Distribution Agreement

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

Signature:

_____ _____

Ran Shi Date

Some Novel Statistical Methods for Neuroimaging Data Analysis

By

Ran Shi
Doctor of Philosophy
Biostatistics

Ying Guo, Ph.D.
Advisor

Jian Kang, Ph.D.
Advisor

Xiaoping Hu, Ph.D.
Committee Member

Suprateek Kundu, Ph.D.
Committee Member

Accepted:

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

Date

Some Novel Statistical Methods for Neuroimaging Data Analysis

By

Ran Shi
B.S., Peking University, 2011

Advisors: Ying Guo, Ph.D. and Jian Kang, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2016

Abstract

Some Novel Statistical Methods for Neuroimaging Data Analysis
by Ran Shi

In this dissertation, we propose three novel statistical methods for analyzing neuroimaging data.

In the first topic, we propose a hierarchical covariate-adjusted ICA (hc-ICA) model that provides a formal statistical framework for estimating covariate effects and testing differences between brain functional networks. Our method provides a more reliable and powerful statistical tool for evaluating group differences in brain functional networks while appropriately controlling for potential confounding effects. We present two EM algorithms to obtain maximum likelihood estimates of our model. We introduce a voxel-wise approximate inference procedure which eliminates the need of computationally expensive covariance matrix estimation and inversion. We demonstrate the advantages of our methods over the existing method via simulation studies. We apply our method to an fMRI study to investigate differences in brain functional networks associated with post-traumatic stress disorder (PTSD).

In the second topic, we propose a spatially varying coefficient model (SVCM) with structured sparsity and region-wise smoothness. A new class of nonparametric Bayesian priors is developed named thresholded Gaussian processes (TGP). We show that TGP has a large prior coverage on the space of region-wise smooth functions with restricted supports, leading to posterior consistency in both estimation and feature selection. Efficient posterior computation algorithms are developed by adopting a kernel convolution approach. Based on simulation studies, we demonstrate that our methods can achieve better performance in estimating functional coefficients and selecting imaging features. The application of our proposed method to a resting state functional magnetic resonance imaging (rs-fMRI) data provides biologically meaningful findings.

In the third topic, we present a new independent component analysis (ICA) model with spatially dependent source signals. We model the conditional expectation of IC source signals using Bayesian nonparametric kernel models, which can generate flexible prior spatial dependence structures. We adopt a fully Bayesian approach to make posterior inference about our model through an efficient Markov chain Monte Carlo algorithm. Simulation studies show that, compared with existing ICA algorithms, our method estimates the mixing matrices more accurately and identifies the spatial activation patterns more precisely. When applied to a real fMRI dataset, our method elicits meaningful scientific findings.

Some Novel Statistical Methods for Neuroimaging Data Analysis

By

Ran Shi
B.S., Peking University, 2011

Advisors: Ying Guo, Ph.D. and Jian Kang, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2016

Acknowledgements

I would like to thank my advisors, Dr. Ying Guo and Dr. Jian Kang, for teaching me how to become a real scholar, for guiding me to do good research, for fostering my passion about academics and for supporting me without reservation throughout my stint at Emory. They are my lifetime mentors.

I would like to thank my dissertation committee members, Dr. Suprateek Kundu and Dr. Xiaoping Hu, for their thoughtful comments, constructive criticism and invaluable suggestions. They help me improve my dissertation work significantly.

I would like to thank all the people in the biostatistics and bioinformatics department at Emory for creating an extraordinary educational environment. I cannot complete this work without their generous help and continuous support.

I would also like to thank my parents and my wife for their understanding, their support and their love. I dedicate this work to them.

Finally, I want to take a chance to salute to those crazy ones, those who see things differently and those who are crazy enough to think that they can change the world. All of you, whether successful or not, are my role models.

Contents

1 Introduction **1**

1.1 Overview 1

1.2 Some current research topics 2

1.2.1 Functional brain connectivity 2

1.2.2 Activation study and feature selection 6

1.2.3 Spatial dependence when performing ICA dempoition 8

1.3 Literature Review 12

1.4 Outlines 16

2 Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis 17

2.1 Methods 17

2.1.1 Preprocessing prior to ICA 17

2.1.2 A hierarchical covariate-adjusted ICA model (hc-ICA) . 18

2.1.3 Source signal distribution assumptions 20

2.1.4 Maximum likelihood estimation 21

2.1.5 Inference for covariate effects in hc-ICA model 27

2.2 Application to fMRI data from Grady PTSD study 30

2.2.1 Experimental design, image acquisition and pre-processing 30
2.2.2 Analysis and findings 31
2.3 Simulation Study 36
2.3.1 Simulation study I: performance of the hc-ICA v.s. TC-GICA 36
2.3.2 Simulation study II: performance of the approximate EM 38
2.3.3 Simulation study III: performance of the proposed inference procedures for covariate effects 42
2.4 Discussion 43
2.5 Appendices 45
2.5.1 The Conditional Expectation Function in the E-step . . 45
2.5.2 The derivation of conditional probabilities in the E-step 46
2.5.3 Details of the M-step in the exact EM 47
2.5.4 Proof of Theorem 1 49
2.5.5 Remarks on the subspace-based approximate EM . . . 51
2.5.6 Thresholding the spatial maps based on the ML estimates for functional brain networks 51
2.5.7 Specifying the initial values for hc-ICA 51
2.5.8 Additional Simulation Studies 52
2.5.9 Checking the stability of our EM algorithm for the PTSD data analysis 56

3 Bayesian Spatial Feature Selection for Massive Neuroimaging Data via Thresholded Gaussian Processes 59

3.1 Feature selection within the spatially varying coefficient functions 59
3.1.1 The spatially varying coefficient model for neuroimaging data 60
3.1.2 The thresholded Gaussian process priors 62
3.2 Theoretical Results 64
3.3 Posterior Inferences 68
3.3.1 Model Representation 68
3.3.2 Hyper Prior Specifications 70
3.3.3 Kernel Expansion for Massive Data Analysis 72
3.3.4 A Markov chain Monte Carlo Algorithm 73
3.3.5 Posterior Inference on SVCFs 74
3.4 Numerical Examples 75
3.4.1 Simulation Study: Synthetic Imaging Data 75
3.4.2 Real Data Application: The Autism Brain Imaging Data Exchange (ABIDE) 81
3.5 Discussion 84
3.5.1 Proof of Theorem 1 86
3.5.2 Proof of Theorem 2 88
3.5.3 Proof of Theorem 3 95
3.5.4 Details about the MCMC algorithm 96

4 Bayesian Independent Component Analysis Involving Spatially Dependent Sources With Application to fMRI Data 99

4.1 Method 99

4.1.1 Preprocessing of fMRI data 99
4.1.2 The spatially dependent ICA model for fMRI data . . . 100
4.1.3 Model representation, hyperprior specification and posterior inference 104
4.2 Data Examples 110
4.2.1 Simulated data 110
4.2.2 Real resting-state fMRI data 113
4.3 Discussion 119
4.4 Appendices 120
4.4.1 Proof of Theorem 5 120
4.4.2 Details about the algorithm to draw from the posterior 122
4.4.3 Functional eigen-decomposition for the generalized SE kernel 125

