := O(a_{n,d})$, where $a_{n,d}$ scales with (n, d) and further suppose*

$\lambda = O(n^{-\frac{1}{2}}[\log(n)\log(K)]^{\frac{1}{4}})$, *then we have*

$$\mathcal{C}(g) - \mathcal{C}(g^*) = O_P(c_s s^2 \omega_0^4 \cdot \frac{\log s \log \omega_0}{n}) + c_s O(a_{n,d}^2),$$

where $c_s := \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_2^2$.

As alluded to previously, we defer the proofs of the above theoretical results to Appendix 4.4.2.

4.3.4 Nonparanormal and Normal Score Transformation

To the extent that both gNSC and NSC are only well-justified when data are Gaussian, their applications to the "real" data, e.g. gene expression data, is highly limited. To attack this problem, Liu et al. [2009] weaken the Gaussian assumption via introducing the Nonparanormal distribution family. In detail, a random variable $X = (X_1, ..., X_d)^T$ is said to follow a *nonparanormal* distribution if and only if there exist a set of univariate monotone transformations $f = \{f_j\}_{j=1}^d$ such that: $f(X) = (f_1(X_1), ..., f_d(X_d))^T := Z \sim N(\mu, \Sigma)$, where $\mu = (\mu_1, ..., \mu_d)^T$, $\Sigma = [\Sigma_{jk}]$ are the mean and covariance matrix of the Gaussian distribution Z. $\{\sigma_j^2 := \Sigma_{jj}\}_{j=1}^d$ are the corresponding marginal variances. To make the model identifiable, we assume, for $1 \leq j \leq d$, $\mathbb{E}(X_j) = \mathbb{E}(f_j(X_j)) = \mu_j$ and $\mathrm{Var}(X_j) = \mathrm{Var}(f_j(X_j)) = \sigma_j^2$. For notational convenience, we denote such X by $X \sim NPN(\mu, \Sigma, f)$. Liu et al. [2009] prove that if the transformation functions are monotone, the nonparanormal family is equivalent to the Gaussian Copula.

In practice, a parallel model to Equation (4.3.1) can be constructed: $\boldsymbol{x}_{ik} \sim^{i.i.d.} \text{NPN}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_{mk}, \Sigma_k, \boldsymbol{f}_k), \forall i \in C_m,\ k \in \{1, \ldots, K\}$, where $\boldsymbol{f}_k = \{f_k^j\}_{j=1}^{d_k}$ is a set of univariate monotone functions common across difference classes. Under this model, a natural data preprocessing approach is to do normal score transformation first on the data and achieve $\boldsymbol{Z} = [z_{ij}] \in \mathbb{R}^{n \times d}$ such that for all $m \in \{1, \ldots, M\}$: $z_{ij} = \widehat{\mu}_{mj} + \widehat{\sigma}_j \cdot \Phi^{-1}(\widetilde{F}(x_{ij}; \{x_{i'j}\}_{i' \in C_m}))$, $\forall i \in C_m$, where $\widetilde{F}(\cdot;\cdot)$ is defined in Equation (4.2.3), $\widehat{\mu}_{mj} = \frac{1}{n_m} \sum_{i \in C_m} x_{ij}$ and $\widehat{\sigma}_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \frac{1}{n} \sum x_{ij})^2$. Its theoretical performance has been deeply studied by Klaassen and Wellner [1997] and Bickel [1998]. Its theoretical and empirical performance in high dimensional settings has been further verified by Liu et al. [2012]. We therefore recommend conducting normal score transformation while preprocessing the data and use $\boldsymbol{Z}$ as the input to the gNSC algorithm. With regard to classification, given a new data point $\boldsymbol{x} \in \mathbb{R}^d$, we transform it to a new data $\boldsymbol{z} = (z_1, \ldots, z_d)^T$ by:

$$z_j = \frac{1}{M} \sum_{i=1}^{M} (\widehat{\mu}_{mj} + \widehat{\sigma}_j \cdot \Phi^{-1}(\widetilde{F}(x_j; \{x_{i'j}\}_{i' \in C_m}))).$$

We then apply $\boldsymbol{z}$ to Equation (4.3.7) to obtain the discriminant score for classification.

4.4 Application

Gene expression is the process by which information encoded *within* a gene is used in the synthesis of a functional gene *product*, such as proteins. After genome sequencing, microarray analysis has become one of the indispensable tools that many biologists use to

monitor genome-wide expression levels of genes in a given organism. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. To the extent that the quantity of data collected from such experiments is overwhelming, efficiently synthesizing these expression levels, via microarray analysis, is an essential part of current methodology. The GPL96 set (Affymetrix GeneChip Human Genome U133 Array Set HG-U133A) [McCall et al., 2010], a collection of publicly available microarray data from hundreds of different experiments, is among the highest accessible microarray datasets. This set includes over 1,000,000 unique oligonucleotide features covering more than 39,000 transcript variants, which in turn represent greater than 33,000 of the best characterized human genes. Sequences were selected from GenBank, dbEST, and RefSeq. Sequence clusters were created from Build 133 of UniGene and refined by analysis and comparison with a number of other publicly available databases including the Washington University EST trace repository and the University of California, Santa Cruz golden-path human genome database.

We will apply our above-discussed technique (cf. Section 4.3), to 20,248 genes and 8,124 microarray samples from Affymetrixs HG-U133A platform. Each sample belong to a certain tissue type (e.g., lung cancers, brain tumor etc.), of which we have 312 types total. We also are interested in certain gene pathways extracted from the one of the largest pathway databases, Molecular Signature Database (MSigDB) [Subramanian et al., 2005]. This database consists of 12,713 genes, notably including information concerning biological pathways and responses to a drug treatment. The pathway information is extracted

from the MSigDB. A total of 6,769 pathways are obtained. The main purpose of our experiment is to test the association between gene pathways and certain diseases or tissue types. To demonstrate the effectiveness of our new approach for high dimensional discriminant analysis, we look at the performance of both classification and feature selection of group Nearest Shrunken Centroids. The task of sample classification is to classify and predict the diagnostic category of a sample on the basis of its gene expression profile. We show our effectiveness of sample classification by comparing the performance of our new method with Nearest Shrunken Centroids. The misclassification rates of both of the two methods are calculated in Section 4.4.1. The performance of gNSC on feature selection is shown in Section 4.4.2, where we apply gNSC on context analysis. We show that the power of feature selection is improved by utilizing the text information.

4.4.1 gNSC for Classifcation

The raw data of GPL96 contain 20,248 genes and 8,124 samples belonging to 312 tissue types. Since the tissue types with too few samples may not follow the asymptotic properties, we exclude from consideration tissue types with fewer than 30 samples. 5,510 samples, belonging to 24 tissue types, form our data set. To explore the association between the gene pathways and the tissues, we utilize the gene structure information extracted from the Molecular Signature Database (MSigDB), which contains 6,769 pathways and 12,713 genes. To preserve efficiency, we exercise data-rich gene pathways – those with more than 50 genes. We finally have 4,005 pathways, containing 10,990 different genes. Conse-

quently, the final dataset we use contains 10,990 genes belonging to 4,005 pathways, 5,510 samples belonging to 24 tissue types. Finally, we arrange the genes by pathways, giving us a matrix with dimension $5,510 \times 88,396$. gNSC is then applied to this data for detecting the association between gene pathways and tissue types. Note that there are 88,396 columns instead of 10,990. This is because the genes can belong to more than one pathway.

Procedures.

In Section 4.3.3, we have shown that the asymptotic variable selection and misclassification consistency results of gNSC hold under the assumption of normality of the data. Therefore we first test the normality of the dataset. For each gene in each sample class we present the Quantile-to-Quantile plot (QQ plot) to visualize the normality. Three of them are shown in Appendix 4.4.6. It can be observed that all the three marginal distributions are severely away from Gaussian. Accordingly, we utilize the idea of normal score transformation (NST) to generalize the model to nonparanormal [Liu et al., 2012].

We calculate $\widetilde{\boldsymbol{\mu}}_k$ and $\widetilde{\boldsymbol{\mu}}_{mk}$ by using the Equation (4.3.5), where $\omega_{mk} = \sqrt{d_k/n_m}$, $\widetilde{\Sigma}_k$ is calculated using Equation (4.3.6), and λ is a tuning parameter. We use two types of cross validation methods to tune the parameter λ. 10-fold cross-validation is used by [Tibshirani et al., 2002] to find the λ with the lowest average misclassification error. In practice, however, the new data points usually come from a new experiment. We therefore propose an alternative way, "leave experiment out" cross-validation, to select λ. In detail, we isolate all samples from a single experiment as our "testing data;" remaining data is used as "training

data." We calculate the discriminant scores of all classes for each data point in the testing data using Equation (4.3.7). The estimated class for each data point is the one that achieves the lowest discriminant score. Then, for each λ, we calculate an average misclassification error by summing up the number of misclassified data points for all experiments and dividing it by the total number of samples. The parameter λ is chosen to be the one with the lowest average misclassification error.

Results.

We say that the pathway k is significantly associated with the sample class m if $\widetilde{\mu}_{mk}$ is greater than 0. We call the combination of one certain pathway and one certain tissue type a block. There are then $M \times K$ blocks.