5 Summary and Future Work **127**

List of Figures

1.1 The empirical (stationary) spatial correlation functions, $\kappa(\cdot)$, of four esitmated brain functional networks based on our resting-state fMRI data example: Each subfigure contains 100 curves; each curve is plotted based on $2,500$ randomly sampled voxels (without replacement) since the total number of voxels is too large to compute the empirical correlation function. 10

2.1 The estimated subpopulational maps for the PTSD$-$ and PTSD$+$ women at the median age (36 year old) and the median depression score (BDI=10): Panel (A) shows the estimates for the network featuring the visual cortex, which has the highest positive correlation with the task time series. Panel (B) shows the estimates for the default mode network, which has the largest negative correlation with the task time series. All IC maps are thresholded at the posterior probability of activation above 0.9. PTSD$+$ women show stronger IC signals in both networks. 32

2.2 p-values, thresholded below 0.01, for comparing the adjusted PTSD group differences (PTSD$-$ < PTSD$+$) in the task-related network: hc-ICA found increased spatial source signals at the central part of the visual cortex among PTSD$+$ women, which remained significant after FDR control; dual-regression found much less group differences in the network, all of which became insignificant with the FDR control. 34

2.3 p-values, thresholded below 0.01, for comparing the adjusted PTSD group differences (PTSD$-$ < PTSD$+$) in the default mode network: hc-ICA finds stronger network activities across all the major regions of this network for PTSD$+$ women. Many of these identified voxels still appear after FDR control; dual-regression findings only dicover a few differences in the PCC and mPFC regions. 35

2.4 Comparison between our method and dual-regression ICA: truth, estimates from our model and estimates from dual-regression (N=10, between-subject variabilities are medium) are displayed based on 100 runs. All the images displayed are averaged across the 100 Monte Carlo datasets. Population-level spatial maps are shown in Figure 2.4(A). The results of dual-regression ICA are contaminated by the covariate effects. The results from our method are more accurate. Covariate effect estimates are shown in Figure 2.4(B1) and Figure 2.4(B2) respectively. The results of dual-regression show clear mismatching while our method provide accurate estimates. . 39

2.5 Evaluating the validity of subspace-based approximation EM for data generated from overlapping ICs: three panels show results from different level of overlapping; the two columns in each panel correspond to different ICs; the three rows in each panel represent true source signals, estimates based on the exact EM and estimates from the approximate EM 54

3.1 Sampling SVCFs from the TGP prior 64

3.2 Simulated $\widehat{\ell}(\lambda_k)$ from 50 synthetic datasets: ground truth ($\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 4$) are marked in the figures. 71

3.3 Column 1-5: true and estimated spatial covariate effects from GLM-t, GLM-FDR, GLM-RFT and SVCM-TGP; Column 6: the selection probability estimated from SVCM-TGP. The result is generated from one simulated dataset with $m = 200$ subjects, $n = 2500$ pixels and noise level $\sigma^2(\mathbf{s}) \in [16, 20]$. 76

3.4 The ROC curves for our method (red curves) compared with GLM fittings using the FDR control (blue curves) or the FWER correction based on the random field theory (cyan curves) under four different distribution assumptions ($n = 2500,\ m = 200,\ \sigma^2(\mathbf{s}) \in [16, 20]$). 79

3.5 $\ell(\lambda_k)$ for specifying λ priors in the analysis of ABIDE data. The colored shades mark the intervals we choose as the range of the uniform priors for λ. 82

3.6 Estimated SVCFs (top row in each subplot) and regional selection probabilities (bottom row in each subplot) based on posterior samples from our MCMC algorithm for “ASD group", “age" and “ASD group$\times$ gender" 83

4.1 The true mean functions of the spatial sources signals used in our simulation studies. The mean function $\mu_1(\mathbf{s})$ is constantly zero. 109

4.2 The new true mean functions of the spatial sources signals used in additional simulation studies (setting three: smaller area of activation). The mean function $\mu_1(\mathbf{s})$ is constantly zero. 113

4.3 The Amari errors from different ICA methods. SDP: our ICA method with spatially dependent sources; Fast: FastICA; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function 114

4.4 The ROC curves for different ICA methods. SDP: our ICA method with spatially dependent sources; Fast: FastICA; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function 115

4.5 The average running time for different ICA methods (FastICA is excluded for its convergence issues). SDP: our ICA method with spatially dependent sources; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function. . . . 116

4.6 Additioanl simulation studies for robustness checking. SDP: our ICA method with spatially dependent sources; Fast: FastICA; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function 117

4.7 Plot the log GCV losses against different values of ρ, the scale parameter in the kernel function, for the PNC dataset. 120

4.8 ICA results for the PNC data example from three representative ICA methods (RPF: right parietal frontal network; DMN: default mode network; PVC: primary visual cortex; AUD; auditory netork) 121

List of Tables

2.1 Simulation results for comparing our hc-ICA method against dual-regression ICA based on 100 runs. Values presented are mean and standard deviation of correlations between the true and estimated: subject-specific spatial maps, population-level spatial maps and subject-specific time courses. The mean and standard deviation of the MSE of the covariate estimates are also provided. 40

2.2 Simulation results for comparing the subspace-based approximate EM and the exact EM based on 50 runs. Mean and standard deviation of correlations between the true and estimated spatial maps and time courses are presented. The mean and standard deviation of the MSE of the covariate estimates are also provided. 41

2.3 Simulation results for the inference of $\beta(v)$ based on 1000 runs. Type-I errors are averaged across all voxels with $\beta_{k\ell}(v) = 0$; powers are averaged across voxels having the same values of $\beta_{k\ell}(v) \neq 0$. 43

2.4 Simulation results for comparing our hc-ICA method (approximate EM, $q = 10$) against the dual-regression ICA based on 100 runs. Values presented are mean and standard deviation of correlations between the true and estimated: subject-specific spatial maps, population-level spatial maps and subject-specific time courses. The mean and standard deviation of the MSE of the covariate estimates are also provided. 57

2.5 Checking the stability of EM algorithm in real data using the resulting correlations between pairs form 50 different initialization (IQR: interquantile range; Q_1: the 25th percentile; Q_3: the 75th percentile) 58

3.1 Summary of the true SVCFs and the TGP prior 65

3.2 Quantitative comparison of SVCM-TGP to voxel-wise GLM fitting results with various thresholdings. All results reported are the means and standard errors based on 50 independently simulated datasets with $e_j(\mathbf{s}) \sim N(0, 1)$. 78

3.3 Quantitative comparison of SVCM-TGP to voxel-wise GLM fitting results with various thresholdings. All results reported are the means and standard errors based on 50 independently simulated datasets with $n = 2500$, three alternative distributions. 80

4.1 Summary of the data generating processes for $\mathbf{s}(\boldsymbol{v})$ in the simulation study: The ean functions referred to here are all displayed in Figure 4.1; the nontaion DE(λ) stands for a double exponential distribution with rate parameter λ; the notation Gamma(k, θ) represents a gamma distribution with k being the shape parameter and θ being the scale parameter. σ parameter here is used to control the noise levels. 109

Chapter 1

Introduction

1.1 Overview

Recent advancements in biomedical imaging technologies have provided abundant information and extensive resources for researchers to study the human brain and neurological diseases. A variety of imaging modalities, such as magnetic resonance imaging (MRI), diffusion tensor imaging (DTI), functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have been developed to measure brain structures and functions from different perspectives, generating various large-scale spatially distributed measurements over a three dimensional (3D) space of the human brain. The data acquired from these measurements, namely neuroimaging data, shares come common characteristics: high dimensionality, complex patio-temporal dependence, delicate edge effects, local discontinuity and low signal-to-noise ratio (SNR). These properties can make traditional statistical analysis inappropriate to handle the neuroimaging data. Therefore, this dissertation is dedicated to develop novel statistical methods that can reasonably resolve these difficulties from certain aspects.

1.2 Some current research topics

1.2.1 Functional brain connectivity

In the past decade, the field of neuroimaging has been moving towards a network-oriented view of brain functions. Functional magnetic resonance imaging (fMRI) is one of the most commonly used technologies to investigate brain functional networks (BFNs). In fMRI studies of functional connectivity, observed data are viewed as mixtures of signals generated from various BFNs. Each of these networks consists of a set of spatially distributed but functionally linked brain regions that present similar blood oxygenation level dependent (BOLD) signals during the scanning sessions. One of the major goals in fMRI data analysis is to decompose the observed fMRI data to identify the underlying functional networks and characterize their spatial distributions and temporal dynamics. Independent component analysis (ICA) has become the most widely used tool in the neuroscience community to investigate these functional networks. As a special case of blind source separation, ICA can separate observed fMRI signals into linear combinations of latent spatial source signals that are statistically as independent as possible. Each of these latent components correspond to a BFN.

Compared with alternative network methods, ICA has several major advantages. As a multivariate approach, ICA can jointly model the relationships among multiple voxels and hence provide a tool for investigating whole brain connectivity. Unlike second-order statistical methods such as PCA, ICA takes into account higher-order statistics and the spatial statistical independence assumption of ICA is well-supported by the sparse nature in typical fMRI activation patterns (Calhoun et al., 2001; Beckmann and Smith, 2004). Furthermore, ICA is a fully data-driven approach that doesn't need *a priori* temporal or spatial models. This makes ICA an important tool for analyzing resting-state fMRI where there is no experimental paradigm (Beckmann et al., 2005). Finally, compared

with other whole brain connectivity methods such as graphical models,a distinctive feature of ICA is that it can partition the whole brain into functionally coherent networks.

In recent years, neuroscience literature has provided evidence that BFNs can vary considerably due to subjects' clinical, biological and demographic characteristics. For example, neuroimaging studies have shown that neural activity and connectivity in specific functional networks are significantly associated with mental disorders and their responses to treatment regimes (Anand et al., 2005; Greicius et al., 2007; Chen et al., 2007; Sheline et al., 2009). Other studies have found activity patterns in major functional networks vary with demographic factors including age and gender (Quiton and Greenspan, 2007; Cole et al., 2010). These findings call for statistical methods that can quantify the effects of subjects' characteristics on the BFNs and can evaluate the differences in BFNs between subject groups (e.g., diseased v.s. normal).

The data example in my first chapter demonstrates the need for incorporating covariate effects when investigating differences in BFNs using group ICA. The data come from a post-traumatic stress disorder (PTSD) study conducted by the Grady Memorial Hospital and Emory University in Atlanta. This PTSD study is one of the largest NIH sponsored ongoing research projects on PTSD in urban population. In our data example, a subgroup of African-American female subjects from the Grady PTSD study were recruited for fMRI acquisitions. One of the main goals of the fMRI study is to investigate PTSD-related differences in BFNs. A major challenge in achieving this goal is that the Grady PTSD study is an observational study in which the PTSD positive and PTSD negative groups were not matched on their demographic or clinical variables. Therefore, between-group comparisons are prone to be biased due to potential confounding factors. For example, it is well-known that PTSD is often comorbid with other mental problems such as major depression disorders (MDD) (Kessler et al., 1995; Campbell et al., 2007). Another example is

that the heterogeneity of age distribution between the two PTSD groups can affect the BFNs, according to findings in Bullmore and Sporns (2009). Thus, to assess PTSD-related brain network alterations, it is necessary to adjust for these potential confounding factors.

Existing group ICA methods, which often assumes the same spatial patterns of BFNs across subjects, do not directly incorporate covariate information in the ICA decomposition. Currently, differences in brain functional networks and their associations with subjects' covariates are assessed through two kinds of heuristic approaches. The first approach is through conducting single-subject ICA separately on each subject's, selecting matched ICs and then performing group analysis on the selected subject-level IC maps (Greicius et al., 2007). A major problem with this approach is that it is often challenging to match ICs across subjects since ICA results are only identifiable up to a permutation of the ICs. Furthermore, since most ICA algorithms are stochastic (Himberg et al., 2004), the levels of ICs extracted from separate ICA runs for different subjects are often not comparable to each other. The second approach is via two-stage analysis based on TC-GICA. Two representative methods in this category are the back-construction (Calhoun et al., 2001) and dual-regression (Beckmann et al., 2009; Filippini et al., 2009). These methods first perform TC-GICA to extract common IC maps at the group level and then reconstruct subject-specific IC maps by post-ICA steps. The covariate effects are evaluated via secondary hypothesis testing or regression analysis on the reconstructed subject-specific maps. These methods do not take into account the random variabilities introduced in reconstructing subject-specific IC maps, which could lead to loss of accuracy and efficiency in estimating and testing covariate effects on functional networks.

In the first topic, we propose a hierarchical covariate-adjusted ICA (hc-ICA) model that directly incorporates covariate effects in group ICA decomposition to investigate differences in BFNs. The hc-ICA model decomposes each subject's fMRI data into linear mixtures of subject-

specific spatial source signals (ICs). These distinct subject-specific ICs are then modeled in terms of population-level baseline source signals, covariate effects and between-subject random variabilities. To the best of our knowledge, hc-ICA is the first model-based group ICA method that captures variabilities in BFNs due to covariates effects. Compared with existing group ICA methods, hc-ICA has several advantages. hc-ICA is more accurate and powerful in terms of detecting brain network differences due to the primary effects of interest, such as disease status, while controlling for other confounding factors. For example, application of hc-ICA to the Grady PTSD study reveals important differences in the brain networks of the two PTSD groups, while the existing group ICA method cannot detect these differences effectively. Results from our simulation studies also corroborate that hc-ICA has better performance than the existing method in terms of both estimation accuracy and statistical power. In addition, hc-ICA can provide model-based estimation or prediction of brain functional networks for subpopulations defined by specific clinical or demographic characteristics. This will promote understandings of both commonalities and distinctions in brain networks across various subgroups within a study cohort.

Our hc-ICA model is developed under the hierarchical probabilistic ICA modeling framework first introduced in Guo and Tang (2013) which proposed a hierarchal random effects ICA model for relaxing the spatial homogeneity assumption in TC-GICA. The hc-ICA model, as well as its estimation and inference procedures, provides several important contributions to hierarchical ICA modeling. First, hc-ICA provides the first statistical framework to evaluate how subjects' demographic and clinical characteristics can affect their brain functional networks. This is not available in any existing group ICA methods including the random effect model in Guo and Tang (2013). Second, we propose a novel subspace-based approximate expectation-maximization (EM) algorithm for obtaining maximum likelihood estimates. The approximate EM algorithm scales

linearly with the number of ICs, which is significantly faster than the exponential growth of the exact EM algorithms used by Guo (2011) and Guo and Tang (2013). Third, our work provides an efficient voxel-wise approximate inference procedure for testing covariate effects on ICs. Such statistical inference procedures are not available in existing group ICA methods including Guo and Tang (2013).