We compare our method of group Nearest Shrunken Centroids with the Nearest Shrunken Centroids. The averaged misclassification error for each λ from 0.1 to 10 is calculated using both "leave fold out" and "leave experiment out" cross-validation, where "leave fold out" represents the commonly used 10-fold cross validation. The λ with the lowest averaged misclassification error is picked up using these two cross validation methods. The corresponding averaged misclassification errors with their standard deviations are calculated. Moreover, we present the corresponding averaged significant numbers of unique genes across different tissue types with their standard deviations. All the results are illustrated in Table 4.1. It can be observed that group Nearest Shrunken Centroids has – on average – lower misclassification errors than Nearest Shrunken Centroids, as the gNSC re-

quires much fewer genes to obtain a better prediction result. We also provide two tables to show the general trend of the averaged misclassification errors and the corresponding genes with increasing of λ in Appendix 4.4.6.

Table 4.1: Leave Fold Out v.s. Leave Experiment Out Cross Validation.

		gNSC	sd	NSC	sd
leave experiment out	gene number	3901	22.50	4495	17.26
	error	0.089	0.0505	0.149	0.0479
leave fold out	gene number	5502	21.33	5670	11.14
	error	0.139	0.0866	0.136	0.0906

By using "leave experiment out" cross validation, the tuning parameter with lowest averaged misclassification error is around $\lambda = 6.5$. To illustrate our result more clearly, we randomly pick 12 tissue types and 50 gene pathways for visualization. Figure 4.1 presents the significant associations, i.e. the threshold term $(1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+$, between gene pathways and tissue types: red color suggests that the corresponding pathway and tissue type are estimated to be associated. The heatplots of the negative "shrinkage amount," i.e. $(1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})$ and the biological expression levels are shown in Appendix 4.4.6.

Using the parameters extracted from our data, we can reclassify the samples associated with different tissue types and compare them with the true labels. The result is shown in Figure 4.2, with the y-axis as true labels and x-axis as predictive labels. The integer with the coordinates (A,B) represents the number of the tissues that is truly B and is predicted to be A. For example, the value 1448 in the left top corner means that there are 1448 samples that are truly acute lymphoma blood and are successfully predicted to be so. We succeed in accurately predicting over 80% of the samples. Moreover, our errors are largely due to

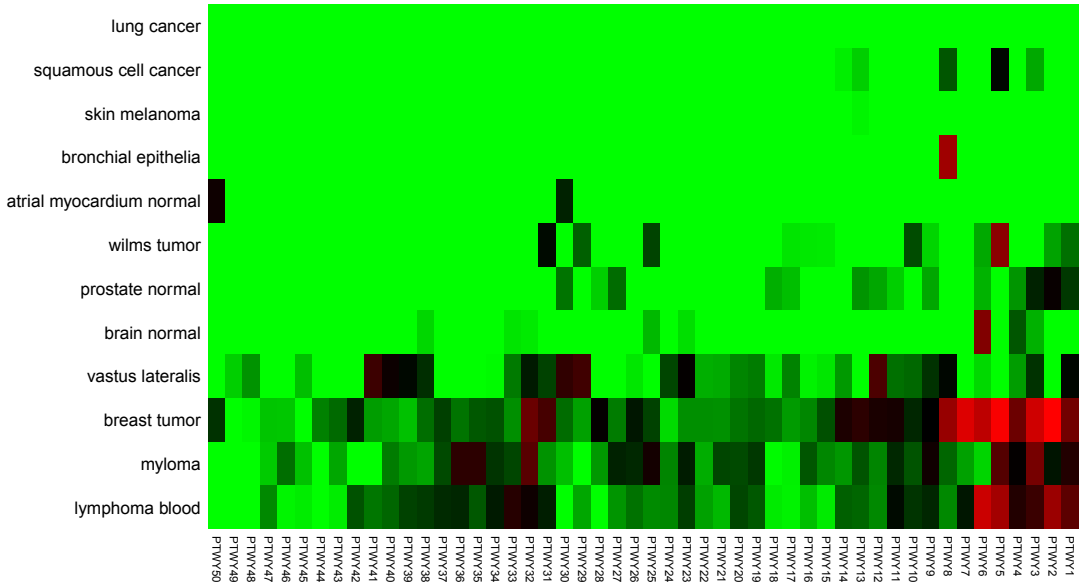


Figure 4.1: Significant Association Between Pathways and Tissue Types

similarity in tissue types, which are also hard to differentiate biologically. For example, we fail to differentiate between wilms tumor that is relapse and the wilms tumor that is non-relapse.

We find 3,220 significant relations in all and 174 significant relations based on the tissue types and pathways we present in Figure 4.5(a). Note that each relation involves one tissue type and one pathway which includes a number of genes. One gene can be involved in several significant relations and one relation involves all the genes in the corresponding pathway. Even for the subset with only 174 significant relations, many have been found to be biologically meaningful. Part of the results are showed in Table 4.2. For example,

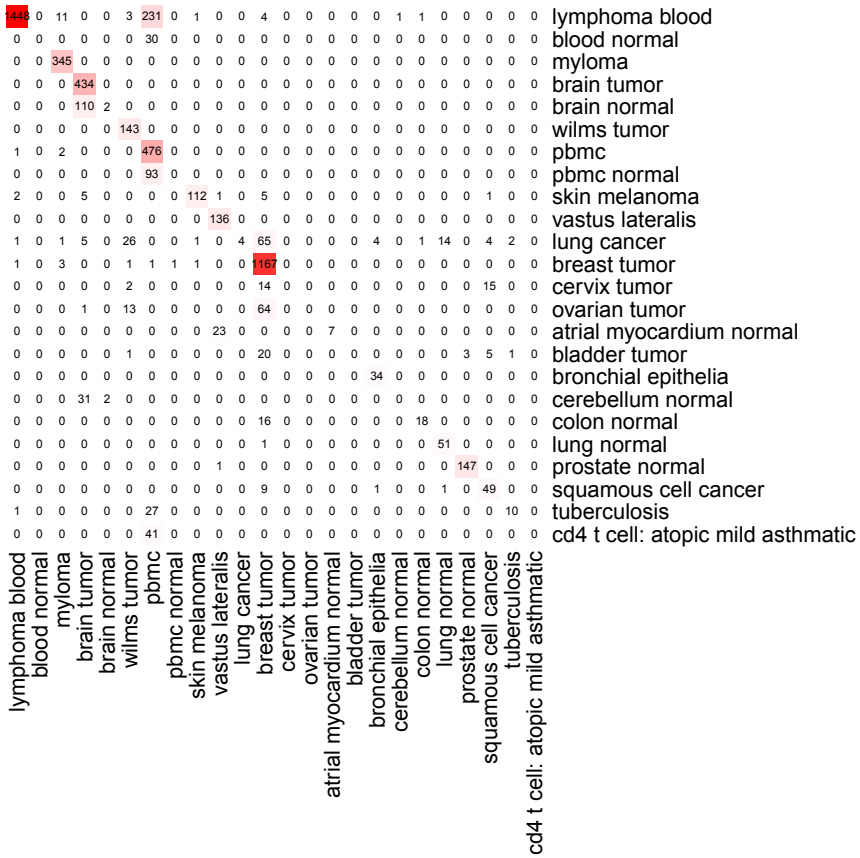


Figure 4.2: The True Tissue Types v.s. the Predictive Tissue Types.

Table 4.2: True relations learnt from the GPL96 data.

(Pathway Name) Disease Related to the Class
(GAUSSMANN MLL AF4 FUSION TARGETS B DN) leukemia
(OZANNE AP1 TARGETS UP) breast tumor
(GOLUB ALL VS AML DN) blasts and mononuclear cells: leukemia
(CHESLER BRAIN D6MIT150 QTL CIS) barin: glioblastoma
(PODAR RESPONSE TO ADAPHOSTIN DN) breast tumor
(HAHTOLA MYCOSIS FUNGOIDES UP) b cell: lymphoma
(HEDVAT ELF4 TARGETS UP) blasts and mononuclear cells: leukemia
(STEIN ESTROGEN RESPONSE NOT VIA ESRRA) breast tumor
(MACLACHLAN BRCA1 TARGETS DN) breast tumor
(VETTER TARGETS OF PRKCA AND ETS1 DN) breast tumor
(NIKOLSKY BREAST CANCER 22Q13 AMPLICON) breast tumor
(DONATO CELL CYCLE TRETINOIN) breast tumor
(CAFFAREL RESPONSE TO THC 8HR 3 DN) b cell: lymphoma
(SINGH NFE2L2 TARGETS) breast tumor
(WANG RESPONSE TO ANDROGEN UP) prostate tumor
(WAGNER APO2 SENSITIVITY) breast tumor
(DE YY1 TARGETS UP) breast tumor
(SABATES COLORECTAL ADENOMA SIZE UP) breast tumor
(MYLLYKANGAS AMPLIFICATION HOT SPOT 11) breast tumor

(The information used here are from http://www.broadinstitute.org/.)

"CHESLER BRAIN D6MIT150 QTL CIS" is a Cis-regulatory quantitative trait loci found at the D6Mit150 region. It is believed to regulate the central nervous system. Therefore, this pathway is considered to be highly related to certain brain diseases [Chesler et al., 2005] and is successfully identified by our techniques.