1.2.2 Activation study and feature selection

Recent advancements in biomedical imaging technologies have provided abundant information and extensive resources for researchers to learn the human brain and neurological diseases. A variety of imaging modalities, such as magnetic resonance imaging (MRI), diffusion tensor imaging (DTI) and functional magnetic resonance imaging (fMRI) have been developed to measure brain structures and functions from different perspectives, generating various large-scale spatially distributed measurements over a three dimensional (3D) space of the human brain. We refer to those massive spatial measurements of the brain as neuroimages. This type of data poses both opportunities and challenges for statisticians to develop efficient analytical methods that extract useful features from neuroimages to characterize the association between the brain activities and neurological diseases. To this end, regression analysis, a flexible modeling framework for studying the association among variables, has been investigated and considered as a powerful tool in the analysis of massive neuroimaging data, where neuroimages can be modeled as outcome variables; and the disease status along with the clinical, biological and demographical information can all potentially be predictors.

In the second chapter, we aim to develop a Bayesian feature selection method that can directly select imaging features associated with covariates while integrating the region-wise smoothness features.

For capturing spatial dependence structures in fMRI data analysis, many Bayesian methods have been proposed by using different spatial

priors. For example, Friston and Penny (2003) used a zero-mean multivariate Gaussian distribution as the prior for the GLM regression coefficients across voxels and estimated the its covariance with restricted maximum likelihood. Woolrich et al. (2004) proposed to assign the Markov random field prior for the autocorrelation parameters of fMRI time series. Penny et al. (2005) considered first-order correlations between neighboring voxels with a covariance matrix known up to a multiplicative constant. Bowman et al. (2008) proposed a spatial Bayesian hierarchical model to include regional effects into the spatial dependence structure of the regression coefficients from MUA. Flandin and Penny (2007) finessed the Bayesian framework by allowing for variations in spatial smoothness using sparse spatial basis functions, which is similar to the well known wavelet shrinkage method in spirit. All these methods, however, mainly focus on modeling the spatial-temporal structure of fMRI data and do not consider the problem of feature selection. Thus, these methods are not suitable for achieving our goal of finding neuroimaging features that are strongly associated with covariates.

To select important features in the analysis of neuroimaging data while incorporating spatial smoothness and jumps, we develop a new family of Bayesian nonparametric priors based on the Gaussian processes (GPs) under the SVCM framework. Some recent works on Bayesian nonparametric priors constructed from GPs include Kim et al. (2005); Gramacy and Lee (2008); Fox and Dunson (2012) and their major goals are modeling non-stationarity, abrupt spatial changes or long-range dependence. Our proposed prior is constructed by combining and thresholding two Gaussian processes: a global GP that accounts for the entire domain spatial dependence and a local GP which accommodates the regional fluctuations. We refer to it as the thresholded Gaussian process (TGP) prior. This construction of TGP can characterize important common features of the neuroimaging data, including sparsity, global spatial dependence, region-wise smoothness and jump discontinuities. The proposed TGP

prior enjoys the large support property within the functional space of SVCF, leading to the posterior consistency in SVCF estimation. More interestingly, we can prove the posterior consistency in feature selection of SVCF. We also develop efficient MCMC posterior computation algorithms based on a kernel convolution approach. A special choice of the kernel function enables the computation scalable to an ultra-high dimensional case.

1.2.3 Spatial dependence when performing ICA dempoition

Independent component analysis (ICA) refers to a family of algorithms for blind source separation. ICA has been applied to many fields including biomedical imaging, telecommunication and signal processing for exploratory data analysis, feature extraction and compression. For a comprehensive exposition of ICA, see Hyvärinen et al. (2001) as a good reference. In general, ICA algorithms aim to recover statistically independent source signals, or independent components (ICs), from their linear mixtures without resorting to *a priori* knowledges. Under the canonical ICA setting, one observes $q(\geqslant 2)$ synchronously measured records at n distinct points, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, in a compact set $\mathcal{V} \subset \mathbb{R}^d$. When $d = 1$, for example, the outcomes are considered as recordings at different time points. Otherwise, if $d = 2$ or 3, the records are collected at different pixels or voxels on 2D or 3D images. Denote by $y_j(\boldsymbol{v}_i)$ the jth ($j = 1, \ldots, q$) record at point $\boldsymbol{v}_i$, ICA model assumes that

$$\mathbf{y}(\boldsymbol{v}_i) = \mathbf{A}\mathbf{s}(\boldsymbol{v}_i), \quad \boldsymbol{v}_i \in \mathcal{V},\ i = 1, \ldots, n, \tag{1.1}$$

in which $\mathbf{y}(\boldsymbol{v}_i) = [y_1(\boldsymbol{v}_i), \ldots, y_q(\boldsymbol{v}_i)]^\top$ is a vector containing all the observed data recorded at the point indexed by coordinate $\boldsymbol{v}_i$; the so-called mixing matrix, $\mathbf{A}$, is a $q \times q$ matrix that mixes the q mutually independent channels, $\mathbf{s}(\boldsymbol{v}_i) = [s_1(\boldsymbol{v}_i), \ldots, s_q(\boldsymbol{v}_i)]^\top$, to generate the observed data vector. Given the ICA model (1.1), ICA algorithms aim to recover the mutually

independent latent sources $s_1(\cdot), \ldots, s_q(\cdot)$ by estimating the mixing matrix $\mathbf{A}$, or equivalently, the unmixing matrix $\mathbf{W}$ such that $\mathbf{WA} = \mathbf{I}$, from the observed multivariate outcomes $\mathbf{y}(\cdot)$.

In recent years, ICA has become a popular tool for analyzing functional magnetic resonance imaging (fMRI) data (Mckeown et al., 1998; Biswal and Ulmer, 1999; Calhoun et al., 2001; Beckmann and Smith, 2005). When applied to fMRI studies, one major disadvantage of most existing ICA methods is that their iid source signal assumption does not reflect the complex spatio-temporal dependence structure in fMRI data. To address this issue, Lee et al. (2011) drop the iid assumption in model (1.2) and propose an ICA model with autoregressive IC sources for $\boldsymbol{\nu} \in \mathbb{Z}^1$. Their model can extract mutually independent time courses (temporal ICA) from fMRI data featuring temporal autocorrelations. However, in many fMRI studies, people are interested in finding mutually independent source signals in the spatial domain (spatial ICA), which represent functionally connected networks in human brains (Mckeown et al., 1998). However, Lee et al. (2011)'s method does not apply to this setting directly. At the same time, how to characterize the spatial dependence structures of fMRI data when performing ICA decomposition remains an unresolved issue.

In the third project, we propose a new spatial ICA model for fMRI data analysis, which features spatial dependence within each IC source channel. Spatial dependence is an important feature of fMRI data. Statistical correlations across voxels can be interpreted as functional connectivities. The most common dependence structure in fMRI data is the neighboring effect: When a voxel is functionally activated, voxels that are closer to it are more likely to be activated due to functional homogeneities of nearby brain regions (Katanoda et al., 2002; Woolrich et al., 2004). In addition, spatial smoothing, as a routinely used pre-processing step of fMRI data in order to filter low pass signals and reduce between-subject functional variations, strengthens the spatial correlations between adja-

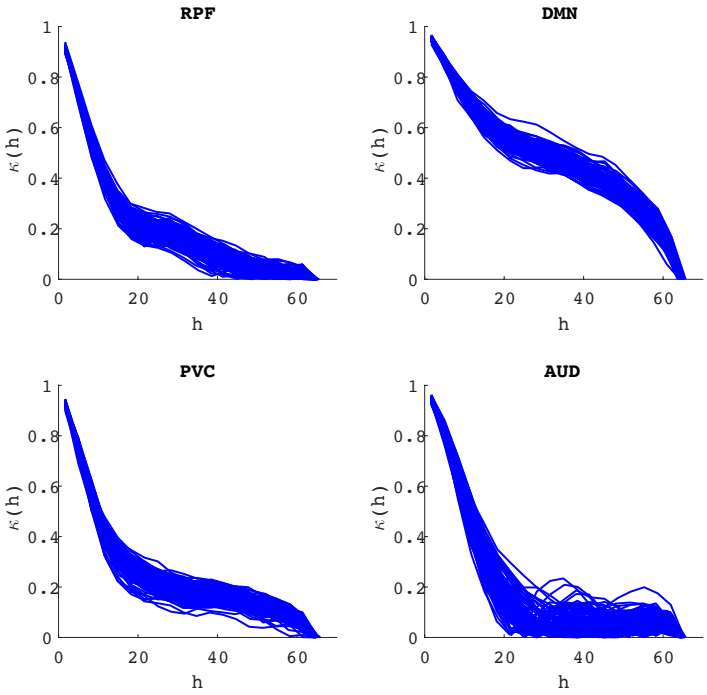


Figure 1.1: The empirical (stationary) spatial correlation functions, $\kappa(\cdot)$, of four esitmated brain functional networks based on our resting-state fMRI data example: Each subfigure contains 100 curves; each curve is plotted based on $2,500$ randomly sampled voxels (without replacement) since the total number of voxels is too large to compute the empirical correlation function.

cent voxels. Kernel methods are ideal tools to capture this type of dependence structure. Specifically, we model the spatial dependence in fMRI data using the Bayesian kernel models (Pillai et al., 2007; Wolpert et al., 2011), enabling flexible prior dependence structure specification. At the same time, some brain regions that are far away from each other can form brain functional networks and exhibit similar activation patterns (Mckeown et al., 1998; Biswal and Ulmer, 1999). This type of dependence structure is difficult to model appropriately *ex-ante*. However, spatial ICA for fMRI data has been proved to be a very useful tool to recover the brain functional networks by “clustering" co-activated brain regions within the same ICs (Mckeown et al., 1998; Calhoun et al., 2001). Thus, in our spatially dependent ICA model, we only include prior knowledge about the neighborhood functional similarities and let the ICA model itself to ex-

tract information about the correlations induced by brain functional networks.

Our work contributes to existing literature from the following four aspects. First, to the best of our knowledge, our model is the first one that can perform spatial ICA decomposition of fMRI data while incorporating spatial dependence structures in the brain. Second, we demonstrate through simulation studies that by incorporating spatial dependence structures, our model can provide more accurate mixing/unmixing matrix estimation than existing ICA algorithms which rely on the iid assumption about the ICs. Third, we can easily make model-based posterior inference about the spatial source signals in order to identify functionally activated brain regions. This is because that we take a fully Bayesian approach for model inference and approximate the joint posterior using samples from Markov chain Monte Carlo (MCMC). Working with the posterior MCMC samples, we can compute credible intervals for the mean processes of the ICs and select functionally activated brain regions whose credible intervals exclude zero. Forth, we establish a new paradigm for ICA modeling beyond the traditional framework based on density estimations. Our treatment of IC source signals can be regarded as nonparametric regressions with location coordinates being the independent variable. Compared with those density-based approaches, the new paradigm can easily incorporate additional dependence structures or association patterns by adding covariates into the kernel function. Compared with the autoregressive model by Lee et al. (2011), our nonparametric model is not restricted to the integer grid support and is potentially more flexible, especially for modeling non-stationary source signals. By separating conditional mean responses from regression residual terms, our method has improved power of detecting activation areas in the brain functional networks.

1.3 Literature Review

ICA and Group ICA

ICA was initially used for analyzing single-subject fMRI data to either characterize spatially independent brain networks, i.e., spatial ICA (Mckeown et al., 1998; Biswal and Ulmer, 1999; Calhoun et al., 2001; Beckmann and Smith, 2005) or separate independent time courses, i.e., temporal ICA (Lee et al., 2011). In this chapter, we consider spatial ICA which is more suitable for our fMRI data example. Denote by $\mathbf{Y}$ the $T \times V$ fMRI data matrix for one subject, where T is the number of fMRI scans and V is the number of voxels in the 3D brain image acquired during each scan. Each row of $\mathbf{Y}$ represents a vectorized 3D image. Classical noise-free spatial ICA decomposes the observed fMRI data for one subject as $\mathbf{Y}_{T\times V} = \mathbf{A}_{T\times q}\mathbf{S}_{q\times V}$, where q is the total number of source signals. Each row of $\mathbf{S}$ represents a vectorized 3D image of a spatial source signal. The q spatial source signals are assumed to be statistically independent and hence are called independent components (ICs). $\mathbf{A}$ is the mixing matrix, the columns of which determine the temporal dynamics of the ICs.

To decompose multi-subject fMRI data, ICA has been extended for group analysis, which is referred to as group ICA (Calhoun et al., 2001). One commonly used group ICA framework in fMRI analysis is the temporal concatenation group ICA (TC-GICA). In TC-GICA, the $T \times V$ fMRI data matrices from N subjects are stacked in the temporal domain to form a tall $TN \times V$ group data matrix. The concatenated group data are then decomposed into the product of a $TN \times q$ group mixing matrix and a $q \times V$ spatial source matrix with independent rows. Many existing group ICA methods (Calhoun et al., 2001; Beckmann and Smith, 2005; Guo and Pagnoni, 2008; Guo, 2011) were developed under the TC-GICA framework. A notable restriction of the TC-GICA models is that they assume the same spatial distribution of BFNs across subjects.

Spatial Regression Models for Neuroimaging Data

A pioneer work using the regression model for the neuroimaging data is the mass univariate analysis (MUA). This approach fits a general linear model (GLM) at each spatial location in the brain (to which is referred as a voxel) and obtains massive test statistics over space to identify voxels/regions that are significantly associated with a specific covariate, which requires multiple comparisons correction. One standard procedure is to calculate the family-wise error rate (FWER) based on the random field theory for statistical parametric maps (Friston et al., 1995; Nichols and Hayasaka, 2003). Another approach is to control the false discovery rate (FDR) using the observed p-values (Benjamini and Yekutieli, 2001; Genovese et al., 2002). A major drawback of MUA is that the models do not borrow information from the spatial dependence across brain locations. In practice, the neuroimaging data are usually pre-processed by a spatial smoothing procedure using a kernel convolution approach. Performing MUA on these pre-smoothed data may lead to inaccuracy and low efficiency in terms of estimating and testing the covariate effects (Chumbley et al., 2009; Li et al., 2011). Recent development in adaptive smoothing methods for preprocessing (Yue et al., 2010) and estimation (Polzehl and Spokoiny, 2000; Qiu, 2007; Tabelow et al., 2008; Li et al., 2011; Wang et al., 2013) may improve the performance in terms of reducing noise and preserving features. It is especially powerful to detect delicate features such as jump discontinuities, which is one of the universal characteristics for neuroimaging data.