4.4.2 Context Analysis of Myc pathway

Myc is a regulator gene that codes for a transcription factor. A mutated version of Myc is found in many cancers. Translocation involving Myc is critical to certain kinds of B-cell lymphoma [Lovec et al., 1994]. A very recent result obtained by Ji et al. [2011] concludes that microarray samples enriched in Wilms tumor have low Myc. A list of 51 genes are believed to be highly positively correlated with Myc in Myc pathways, 37 of which are included in GPL96. To show the effectiveness of feature selection, we use context analysis to identify most related genes of medical terms including "wilms tumor" and "b-cell lymphoma". Since it is not the main part of this paper, we put the detailed procedure of conducting context analysis in the Appendix 4.4.6. We will discuss more about it in our future papers.

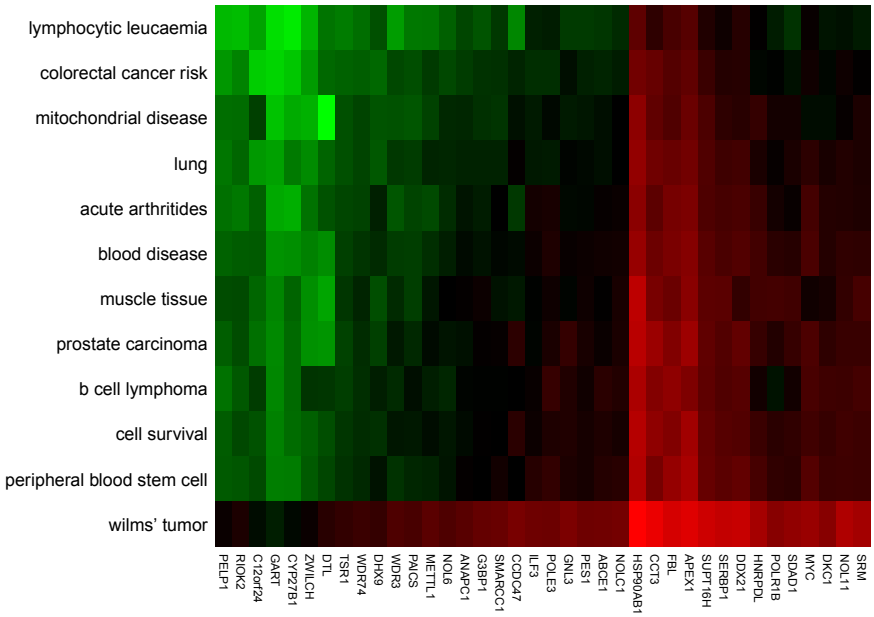
The 37 positively related genes in Myc pathways are believed to be related to both "Wilms tumor" and "b-cell lymphoma". Both NSC and gNSC are used in our analysis. By including the text information, we show that both "Wilms tumor" and "b-cell lymphoma" are predicted to be significantly related to the Myc pathway which coincides with the finding of Ji et al. [2011]. For the sake of comparison, we used both NSC and gNSC

without including the text information. To the extent that the results of these "control" experiments were insignificant, we conjecture that text information is necessary to the discernative power of context analysis in feature selection. We show the results of context analysis using NSC and gNSC separately below.

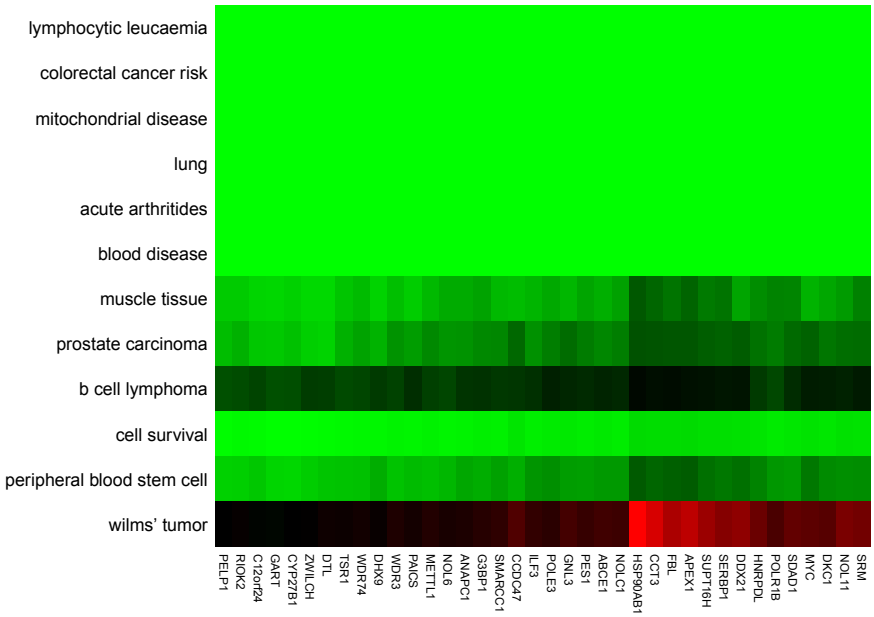
Context Analysis using NSC

Similar to Section 4.4.1, we consider only tissue types with more than 30 samples. After further screening out the samples without document information, there remain 5,484 samples for study, leading to an expression matrix with a dimension of $5,484 \times 12,713$. We extract 11,220 meaningful single terms from the text information of GPL96, from which 4,308 terms are included in the 5,484 sample documents we have in the gene expression matrix. The dictionary we use consists of 1,048,576 words and phrases in total. Among them we only need nouns, resulting in 565,308 words and phrases left. Each word and phrase has been indexed to a specific synonym cluster. We exclude those with no terms belonging to any sample document. There are 4,560 synonym clusters left with different indices. In summary, we end up with an index-doc matrix with a dimension of $4,560 \times 5,484$ (see Appendix 4.4.6 for more details). Based on the expression matrix and the index-doc matrix, we can construct the index-gene relevance matrix, which has a dimension of $5,484 \times 4,560 \times 12,713$. We then implement NSC to analyze the associations between synonym clusters and the genes.

Remark 4.4.1. *Although the dimension of the index-gene relevance matrix is over* $3 \times$



(a)



(b)

Figure 4.3: gNSC Results of Keywords v.s. Myc pathway. (a) The mean relevance levels of synonym word groups with the 37 genes in Myc pathway; (b) The figure illustrates $\widetilde{\mu}_{mk}$ calculated by Equation (4.3.5).

10^{11}*, the clean encoding of the data in* R *allows for efficient analysis of this large-scale information: by calculating sufficient statistics of the original microarray data, we were able to finish our whole procedure in minutes.*

Since it is impossible to obtain all true relations between genes and words , we use a simpler algorithm to choose the amount of shrinkage instead of doing cross validation. We choose λ to be the 95% quantile of $|d_{mj}|$ with $m = 1, 2, \cdots, M$ and $j = 1, 2, \cdots, d$. Here we have $M = 4,560$ and $d = 12,713$. Therefore 5% of the index-gene relations are considered to be significant. This gives us a λ around 0.00348.

To show the effectiveness of our method, we count the number of genes in the list that are significantly related to the word "wilms tumor" and "b-cell lymphoma". All the 37 genes are significant related with "wilms tumor" and 32 of them are significant related to "b-cell lymphoma". Both words are significantly related to Myc. The heatplot of the relevance of certain words, including "wilms tumor" and "b-cell lymphoma", with the 37 genes is shown in Appendix 4.4.6.

Context Analysis using gNSC

We can also include the pathway information into the context analysis and use gNSC to identify the most related pathways of certain words. Similar to Section 4.4.1, we screen out the gene pathways with more than 50 genes and sort the matrix of expression levels by pathways. We end up with a matrix with a dimension of $5,484 \times 88,396$. As above, we can construct the index-gene relevance matrix, which has a dimension of $5,484 \times 4,560 \times$

$88, 396$. The mean relevance levels in one synonym block of the relevance matrix are defined to be the means of all relevance values restricted to the pertaining block. The result can be visualized in Figure 4.3(a). Figure 4.3(b) shows the relevance of certain words and the 37 genes in Myc pathway. The red block shows high relevance of the word and the genes. As we can see, the Myc gene pathway is significantly related to both "wilms tumor" and "b-cell lymphoma".

Appendix.1. Proofs of the Theorems

4.4.3 Proof of Theorem 2

Proof. It is not difficult to observe that

$$\begin{aligned} \arg\min_{\boldsymbol{\mu}_k, \boldsymbol{\mu}_{mk}} L = (\{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \boldsymbol{x}_{ik}, k = 1, ..., K\}, \\ \{\arg\min_{\boldsymbol{\mu}_{mk}} L_m, k = 1, ..., K\}), \end{aligned} \tag{4.4.1}$$

where

$$L_m := \frac{1}{2n_m} \sum_{k=1}^{K} \sum_{i \in C_m} \|\boldsymbol{x}_{ik}^* - \boldsymbol{\mu}_{mk}\|_k^2 + \lambda \sum_{k=1}^{K} \omega_{mk} \|\boldsymbol{\mu}_{mk}\|_k. \tag{4.4.2}$$

The solution path to Equation (4.4.2) could be achieved by using Karush-Kuhn-Tucker conditions, presented as the following lemma:

LEMMA 1. *A necessary and sufficient condition for a vector* $\boldsymbol{\mu}_m = (\boldsymbol{\mu}'_{m1}, ..., \boldsymbol{\mu}'_{mK})'$ *with*

sparsity pattern $S_m := \{k : \boldsymbol{\mu}_{mk} \neq 0\}$ *to be a solution to Equation* (4.4.2) *is:*

$$-(\boldsymbol{x}^*_{mk}/n_m - \boldsymbol{\mu}_{mk}) + \lambda \cdot \frac{\omega_{mk}\boldsymbol{\mu}_{mk}}{\|\boldsymbol{\mu}_{mk}\|_k} = 0, \qquad \forall \quad k \in S_m; \tag{4.4.3}$$

$$\| - \boldsymbol{x}^*_{mk}/n_m\|_k \leq \lambda\omega_{mk}, \qquad \forall \quad k \in S^c_m, \tag{4.4.4}$$

where $\boldsymbol{x}^*_{mk} = \sum_{i\in C_m} \boldsymbol{x}^*_{ik}$.