To achieve a goal similar to MUA when analyzing neuroimaging data, Zhu et al. (2014) recently developed a systematic modeling approach using a novel spatially varying coefficient model (SVCM) which incorporates both spatial smoothness and jump discontinuity in covariate effects. General SVCMs have been extensively investigated for different applications in environmental heath, epidemiology and ecology as demon-

strated by Cressie and Cassie (1993); Diggle et al. (1998); Gelfand et al. (2003). The SVCM encompasses a wide range of regression models with the outcome variable observed over space and the regression coefficients modeled as functions varying spatially. We refer to this type regression coefficients as spatially varying coefficient functions (SVCFs). SVCFs are commonly assumed to be smooth functions or ρ times continuously differentiable functions with $\rho \geqslant 1$ (we will not make this distinction throughout the rest of this chapter unless noted). Zhu et al. (2014) extended the general SVCMs by introducing jump discontinuities into the SVCFs, making the model especially useful for neuroimaging data analysis. Based on stepwise estimating procedures and asymptotic Wald tests, Zhu et al. (2014)'s SVCM also can identify brain regions that are significantly associated with the given covariates, although it is not developed particularly for feature selection.

Bayesian Variable/Feature/Model Selection Methods

For variable selection in regression models, regularization methods have been studied extensively (Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Yuan and Lin, 2006). Bayesian methods have also been developed based on various prior specifications. Mitchell and Beauchamp (1988) developed a prior model for linear model coefficients using the mixture of a uniform distribution (slab) and a point mass at zero (spike), which is broadly referred to as the spike-and-slab type of priors. George and McCulloch (1993) proposed to use the scale mixture of two zero-mean Gaussian distributions and developed posterior computation algorithm based on Gibbs sampling. Relative works also include Liang et al. (2008); Park and Casella (2008); Bondell and Reich (2012); Johnson and Rossell (2012); Narisetty et al. (2014); Bhattacharya et al. (2014). For the analysis of physical activity and environmental health data, Reich et al. (2010) developed an multivariate SVCM along with a Bayesian variable selection procedure to identify important SVCFs, using the spike-and-

slab prior. Their focus, however, was on distinguishing covariate effects that were zero constant, nonzero constant and spatially varying instead of selecting features within the varying coefficient functions. In light of the needs for integrating complex data structure in many applications, recent development of Bayesian variable selection incorporates dependence structures into the prior model. Li and Zhang (2010) assumed that covariates lay on an undirected graph and used the Ising prior to incorporate this information to the model space and applied this method to analyze the genomics data. For the modeling of spatial data, Markov random field (MRF) is one of the commonly used priors for to model dependence structure of the regression coefficients. For instance, Smith and Fahrmeir (2007) applied this type of priors to fMRI data analyses.

ICA Algorithms

Most existing ICA methods rely on the assumption that the ICs are iid draws from probabilistic distributions at recording points $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, i.e.,

$$\mathbf{s}(\boldsymbol{v}_1), ..., \mathbf{s}(\boldsymbol{v}_n) \overset{\text{iid}}{\sim} \prod_{\ell=1}^{q} g_\ell(\cdot), \tag{1.2}$$

where $g_\ell(\cdot)$ is the probability density of $s_\ell(\cdot)$, the ℓth IC channel. By adopting restricted parametric assumptions about the source signal distributions, g_ℓ, $\ell = 1, \ldots, q$, Bell and Sejnowski (1995); Cardoso (1999); Hyvärinen (1999) estimate the unmixing matrix through maximizing mutual information, diagonalizing higher order cumulants or maximizing non-Gaussianity of the ICs. In addition, flexible semiparametric or nonparametric approaches are also developed for ICA estimation and inference (Tibshirani and Hastie, 2002; Bach and Jordan, 2003; Samarov and Tsybakov, 2004; Chen and Bickel, 2005, 2006). While these methods require the existence of IC density functions with certain degrees of continuity, Samworth and Yuan (2012); Ilmonen et al. (2011); Hallin and Mehta (2015) relax this limitation using log-concave density estimation or rank based

methods.

1.4 Outlines

We introduce in Section 2.1 the hc-ICA framework including data preprocessing, model specification, estimation and inference. Section 2.2 presents an analysis of the PTSD dataset using our method. Section 2.3 reports simulation results for comparing hc-ICA to the existing TC-GICA two-stage method, comparing the subspace-based EM to the exact EM algorithms and comparing the proposed inference method to the existing TC-GICA two-stage method for testing covariate effects. Conclusions and discussions are presented in Section 2.4. Derivations, proofs, additional simulation studies and details for the analysis of the PTSD data are provided in the web supplementary materials.

The remaining parts of the second chapter is organized as follows: Section 3.1 introduces the SVCMs for neuroimaging data analysis and particularly discuss conditions on SVCFs in the proposed model. This section also presents the construction of TGP which serves as a prior model for SVCFs. We study the theoretical properties of TGP and the proposed SVCMs in 3.2. We develop an efficient and scalable posterior computation algorithm based on a kernel convolution approach in Section 3.3. We evaluate the performance of proposed method via simulation studies and and analyze the ABIDE data in Section 3.4. We conclude our work with a brief discussion on future work in Section 3.5.

We provide more details about the Bayesian spatially dependent ICA model and its posterior inference procedures in Section ??. Simulation studies are included in Section ?? to make comparison of our method against some existing ICA algorithms. Analysis of a real resting-state fMRI dataset is also reported in Section ??. Section ??.

In the final chapter, we present concluding remarks and discussions about future research.

Chapter 2

Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis

This chapter is joint work with Dr. Ying Guo.

2.1 Methods

This section introduces the hc-ICA framework, which includes the preprocessing step, the hc-ICA model, estimation algorithms and the inference procedure.

2.1.1 Preprocessing prior to ICA

Prior to an ICA algorithm, some preprocessing steps such as centering, dimension reduction and whitening of the observed data are usually per-

formed to facilitate the subsequent ICA decomposition (Hyvärinen et al., 2001). Suppose that the fMRI study consists of N subjects. For each subject, the fMRI signal is acquired at T time points across V voxels. Let $\widetilde{y_i}(v) \in \mathbb{R}^T$ be the centered time series recorded for subject i at voxel v; $\widetilde{Y}_i = [\widetilde{y_i}(1), ..., \widetilde{y_i}(V)]$ is the $T \times V$ fMRI data matrix for subject i.

Under the paradigm of group ICA, we perform the following dimension reduction and whitening procedure on the original fMRI data: for $i = 1, ..., N$,

$$Y_i = (\Lambda_{i,q} - \tilde{\sigma}^2_{i,q} I_q)^{-\frac{1}{2}} U'_{i,q} \widetilde{Y}_i, \tag{2.1}$$

where $U_{i,q}$ and $\Lambda_{i,q}$ contain the first q eigenvectors and eigenvalues based on the singular value decomposition of $\widetilde{Y}_i$. The residual variance, $\tilde{\sigma}^2_{i,q}$, is the average of the smallest $T - q$ eigenvalues that are not included in $\Lambda_{i,q}$ representing the variability in $\widetilde{Y}_i$ that is not accounted by the first q components. The parameter q, which is the number of ICs, can be determined using the Laplace approximation method (Minka, 2000). Throughout the rest of our chapter, we will present the model and methodologies based on the preprocessed data $Y_i = [y_i(1), ..., y_i(V)]$ $(i = 1, ..., N)$, which are $q \times V$ matrices.

2.1.2 A hierarchical covariate-adjusted ICA model (hc-ICA)

In this section, we present a hierarchical covariate-adjusted ICA (hc-ICA) model for evaluating covariate effects on brain functional networks using multi-subject fMRI data. The first-level model of hc-ICA decomposes a subject's observed fMRI signals into a product of subject-specific spatial source signals and a temporal mixing matrix to capture between-subject variabilities in the spatio-temporal processes in the functional networks. We include a noise term in this ICA model to account for residual variabilities in the fMRI data that are not explained by the extracted ICs, which is known as probabilistic ICA (Beckmann and Smith, 2004). To

be specific, the first-level of hc-ICA is defined as,

$$\mathbf{y}_i(v) = \mathbf{A}_i \mathbf{s}_i(v) + \boldsymbol{e}_i(v), \tag{2.2}$$

where $\mathbf{s}_i(v) = [s_{i1}(v), ..., s_{iq}(v)]'$ is a $q \times 1$ vector with $s_{i\ell}(v)$ representing the spatial source signal of the ℓth IC (i.e., brain functional network) at voxel v for subject i. The q elements of $\mathbf{s}_i(v)$ are assumed to be independent and non-Gaussian. $\mathbf{A}_i$ is the $q \times q$ mixing matrix for subject i which mixes $\mathbf{s}_i(v)$ to generate the observed (preprocessed) fMRI data. Since $\mathbf{Y}_i$ is whitened, the mixing matrix, $\mathbf{A}_i$, should be orthogonal (Hyvärinen and Oja, 2000). $\boldsymbol{e}_i(v)$ is a $q \times 1$ vector that represents the noise in the subject's data and $\boldsymbol{e}_i(v) \sim N(\mathbf{0}, \mathbf{E}_v)$ for $v = 1, ..., V$. The noise term is assumed to be independent across voxels because the spatial correlation across voxels is modelled by the spatial source signals (Hyvärinen et al., 2001; Beckmann and Smith, 2004; Guo, 2011). Prior to ICA, preliminary analysis such as prewhitening (Bullmore et al., 1996) can be performed to remove temporal correlations in the noise term and to standardize the variability across voxels. Therefore, following previous work (Beckmann and Smith, 2004, 2005; Guo and Pagnoni, 2008; Guo, 2011), we assume that the covariance for the noise term is the same across voxels and isotropic, i.e. $\mathbf{E}_v = \nu_0^2 \mathbf{I}_q$. The ICA decomposition in the first-level model is a spatial ICA model since statistical independence is assumed for the spatial maps of brain functional networks. For fMRI data, spatial ICA has become dominant because the spatial independence assumption is well suited to the spatial patterns of most cognitive activation paradigms (Mckeown et al., 1998).

At the second-level of hc-ICA, we further model subject-specific spatial source signals $\mathbf{s}_i(v)$ as a combination of the population-level source signals, the covariate effects and additional between-subject random variabilities:

$$\mathbf{s}_i(v) = \mathbf{s}_0(v) + \boldsymbol{\beta}(v)' \mathbf{x}_i + \boldsymbol{\gamma}_i(v), \tag{2.3}$$

where $\mathbf{s}_0(v) = [s_{01}(v), ..., s_{0q}(v)]'$ is the population-level spatial source signals of the q statistically independent and non-Gaussian ICs; $\mathbf{x}_i = [x_{i1}, ..., x_{ip}]'$ is the $p \times 1$ covariate vector containing subject-specific characteristics such as the treatment or disease group, demographic variables and biological traits; $\boldsymbol{\beta}(v)$ is a $p \times q$ matrix where the element $\beta_{k\ell}(v)$ $(k = 1, ...p, \ell = 1, ..., q)$ in $\boldsymbol{\beta}(v)$ captures the effect of the kth covariate on the ℓth functional network at voxel v; $\boldsymbol{\gamma}_i(v)$ is a $q \times 1$ vector reflecting the random variabilities among subjects after adjusting for covariate effects. We assume $\boldsymbol{\gamma}_i(v) \overset{\text{iid}}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{D})$ where $\mathbf{D} = \text{diag}(v_1^2, ..., v_q^2)$. IC-specific variances specified in $\mathbf{D}$ allow us to accommodate different levels of between-subject random variability.

2.1.3 Source signal distribution assumptions

Following Guo (2011); Guo and Tang (2013), we choose mixtures of Gaussians (MoG) as our source distribution model for the population-level spatial source signals, $\mathbf{s}_0(v)$, in (2.3). MoG hae several desirable properties for modeling fMRI signals. Within each BFN, only a small percentage of locations in the brain are activated or deactivated whereas most brain areas exhibit background fluctuations (Biswal and Ulmer, 1999). MoG are well suited to model such mixed patterns. Furthermore, MoG can capture various types of non-Gaussian signals (Xu et al., 1997; Kostantinos, 2000) and also offer tractable likelihood-based estimation (McLachlan and Peel, 2004).

Specifically, for $\ell = 1, \ldots, q$ we assume that

$$s_{0\ell}(v) \sim \text{MoG}(\boldsymbol{\pi}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell^2), \qquad v = 1, ..., V, \tag{2.4}$$

where $\boldsymbol{\pi}_\ell = [\pi_{\ell,1}, ..., \pi_{\ell,m}]'$ with $\sum_{j=1}^{m} \pi_{\ell,j} = 1$, $\boldsymbol{\mu}_\ell = [\mu_{\ell,1}, ..., \mu_{\ell,m}]'$ and $\boldsymbol{\sigma}_\ell^2 = [\sigma_{\ell,1}^2, ..., \sigma_{\ell,m}^2]'$; m is the number of Gaussian components in MoG. The probability density of $\text{MoG}(\boldsymbol{\pi}_\ell, \boldsymbol{\mu}_\ell, \boldsymbol{\sigma}_\ell^2)$ is $\sum_{j=1}^{m} \pi_{\ell,j}\, g(s_{0\ell}(v); \mu_{\ell,j}, \sigma_{\ell,j}^2)$ where $g(\cdot)$ is the pdf of the (multivariate) Gaussian distribution. In fMRI applications, mixtures of two to three Gaussian components are sufficient to

capture the distribution of fMRI spatial signals, with the different Gaussian components representing the background fluctuation and the negative or positive fMRI BOLD effects respectively (Beckmann and Smith, 2004; Guo and Pagnoni, 2008). Without loss of generality, we denote by $j = 1$ the background fluctuation state throughout the rest of the chapter.

To facilitate derivations in models involving MoG, latent state variables are often used (McLachlan and Peel, 2004). Here we define latent states $z(v) = [z_1(v), ..., z_q(v)]'$ at voxel v as follows. For $\ell = 1, ..., q$, $z_\ell(v)$ takes a value in $\{1, \ldots, m\}$ with probability $p[z_\ell(v) = j] = \pi_{\ell,j}$ for $j = 1, .., m$. Conditional on $z(v)$, we can rewrite our source distribution model as,

$$\mathbf{s}_0(v) = \boldsymbol{\mu}_{z(v)} + \boldsymbol{\psi}_{z(v)}, \tag{2.5}$$

where $\boldsymbol{\mu}_{z(v)} = [\mu_{1,z_1(v)}, ..., \mu_{q,z_q(v)}]'$, $\boldsymbol{\psi}_{z(v)} = [\psi_{1,z_1(v)}, ..., \psi_{q,z_q(v)}]'$, $\boldsymbol{\psi}_{z(v)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{z(v)})$ with $\boldsymbol{\Sigma}_{z(v)} = \text{diag}(\sigma^2_{1,z_1(v)}, ..., \sigma^2_{q,z_q(v)})$.