It is not difficult to verify that the solution to Equations (4.4.3) and (4.4.4) can be presented as:

$$\widetilde{\boldsymbol{\mu}}_{mk} = (1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+\widehat{\boldsymbol{\mu}}_{mk}. \tag{4.4.5}$$

This complete the proof. □

4.4.4 Proof of Theorem 3

Under assumption (A1) to (A3), we provide the following two Lemmas, which are necessary to prove the estimation consistency result.

LEMMA 2. *Under assumptions (A1), for all* $m \in \{1, \ldots, M\}$,

for all $k \in S_m$, *we define*

$$A_{mk} := \{c_1 < \|\widehat{\boldsymbol{\mu}}_{mk}\|_k < c_2\}.$$

$\forall\ k \in S^c_m$, *we define*

$$B_{mk} := \{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k > \lambda\omega_{mk}\}.$$

Then for large enough n_m, we have

$$\begin{aligned}\mathbb{P}(A_{mk}^c) \leq & \exp[-(\frac{-\sqrt{d_k+2v_{mk}}+\sqrt{-d_k+2n_mc_2^2}}{2})^2] \\ & + \exp[-\frac{(d_k+v_{mk}-n_mc_1^2)^2}{4(d_k+2v_{mk})}]\end{aligned}$$

and

$$\mathbb{P}(B_{mk}) \leq \exp[-\frac{3}{16}d_k \cdot (\frac{n_m\lambda^2\omega_{mk}^2}{d_k}-1)^2] \tag{4.4.6}$$

where $v_{mk} = n_m\boldsymbol{\mu}_{mk}^T\Sigma_k^{-1}\boldsymbol{\mu}_{mk}$.

LEMMA 3. *Under assumptions (A1) and (A2), for all $m \in \{1,\ldots,M\}$, for large enough n_m, and $\forall\ k \in S_m$, $\forall\ \epsilon > 0$, we have*

$$\begin{aligned}\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2 > 2\epsilon|A_{mk}) \leq & \exp[-\tfrac{3}{16d_k}(\tfrac{c_3n_m\epsilon}{(1-\lambda_2)^2}-d_k)^2] \\ & +\phi(\|\boldsymbol{\mu}_{mk}\|^2 > \tfrac{\epsilon}{\lambda_1^2}),\end{aligned} \tag{4.4.7}$$

where $\lambda_2 = \frac{\lambda\omega_{mk}}{c_2}$, $\lambda_1 = \frac{\lambda\omega_{mk}}{c_1}$ and $\phi(\cdot)$ is the indicator function.

Combining Lemma 2 and Lemma 3, we have Theorem 3 which estimates the rate of convergence of $\widetilde{\boldsymbol{\mu}}_{mk}$ to $\boldsymbol{\mu}_{mk}$. We prove Theorem 3 as follows:

Proof. For any $m \in \{1,\ldots,M\}$ and large enough n_m, $\forall\ k \in S_m$, $\forall\epsilon > 0$, we could use

Lemma 2 and Lemma 3:

$$
\begin{aligned}
&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2>2\epsilon)\\
=&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2>2\epsilon|A_{mk})\mathbb{P}(A_{mk})\\
&\quad+\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2>2\epsilon|A_{mk}^c)\mathbb{P}(A_{mk}^c)\\
\leq&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2>2\epsilon|A_{mk})+\mathbb{P}(A_{mk}^C)\\
\leq&\exp(-\frac{3}{16d_k}(\frac{c_3n_m\epsilon}{(1-\lambda_2)^2}-d_k)^2)+\phi(\|\boldsymbol{\mu}_{mk}\|^2>\frac{\epsilon}{\lambda_1^2})\\
&+\exp[-(\frac{-\sqrt{d_k+2v_{mk}}+\sqrt{-d_k+n_mc_2^2}}{2})^2]\\
&+\exp[-\frac{(d_k+v_{mk}-2n_mc_1^2)^2}{4(d_k+2v_{mk})}]
\end{aligned}
$$

$\forall\ k\in S_m^c$, $\forall\epsilon>0$, similarly we have

$$
\begin{aligned}
&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2>2\epsilon)\\
\leq&\mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk}-\boldsymbol{\mu}_{mk}\|_2^2>2\epsilon|B_{mk}^C)+\mathbb{P}(B_{mk})\\
\leq&\exp[-\frac{3}{64d_k}\cdot(n_m\lambda^2\omega_{mk}^2-2d_k)^2]
\end{aligned}
$$

The followings are straight forward calculations. □

4.4.5 Proof of Theorem 4

Proof. Using Equation (4.3.11) and assumption (A4), we have

$$[(\widehat{S}_1 = S_1) \cap (\widehat{S}_2 = S_2)] \subset [\widehat{S}_1 \cup \widehat{S}_2 = S_1 \cup S_2] \subset [\widehat{\mathcal{S}} = \mathcal{S}],$$

which implies that

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) &\leq \mathbb{P}((\widehat{S}_1 \neq S_1) \cup (\widehat{S}_2 \neq S_2)) \\ &\leq \mathbb{P}(\widehat{S}_1 \neq S_1) + \mathbb{P}(\widehat{S}_2 \neq S_2).\end{aligned}$$

Using the same argument as in Theorem 3, we further have for $m \in \{1, 2\}$,

$$\mathbb{P}(\widehat{S}_m \neq S_m) \leq \sum_{k \in S_m} \mathbb{P}(A_{mk}^c) + \sum_{k \in S_m^c} \mathbb{P}(B_{mk}).$$

Choosing $\epsilon = \gamma\lambda^2$ and λ^2 sufficiently small, using Theorem 3, we have

$$\begin{aligned}\mathbb{P}(\widehat{S}_m \neq S_m) &\asymp \sum_{k=1}^{K} \mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk}\|_2^2 \geq 2\epsilon) \\ &\asymp K \cdot \max_k \mathbb{P}(\|\widetilde{\boldsymbol{\mu}}_{mk} - \boldsymbol{\mu}_{mk}\|_2^2 \geq 2\epsilon) \\ &\asymp K \cdot O(\exp(-c_4 n^2 \lambda^4)),\end{aligned}$$

where $c_4 = \min_k(\frac{3\gamma^2 c_3^2}{16 d_k}, \frac{3\omega_0^4}{64 d_k})$. Therefore

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{S}} \neq \mathcal{S}) &\leq \mathbb{P}(\widehat{S}_1 \neq S_1) + \mathbb{P}(\widehat{S}_2 \neq S_2) \\ &\leq 2K \cdot O(\exp(-c_4 n^2 \lambda^2)) \to 0.\end{aligned}$$

This completes the proof. □

4.4.6 Proof of Theorem 5

Proof. Denote $\mathcal{A} = \{\widehat{\mathcal{S}} = \mathcal{S}\}$, then by Theorem 4,

$$\mathbb{P}(\mathcal{A}) = 1 - o(1),$$

and according to the proof in Theorem 4, we have for $j \in \{1, 2\}$,

$$\|\widetilde{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\|_2^2 = O_P(\epsilon) = O_P(\frac{\omega_0 \sqrt{\log(n)\log(K)}}{n}). \tag{4.4.8}$$

Using the same techniques in [Bickel and Levina, 2004]'s proof on theorem 2, we could prove that,

$$|\Psi_\Sigma(\widetilde{\Delta}, \widehat{\xi}) - \Psi_\Sigma(\Delta, \xi)| \leq C(\|\widetilde{\Delta} - \Delta\|_2^2 + \|\widehat{\xi} - \xi\|_2^2),$$

for all such that $\|\widetilde{\Delta} - \Delta\|_2 \leq \epsilon_1$, $\|\widehat{\xi} - \xi\|_2 \leq \epsilon_2$ with ϵ_1, ϵ_2 small enough. C is a constant only depending on the model assumptions.

And because of $\widetilde{\Delta} \to \Delta$ and $\widehat{\xi} \to \xi$, given n large enough such that both $\|\widetilde{\Delta} - \Delta\|_2 \leq \epsilon_1$ and $\|\widehat{\xi} - \xi\|_2 \leq \epsilon_2$ hold, we have

$$\mathcal{C}(g) - \mathcal{C}(g^*) \leq C\big(\|\widetilde{\Delta} - \Delta\|_2^2 + \|\widehat{\xi} - \xi\|_2^2\big).$$

For the first term, applying Equation (4.4.8), we have

$$\|\widetilde{\Delta} - \Delta\|_2^2 = O_P(s \cdot \frac{\omega_0 \sqrt{\log(n)\log(K)}}{n}).$$

For the second term, given that $\mathcal{A}$ holds, we have

$$\begin{aligned}
&\|\widehat{\xi} - \xi\|_2^2 \\
=&\|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}}(\widetilde{\boldsymbol{\mu}}_2 - \widetilde{\boldsymbol{\mu}}_1)_{\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2 \\
&\qquad \text{(for large enough } n\text{)} \\
=&\|((\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}})(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}} + \\
&(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} \cdot ((\widetilde{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2)_{\mathcal{M}} - (\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)_{\mathcal{M}})\|_2^2 \\
\leq&\|(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2 \cdot \|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|^2 + \\
&(\|(\widetilde{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2)_{\mathcal{M}}\|_2^2 + \|(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2) \cdot \|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}}\|^2 \\
\leq&\|(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2 \cdot \|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|^2 + \\
&(\|(\widetilde{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2)_{\mathcal{M}}\|_2^2 + \|(\widetilde{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)_{\mathcal{M}}\|_2^2) \cdot \\
&(\|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\| + \|(\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|)^2.
\end{aligned}$$

Because

$$\begin{aligned}
&\|(\widehat{\Sigma}^{-1})_{\mathcal{M}\mathcal{M}} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|^2 \\
&\leq (\|(\widehat{\Sigma}_{\mathcal{M}\mathcal{M}})^{-1} - (\Sigma_{\mathcal{M}\mathcal{M}})^{-1}\| + \|(\Sigma_{\mathcal{M}\mathcal{M}})^{-1} - (\Sigma^{-1})_{\mathcal{M}\mathcal{M}}\|)^2 \\
&= (O_P(s\omega_0^2\sqrt{\frac{\log s \log \omega_0}{n}}) + a_{n,d})^2,
\end{aligned}$$

where the last inequality is by applying Lemma (A.3) of [Bickel and Levina, 2008], we have

$$\|\widehat{\xi} - \xi\|_2^2 = O_P(c_s s^2 \omega_0^4 \cdot \frac{\log s \log \omega_0}{n}) + c_s O(a_{n,d}^2). \tag{4.4.9}$$

This completes the proof. □

Appendix.2. The General Trend by Changing λ

The minimum averaged misclassification errors for gNSC and NSC are highlighted in bold for both cross validation procedures. The results are shown in Table 4.3 and Table 4.4.

Appendix.3. QQ Plots of the Data

For each gene in each sample class we present the Quantile-to-Quantile plot (QQ plot) to visualize the normality. Three of them are shown in Figure 4.4.

Table 4.3: Leave experiment out cross validation method is used on the GPL96 data. Averaged misclassification errors and the corresponding averaged gene numbers across tissue types are provided with standard deviations included. We highlight the minimum values in bold.

λ	gNSC	gene number	NSC	gene number
1.0	0.158(0.3644)	10982(1.3)	0.301(0.0823)	8611(23.39)
2.5	0.150(0.3682)	10189(74.07)	0.240(0.0834)	5778(21.62)
3.5	0.175(0.3715)	8376(16.51)	**0.149(0.0479)**	4495(17.26)
4.0	0.168(0.3488)	7380(15.94)	0.150(0.0487)	3975(14.49)
5.5	0.085(0.1054)	5179(13.48)	0.251(0.0455)	2888(9.89)
6.5	**0.089(0.0505)**	3901(22.50)	0.275(0.0925)	2251(9.53)
7.0	0.120(0.0505)	3515(26.12)	0.275(0.0925)	2081(10.43)
8.5	0.120(0.0505)	2265(40.39)	0.245(0.0950)	1594(10.72)
10.0	0.120(0.0505)	1182(38.05)	0.222(0.0962)	1196(10.38)

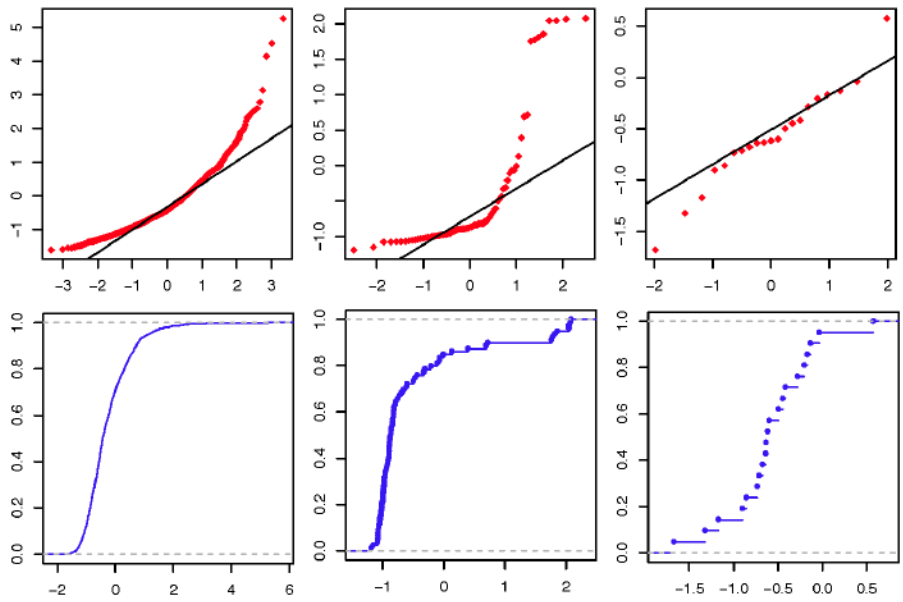


Figure 4.4: Non-Gaussian Data. The two rows are QQ-plot and empirical cdf plot of three different genes.

Appendix.4. Heatplots of gNSC on GPL96

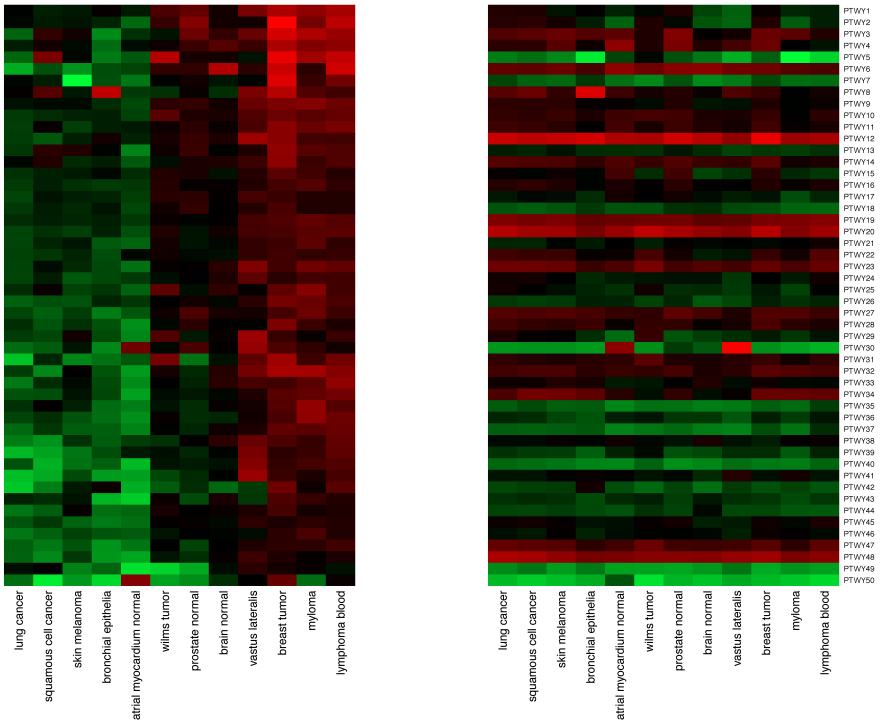
To illustrate our result more clearly, we randomly pick 12 tissue types and 100 gene pathways for visualization. Figure 4.5(a) presents the negative "shrinkage amount," i.e.

Table 4.4: 10-fold cross validation method is used on the GPL96 data. Averaged misclassification errors and the corresponding averaged gene numbers across tissue types are provided with standard deviations included. We highlight the minimum values in bold.

λ	gNSC	gene number	NSC	gene number
1.0	0.141(0.0945)	10982(0.52)	0.145(0.0892)	8538(8.36)
2.5	0.143(0.0892)	10054(19.97)	**0.136(0.0906)**	5670(11.14)
3.0	0.144(0.0843)	9188(26.95)	0.140(0.0900)	4975(9.88)
4.0	0.146(0.0806)	7154(26.06)	0.137(0.0899)	3871(9.17)
5.0	**0.139(0.0866)**	5502(21.33)	0.139(0.0852)	3057(7.86)
5.5	0.181(0.0866)	4874(17.54)	0.144(0.0882)	2735(7.77)
7.0	0.182(0.0923)	3356(15.22)	0.156(0.0852)	2009(5.86)
8.5	0.184(0.0911)	2021(14.62)	0.164(0.0827)	1503(5.53)
10.0	0.189(0.0920)	1052(11.19)	0.170(0.0858)	1143(4.97)

$(1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})$ shown in Equation (4.3.5), of the combination of one certain gene pathway indexed by k and one certain tissue type indexed by m. closer the color to green, the lower the negtive shrinkage amount is. The Figure 4.1 presents the significant associations, i.e. the threshold term $(1 - \frac{\lambda\omega_{mk}}{\|\widehat{\boldsymbol{\mu}}_{mk}\|_k})_+$, between gene pathways and tissue types: red color suggests that the corresponding pathway and tissue type are estimated to be associated, while green suggests not. Moreover, the expression levels in one block are summarized to be the mean of all gene expression values confined in this block and the result can be observed in Figure 4.5(b). Figures 4.5(a) and 4.1 illustrate the statistical significance levels and Figure 4.5(b) illustrates the biological expression levels. A detailed index for all gene pathways are presented in the Appendix 4.4.6.

Appendix.5. Heatplot of the Keyword-Gene Relevance



(a) (b)

Figure 4.5: Heatplots for gNSC. The 100 pathways in this figure are randomly chosen.

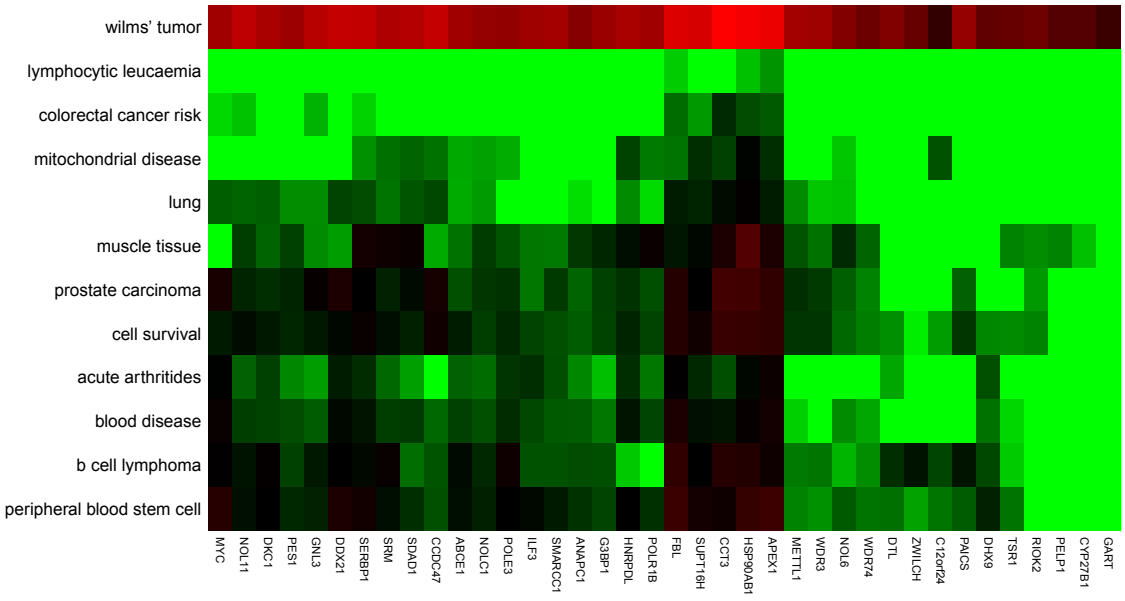


Figure 4.6: NSC Results of Keywords v.s. Genes.

Appendix.6. Gene Pathways

Table 4.5: The index of all gene pathways IDs.

ID	Pathway name
PTWY1	landis erbb2 breast preneoplastic up
PTWY2	waesch anaphase promoting complex
PTWY3	sanchez mdm2 targets
PTWY4	naderi breast cancer prognosis dn
PTWY5	begum targets of pax3 foxo1 fusion and pax3
PTWY6	der ifn alpha response dn
PTWY7	negative regulation of cytokine biosynthetic process
PTWY8	aldo keto reductase activity
PTWY9	xu hgf signaling not via akt1 6hr
PTWY10	valk aml cluster 15
PTWY11	nam fxyd5 targets dn
PTWY12	creighton akt1 signaling via mtor dn
PTWY13	gargalovic response to oxidized phospholipids green up
PTWY14	appierto response to fenretinide up
PTWY15	chiaradonna neoplastic transformation kras cdc25 dn
PTWY16	yao temporal response to progesterone cluster 8
PTWY17	tonks targets of runx1 runx1t1 fusion granulocyte up
PTWY18	ebauer targets of pax3 foxo1 fusion dn
PTWY19	zhan v2 late differentiation genes
PTWY20	hu angiogenesis dn
PTWY21	shipp dlbcl vs follicular lymphoma dn
PTWY22	mori mature b lymphocyte dn
PTWY23	caffarel response to thc dn
PTWY24	nucleotide metabolic process
PTWY25	yang breast cancer esr1 bulk dn
PTWY26	hoegerkorp cd44 targets temporal dn
PTWY27	vesicle membrane
PTWY28	nikolsky breast cancer 1q21 amplicon
PTWY29	naderi breast cancer prognosis up
PTWY30	rickman head and neck cancer f
PTWY31	myllykangas amplification hot spot 9
PTWY32	scheidereit ikk targets
PTWY33	radaeva response to ifna1 dn
PTWY34	zhang interferon response
PTWY35	regulation of t cell activation
PTWY36	finetti breast cancers kinome gray
PTWY37	positive regulation of lymphocyte activation
PTWY38	cortical cytoskeleton
PTWY39	rna polymerase ii transcription mediator activity
PTWY40	meiotic cell cycle
PTWY41	kim response to tsa and decitabine dn
PTWY42	vitamin transport
PTWY43	regulation of rho gtpase activity
PTWY44	brain development
PTWY45	regulation of binding
PTWY46	ginestier breast cancer znf217 amplified up
PTWY47	zhan multiple myeloma subgroups
PTWY48	proteasome complex
PTWY49	organic anion transmembrane transporter activity
PTWY50	krishnan furin targets dn

Appendix.7. Context Analysis

Microarray data often have text documents of samples (e.g. sample description, experiment description, etc.). We propose a new "index-gene relevance" approach to applying NSC or gNSC to context analysis.

More precisely, we encode relationships between representative medical terms, and pertaining genes in matrix form, to which we apply NSC or gNSC. The procedure of creating the index-gene relevance matrix includes four steps: preparing documents for each microarray sample; creating a tf-idf (defined later) based term-document matrix; creating an index-document matrix; and lastly combining text information with the microarray data. We describe the details of these steps below.

Step 1. Document Preparation. In order to use the text information efficiently, we extract the biologically meaningful words and phrases from the text description files and map them into existing knowledge sources. For each sample, we produce a text file with sample specific information, in which "meaningful phrases" and related words are organized into rows. Rather than describing the semantic criteria by which we extract "meaningful phrases," we simply provide an example.

To utilize the text information of GPL96 data, we prepare the sample documents in five steps as shown in Table 4.6.

Step 2. TF-IDF based Term-doc Matrix. "Term frequency-inverse document frequency" (tf-idf) [Wu et al., 2008] is one of the most commonly used relevance weighting factors in today's information retrieval and text mining systems, and is our preferred relevance

Table 4.6: Document Preparation

Procedure
1: Download the file GPL96_family.soft.gz (`www.affymetrix.com`) with text description of the samples and the experiments information of the samples. 2: Retrieve sample information (GSM files, description of the specific sample information involved in one entire experiment) for all experiments from the GPL96 family.soft file. 3: Extract the biologically meaningful sample and experiment information from GSM files. Sample title, sample source, sample organism, sample characteristics and sample description are extracted. 4: Use MetaMap to map all the information extracted from previous step to several knowledge sources, including GO, MSH, HUGO, OMIM and NCI. MetaMap can break the input text into several phrases by its lexical/syntactic analysis and then map those phrases to the knowledge source. 5: We denote each phrase together with its all related words (all the words are lower-cased) as a block (one row in the file) in the text document file, which has the same name as the original GSM files.

metric. The tf-idf value increases proportionally to the number of times a word appears in a specific document, but is offset by the frequency of the word in the corpus. This provides a good measure of relevance which controls for the fact that some words are generally more common than others.

Let doc_i be the text document of sample i and $Doc = \{doc_i : i = 1, 2, \cdots, n\}$ be the set of all documents. Each document is represented as a list of words:

$$doc = (w_1, ..., w_{N_{doc}}),$$

where w_i $(i = 1, 2, \cdots, N_{doc})$ are the words in the doc document, including repetition. N_{doc} is the total number of words in doc. We extract all *distinct* words from all documents

and use W to represent the set of all words:

$$W = \cup_{doc \in Doc} \cup_{j=1}^{N_{doc}} \{doc(j)\}, \tag{4.4.10}$$

where $doc(j)$ is the j-th word of doc. Note that W here is a set of word where each elements is unique. For each extracted term w and the document doc, we define the term-count $tc(w, doc)$ to be the number of times that the term w appears in doc.

To prevent the bias towards longer documents, the term-frequency(tf) is defined as:

$$tf(w, doc) = \frac{tc(w, doc)}{|doc|}, \tag{4.4.11}$$

where $|doc| = N_{doc}$ is the length (total number of words) of the document doc.

In addition, we introduce the inverse document frequency for each word w which is a measure of the general importance of w:

$$idf(w) = \log \frac{|Doc|}{|\{doc : w \in doc\}|}, \tag{4.4.12}$$

where $|Doc|$ is the total number of documents which equals the number of samples and $|\{doc : w \in doc\}|$ is the number of documents in which the word w appears. Based on tf-idf score we build the term-doc matrix tdM:

$$tdM(w, doc) = tf(w, doc) idf(w). \tag{4.4.13}$$

As we can see, tf-idf score is

1. highest when the term w appears many times within a small number of documents;
2. lower when the term w appears fewer in a document, or appears in too many documents;
3. lowest when the term appears in nearly all documents.

Step 3. Index-doc Matrix. The tf-idf based term-doc matrix reflects how important a word is to a document in the corpus. However, we are not only interested in words but also biologically pertinent phrases with multiple terms (e.g. "breast cancer," "brain tumor," etc.). We measure the relevance of such phrases as follows:

Let $p_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,|p_i|})$ be any word ($|p_i| = 1$) or phrase ($|p_i| \geq 2$) in the dictionary, where $w_{i,j}$ ($j = 1, 2, \cdots, |p_i|$) are the single terms in p_i and $|p_i|$ is the number of terms. Let $P = (p_1, p_2, \cdots, p_N)$ be the list of all words and phrases. Let ind_i be the index of p_i such that words or phrases with same meaning (e.g., "brain" and "brains") in the dictionary have the same index. Therefore, each index can represent a synonym word group. Let $I = \{ind_1^*, ind_2^*, \cdots, ind_M^*\}$ be the set of indices, where each element is unique. Note that since the indices of the words and phrases in the word list P are not unique, M need not equal N. We build a index-doc matrix, idM, based on the tf-idf score:

$$idM(ind^*, doc) = \max_{i:ind_i = ind^*} \left\{ \frac{\sum_j tdM(w_{i,j}, doc)}{N_{p_i}} \right\}, \tag{4.4.14}$$

where each element is the maximum of the mean tf-idf score for words and phrases with same meaning.

Step 4. Gene-Index Relevance Matrix. In order to apply NSC to context analysis, we combine the text information with our microarray data. We generate the relevance matrix of words/phrases with genes based on the gene expression levels in microarray data and the index-doc matrix generated above. Let the m-th synonym group be the words or phrases with their index $= ind_m^*$. We can use ind_m^* can represent this group.

The index-doc matrix we get from Step 3 can be seen as a measure of relevance of synonym word groups and samples, while the gene expression levels in microarray data can be seen as a measure of gene-sample relevance. We measure the connection of gene and synonym groups by multiplying the elements of the above two relevance matrix.

In mathematics, we measure the connection of the m_{th} synonym group and the j_{th} gene for a given sample i as:

$$R_m(g_j, doc_i) = idM(ind_m^*, doc_i) \times x_{ij}, \tag{4.4.15}$$

where x_{ij} is the expression level of j_{th} gene in sample i and idM is the index-doc matrix in Step 3.

To respect the structure of the NSC input data, we collect the R matrices to form the gene-index relevance matrix:

$$R = (R_1, R_2, \cdots, R_M). \tag{4.4.16}$$

Here each synonym group is regarded as a sample class, i.e. C_m, in NSC.

By applying NSC to the gene-index relevance matrix, we can select the most relevant genes with all words and phrases in the dictionary. Moreover, we can trivially generalize the NSC approach to context analysis to the group version by first combining the pathway information with the microarray data before using gNSC to the corresponding gene-index relevance matrix.

Remark 4.4.2. *As alluded to previously (cf. Remark 4.4.1), although the dimension of the gene-index relevance matrix R is $d \times (n \times M)$, which can grow prohibitively large, our process can still be completed in minutes. This efficiency is due to the use of sufficient statistics.*

Table 4.7: NSC(/gNSC) for Context Analysis

Algorithm
1: Extract the biological meaningful words and phrases from the text description files of each microarray sample and map them into existing knowledge sources. Prepare a text document file for each sample.
2: Calculate the term-frequency(tf) and the inverse document frequency(idf) for each word w, then multiply them to get the tf-idf score for each word and each document. Build the term-doc matrix(tdM) based on the score.
3: Divide the words into synonym groups and calculate the index-doc matrix(idM), based on equation 4.4.14, for each synonym group and each document.
4: For gNSC, sort the genes in the microarray data by pathways.
5: Calculate the R matrixes and bind them together to get the gene-index relevance matrix.
6: Apply NSC or gNSC on the gene-index relevance matrix to selection the significant genes or pathways for each word or phrase in the dictionary.

Chapter 5

Discussion

CHAPTER 5. DISCUSSION

In the first part of the dissertation, we presented a new group ICA method called homotopic group ICA (H-gICA). It improves the power of finding the underlying brain networks by rearranging the data structure and effectively doubles the sample size. Both the simulation study and the application on ADHD 200 data shows the improvement of H-gICA comparing with with ordinary group ICA . The main networks found included visual networks, the default mode network and the auditory network. In addition, H-gICA enables measurement of the functional homotopy of the underlying functional networks and offers the opportunity to analyze the relation of the brain functional homotopy between the left and right hemisphere with diseases.

In the second part of the dissertation, we posited a different paradigm for statistically evaluating task related activation before and after learning. The simulation results suggest that tests of dimension are a reasonable exploratory procedure for investigating the distribution of paired activation maps. The investigation of activation distribution represents a complementary procedure to voxel-level testing and does not represent a form of omnibus test to be performed prior to voxel-level testing.

In the third part of the dissertation, we propose a group nearest shrunken centroid method (gNSC) for classification and feature selection to deal with big data of microarray expressions. GNSC enables us to use gene pathway information and the sample text information. The application of gNSC on the GPL96 data [McCall et al., 2010] shows a higher power of classification comparing with NSC.

Bibliography

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567, 2002.

Ani Eloyan, Shanshan Li, John Muschelli, Jim J Pekar, Stewart H Mostofsky, and Brian S Caffo. Analytic programming with fmri data: A quick-start guide for statisticians using r. *PloS one*, 9(2):e89470, 2014.

V. D. Calhoun, T. Adali, G.D. Pearlson, and J. J. Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human Brain Mapping*, 14:140–151, 2001a.

V.D. Calhoun, J. Liu, and T. Adalı. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1 Suppl):S163, 2009.

VD Calhoun, T. Adali, GD Pearlson, and JJ Pekar. A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001b.

BIBLIOGRAPHY

Shanshan Li, Ani Eloyan, Suresh Joel, Stewart Mostofsky, James Pekar, Susan Spear Bassett, and Brian Caffo. Analysis of group ica-based connectivity measures from fmri: application to alzheimer's disease. *PloS one*, 7(11):e49340, 2012.

X.N. Zuo, C. Kelly, A. Di Martino, M. Mennes, D.S. Margulies, S. Bangaru, R. Grzadzinski, A.C. Evans, Y.F. Zang, F.X. Castellanos, et al. Growing together and growing apart: regional and sex differences in the lifespan developmental trajectories of functional homotopy. *The Journal of Neuroscience*, 30(45):15034–15043, 2010.

E. Zarahn. Using larger dimensional signal subspaces to increase sensitivity in fmri time series analyses. *Human Brain Mapping*, 17(1):13–16, 2002.

K.J. Worsley, J.B. Poline, K.J. Friston, and AC Evans. Characterizing the response of pet and fmri data using multivariate linear models. *NeuroImage*, 6(4):305–319, 1997.

K.V. Mardia, J.T. Kent, and J.M. Bibby. Multivariate analysis. 1980.

B. Biswal, F. Zerrin Yetkin, V.M. Haughton, and J.S. Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4):537–541, 1995.

D.A. Gusnard, M.E. Raichle, ME Raichle, et al. Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience*, 2(10):685–694, 2001.

JS Damoiseaux, CF Beckmann, E.J.S. Arigita, F. Barkhof, P. Scheltens, CJ Stam,

SM Smith, and S. Rombouts. Reduced resting-state brain activity in the default network in normal aging. *Cerebral Cortex*, 18(8):1856–1864, 2008.

S.A.R.B. Rombouts, F. Barkhof, R. Goekoop, C.J. Stam, and P. Scheltens. Altered resting state networks in mild cognitive impairment and mild alzheimer's disease: an fmri study. *Human brain mapping*, 26(4):231–239, 2005.

Y. Guo and G. Pagnoni. A unified framework for group independent component analysis for multi-subject fmri data. *NeuroImage*, 42(3):1078–1093, 2008.

C.F. Beckmann and S.M. Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *Medical Imaging, IEEE Transactions on*, 23(2): 137–152, 2004.

M.J. McKeown, S. Makeig, G.G. Brown, T.P. Jung, S.S. Kindermann, A.J. Bell, and T.J. Sejnowski. Analysis of fmri data by blind separation into independent spatial components. Technical report, DTIC Document, 1997.

C.F. Beckmann and S.M. Smith. Tensorial extensions of independent component analysis for multisubject fmri analysis. *Neuroimage*, 25(1):294–311, 2005.

VD Calhoun, T. Adali, VB McGinty, JJ Pekar, TD Watson, GD Pearlson, et al. fmri activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *NeuroImage*, 14(5):1080–1088, 2001c.

F. Esposito, T. Scarabino, A. Hyvarinen, J. Himberg, E. Formisano, S. Comani,

BIBLIOGRAPHY

G. Tedeschi, R. Goebel, E. Seifritz, F. Di Salle, et al. Independent component analysis of fmri group studies by self-organizing clustering. *Neuroimage*, 25(1):193–205, 2005.

C. Sorg, V. Riedl, M. Mühlau, V.D. Calhoun, T. Eichele, L. Läer, A. Drzezga, H. Förstl, A. Kurz, C. Zimmer, et al. Selective changes of resting-state networks in individuals at risk for alzheimer's disease. *Proceedings of the National Academy of Sciences*, 104(47): 18760–18765, 2007.

A. Garrity, G. Pearlson, K. McKiernan, D. Lloyd, K. Kiehl, and V. Calhoun. Aberrant default mode functional connectivity in schizophrenia. *American Journal of Psychiatry*, 164(3):450–457, 2007.

F. Sambataro, V.P. Murty, J.H. Callicott, H.Y. Tan, S. Das, D.R. Weinberger, and V.S. Mattay. Age-related alterations in default mode network: impact on working memory performance. *Neurobiology of aging*, 31(5):839, 2010.

R. Toro, P.T. Fox, and T. Paus. Functional coactivation map of the human brain. *Cerebral cortex*, 18(11):2553–2559, 2008.

M.P. Milham. Open neuroscience solutions for the connectome-wide association era. *Neuron*, 73(2):214–218, 2012.

Ani Eloyan, John Muschelli, Mary Beth Nebel, Han Liu, Fang Han, Tuo Zhao, Anita D Barber, Suresh Joel, James J Pekar, Stewart H Mostofsky, et al. Automated diagnoses

of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6:61, 2012.

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

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634, 1999.

P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

E. Oja, A. Hyvarinen, and J. Karhunen. Independent component analysis, 2001.

S.U. Pillai et al. *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education, 2002.

Ani Eloyan, Ciprian M. Crainiceanu, and Brian S. Caffo. Likelihood-based population independent component analysis. *Biostatistics*, 2013.

M. Mennes, B. Biswal, F.X. Castellanos, and M.P. Milham. Making data sharing work: The fcp/indi experience. *NeuroImage*, 2012.

VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *Neuroimage*, 47:S102, 2009.

BIBLIOGRAPHY

V. Fonov, A.C. Evans, K. Botteron, C.R. Almli, R.C. McKinstry, and D.L. Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313, 2011.

K.J. Friston, J.T. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny. *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Academic Press, 2011.

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

M.A. Lindquist. The statistical analysis of fmri data. *Statistical Science*, 23(4):439–464, 2008.

P. Diggle, P. Heagerty, K.Y. Liang, and S. Zeger. *Analysis of longitudinal data*, volume 25. Oxford University Press, USA, 2002.

T.E. Nichols and A.P. Holmes. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25, 2001.

G. McLachlan and D. Peel. *Finite mixture models*, volume 299. Wiley-Interscience, 2000.

J.R. Moeller and C.G. Habeck. Reciprocal benefits of mass-univariate and multivariate modeling in brain mapping: Applications to event-related functional mri, h 2 15 o-, and fdg-pet. *International Journal of Biomedical Imaging*, 2006, 2006.

BIBLIOGRAPHY

Lior Shmuelof, John W Krakauer, and Pietro Mazzoni. How is a motor skill learned? change and invariance at the levels of task success and trajectory control. *Journal of neurophysiology*, 108(2):578–594, 2012.

Lior Shmuelof, Juemin Yang, Brian Caffo, Pietro Mazzoni, and John W Krakauer. The neural correlates of learned motor acuity. *Journal of neurophysiology*, pages jn–00897, 2014.

Jean Talairach and Pierre Tournoux. Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: an approach to cerebral imaging. 1988.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers. Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute, May*, 2011.

P.J. Bickel and E. Levina. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using ?1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

BIBLIOGRAPHY

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in neural information processing systems*, 16(1):49–56, 2004.

J. Friedman, T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. [consistency in boosting]: Discussion. *The Annals of Statistics*, 32(1):102–107, 2004.

Y. Kobayashi, D.M. Absher, Z.G. Gulzar, S.R. Young, J.K. McKenney, D.M. Peehl, J.D. Brooks, R.M. Myers, and G. Sherlock. Dna methylation profiling reveals novel biomarkers and important roles for dna methyltransferases in prostate cancer. *Genome research*, 21(7):1017–1027, 2011.

J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High dimensional semiparametric gaussian copula graphical models. *Arxiv preprint arXiv:1202.2169*, 2012.

M.N. McCall, B.M. Bolstad, and R.A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.

A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43): 15545, 2005.

BIBLIOGRAPHY

H. Ji, G. Wu, X. Zhan, A. Nolan, C. Koh, A. De Marzo, H.M. Doan, J. Fan, C. Cheadle, M. Fallahi, et al. Cell-type independent myc target genes reveal a primordial signature involved in biomass accumulation. *PloS one*, 6(10):e26057, 2011.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10: 2295–2328, 2009.

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

S. Wang and J. Zhu. Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23(8):972–979, 2007.

Trevor. Hastie, Robert. Tibshirani, and JH (Jerome H.) Friedman. *The elements of statistical learning*. Springer, 2009.

C.A.J. Klaassen and J.A. Wellner. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, 1997.

P.J. Bickel. *Efficient and adaptive estimation for semiparametric models*. Springer Verlag, 1998.

E.J. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H.C. Hsu, J.D. Mountz, N.E. Baldwin, M.A. Langston, et al. Complex trait analysis of gene expression uncovers polygenic

and pleiotropic networks that modulate nervous system function. *Nature genetics*, 37(3): 233–242, 2005.

H. Lovec, A. Grzeschiczek, MB Kowalski, and T. Möröy. Cyclin d1/bcl-1 cooperates with myc genes in the generation of b-cell lymphoma in transgenic mice. *The EMBO journal*, 13(15):3487, 1994.

P.J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

H.C. Wu, R.W.P. Luk, K.F. Wong, and K.L. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3): 13, 2008.

Chapter 6

Curriculum Vitae

JUEMIN YANG

Johns Hopkins School of Public Health	Office: (410)-502-3364
Department of Biostatistics	Email: juyang@jhsph.edu
615 North Wolfe Street	Homepage:
Baltimore MD 21205	http://www.biostat.jhsph.edu/ juyang

Education

- **Ph.D. Candidate**, Biostatistics, Johns Hopkins Bloomberg School of Public Health, *expected* 2015
  - *Advisors:* Dr. Brian S. Caffo
- **M.S.**, Computer Science, Johns Hopkins University, United States, 2013-2014
- **B.S.**, Statistics, Peking University, China, 2006-2010
- **B.S.**, Economics, Peking University, China, 2007-2010

Professional Experience and Activities

- 2007-Present **Research Assistant**, Johns Hopkins University

  • Conducted large-scale data analysis on longitudinal imaging data using Independent Component Analysis

  • Developed procedures to examine the change-point in biomarkers

  • Proposed a scalable discriminant analysis approach for feature selection and classification

- 2007-Present **Teaching Assistant**, Johns Hopkins University

  • Statistical Computing, Survival Analysis, Statistical Methods in Public Health I - IV, and Public Health Biostatistics

- 2014 Summer **Intern**, AT&T Research Labs

  • Conducted text analysis on TV service call data using Latent Dirichlet Allocation (LDA)

- 2013 Summer **Intern**, AT&T Labs

  • Conducted dimension reduction, blind source separation and feature selection of high dimensional data using Principal Component Analysis (PCA) and principal component regression (PCR)

  • Identified underlying causing factors of AT&T network performance indicators using Bayesian Network

- 2013 **Visiting Scholar**, Department of Biostatistics and Epidemiology, University of Pennsylvania

- 2013 Summer **Summer of Code Student Participant**, Google Inc.

  • Developed an R package for a modern nonparametric predictive method, Sparse Additive Models (SpAM)

Publications

- **Yang, J.**,Han, F., Irrizary, R. and Liu, H. (2014), Group Nearest Shrunken Centroids in Large-Scale Genomic Studies, AISTATS.

- **Yang, J.**, Eloyan, A., Barber, A., Nebel, M., Mostofsky, S., Pekar, J., Crainiceanu, C. and Caffo, B., Homotopic Group ICA for Multi-Subject Brain Imaging Data, Submitted.

- **Yang, J.**, Shmuelof, L. , Luo, X., Krakauer, J. and Caffo, B., Lessons Learned from Activation Studies of Learning using Tests of Dimension Reduction, Submitted.

- Shmuelof, L., **Yang, J.**, Caffo, B., Mazzoni, P., Krakauer, JW. (2014), The neural correlates of learned motor acuity, Journal of neurophysiology.

- Sasaki, T., Miller, C., Hansford, R., **Yang, J.**, Caffo, B., Zviman, M. et al. (2012), Myocardial Structural Associations with Local Electrograms: A Study of Post-Infarct

Ventricular Tachycardia Pathophysiology and Magnetic Resonance Based Non-Invasive Mapping, Circulation: Arrhythmia and Electrophysiology.

Presentations

- 2013 "Homotopic group ICA for Multi-subject Brain Imaging Data", Joint Statistical Meetings, Montreal, Canada
- 2014 "On the Road to an Effortless Experience - Reducing Repeat Calls to Agents from U-Verse Customers", AT&T Research Labs
- 2013 "Identify relation and underlying causing factors of network performance indicators - Identify relation and underlying causing factors of network performance indicators", AT&T Labs
- 2012 "Homotopic group ICA for Multi-subject Brain Imaging Data", Hopkins Imaging Conference

Honors and Awards

2013	June B. Culley Award for Outstanding Achievement on the Schoolwide Exam Paper
2012	The Peer Choice Award for the Audience's Favorite Poster, Hopkins Imaging Conference
2010	1^{st} prize, Student Paper Competition, Peking University
2007	National Undergraduate Student Research scholarship
2007	Guanghua scholarship, Peking University
2006	Freshmen Scholarship, Peking University
2006	Member of Chinese National Training Team for the 47th International Mathematical Olympiad (IMO)
2006	Gold Medal, Chinese Mathematical Olympiad (CMO)

Professional Development

- **Language Skills:** English (Fluent), Chinese (Native)
- **Computer Skills:** R, JAVA, Python, C/C++, Stata, SAS, SQL