2.1.4 Maximum likelihood estimation

We develop a unified maximum likelihood estimation method via the EM algorithm that simultaneously estimates all parameters in the hc-ICA model. Based on (2.2), (2.3) and (2.5), the complete data log-likelihood for hc-ICA model is

$$l(\Theta; \mathcal{Y}, \mathcal{X}, \mathcal{S}, \mathcal{Z}) = \sum_{v=1}^{V} l_v(\Theta; \mathcal{Y}, \mathcal{X}, \mathcal{S}, \mathcal{Z}), \tag{2.6}$$

where $\mathcal{Y} = \{\mathbf{y}_i(v) : i = 1, ..., N; v = 1, \ldots, V\}$, $\mathcal{X} = \{\mathbf{x}_i : i = 1, ..., N\}$, $\mathcal{S} = \{\mathbf{s}_i(v) : i = 0, ..., N, v = 1, ..., V\}$ and $\mathcal{Z} = \{z(v) : v = 1, ..., V\}$; the parameters are $\Theta = \{\{\boldsymbol{\beta}(v)\}, \{\mathbf{A}_i\}, \mathbf{E}, \mathbf{D}, \{\boldsymbol{\pi}_\ell\}, \{\boldsymbol{\mu}_\ell\}, \{\boldsymbol{\sigma}^2_\ell\} : i = 1, ..., N, v = 1, ..., V, \ell = 1, ..., m\}$. The detailed expression for the complete data log-likelihood function at

each voxel v is:

$$l_v(\Theta; \mathcal{Y}, \mathcal{X}, \mathcal{S}, \mathcal{Z}) = \sum_{i=1}^{N} \left[\log g\left(\mathbf{y}_i(v); \mathbf{A}_i \mathbf{s}_i(v), \mathbf{E}\right) + \log g\left(\mathbf{s}_i(v); \mathbf{s}_0(v) + \boldsymbol{\beta}(v)' \mathbf{x}_i, \mathbf{D}\right) \right] + \log g\left(\mathbf{s}_0(v); \boldsymbol{\mu}_{z(v)}, \boldsymbol{\Sigma}_{z(v)}\right) + \sum_{\ell=1}^{q} \log \pi_{l, z_l(v)}. \quad (2.7)$$

The exact EM algorithm

We first present an exact EM which has an explicit E-step and M-step to obtain ML estimates for the parameters in hc-ICA.

E-step: In the E-step, given the parameter estimates $\hat{\Theta}^{(k)}$ from the last step, we derive the conditional expectation of the complete data log-likelihood given the observed data as follows:

$$Q(\Theta|\hat{\Theta}^{(k)}) = \sum_{v=1}^{V} E_{\mathbf{s}(v), z(v)|\mathbf{y}(v)} \left[l_v(\Theta; \mathcal{Y}, \mathcal{X}, \mathcal{S}, \mathcal{Z}) \right], \quad (2.8)$$

where $\mathbf{y}(v) = [\mathbf{y}_1(v)', ..., \mathbf{y}_N(v)']'$ represents the group data vector from the N subjects at voxel v, $\mathbf{s}(v) = [\mathbf{s}_1(v)', ..., \mathbf{s}_N(v)', \mathbf{s}_0(v)']'$ is the vector containing latent source signals on both the population and individual level. The detailed definition of $Q(\Theta|\hat{\Theta}^{(k)})$ is available in section 1 of web supplementary materials. The evaluation of $Q(\Theta|\hat{\Theta}^{(k)})$ relies on $p\left[\mathbf{s}(v), z(v) \mid \mathbf{y}(v); \hat{\Theta}^{(k)}\right]$ as well as its marginal distributions, which consists of the following three steps. First, we determine $p\left[\mathbf{s}(v) \mid z(v), \mathbf{y}(v); \hat{\Theta}^{(k)}\right]$, which is a multivariate Gaussian distribution. Second, we evaluate the probability mass functions, $p\left[z(v) \mid \mathbf{y}(v); \hat{\Theta}^{(k)}\right]$ through an application of Bayes' Theorem. We finally obtain $p\left[\mathbf{s}(v) \mid \mathbf{y}(v); \hat{\Theta}^{(k)}\right]$ by convolving the distributions derived in the previous two steps. More details can be found in section 2 of the supplementary material.

Given these probability distributions, we can derive the analytical forms for the conditional expectation in (2.8). For illustration purposes, two

main quantities of interest in (2.8) are given as follows:

$$\begin{aligned}
E[\mathbf{s}(v) \mid \mathbf{y}(v); \Theta] &= \sum_{z(v)\in\mathcal{R}} p[z(v) \mid \mathbf{y}(v); \Theta] E[\mathbf{s}(v) \mid \mathbf{y}(v), z(v); \Theta], \\
E[\mathbf{s}(v)^{\otimes 2} \mid \mathbf{y}(v); \Theta] &= \sum_{z(v)\in\mathcal{R}} p[z(v) \mid \mathbf{y}(v); \Theta] E[\mathbf{s}(v) \mid \mathbf{y}(v), z(v); \Theta]^{\otimes 2} \\
&\quad + \sum_{z(v)\in\mathcal{R}} p[z(v) \mid \mathbf{y}(v); \Theta] \mathrm{Var}[\mathbf{s}(v) \mid \mathbf{y}(v), z(v); \Theta],
\end{aligned}$$

where $\mathcal{R}$ represents the set of all possible values of $z(v)$, i.e., $\mathcal{R} = \{z^r\}_{r=1}^{m^q}$ where $z^r = [z_1^r, ..., z_q^r]'$ and $z_\ell^r \in \{1, ..., m\}$ for $\ell = 1, ...q$; the notation $a^{\otimes 2}$ for a vector a stands for aa'.

Based on the results presented above, our E-step is fully tractable without the need for iterative numerical integrations.

M-step: In the M-step, we update the current parameters estimates $\hat{\Theta}^{(k)}$ to

$$\hat{\Theta}^{(k+1)} = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta|\hat{\Theta}^{(k)}). \tag{2.9}$$

We have derived explicit formulas for all parameter updates. The updating rules are provided in section 3 of our supplementary material.

The estimation procedure for the exact EM algorithm is summarized in Algorithm 1. See section 1-3 of the supplementary material for more details. After obtaining $\hat{\Theta}$, we can estimate the population- and individual-level source signals as well as their variability based on the mean and variance of their conditional distributions, i.e., $[\mathbf{s}_0(v) \mid \mathbf{y}(v); \hat{\Theta}]$ and $[\mathbf{s}_i(v) \mid \mathbf{y}(v); \hat{\Theta}]$. These conditional moments are directly obtainable from the E-step of our algorithm upon convergence and no separate post-ICA steps are required. As a referee pointed out, one could also estimate the source signals using the MAP estimator. As a major difference from TC-GICA, our subject-specific ICs $\{\mathbf{s}_i\}$ are estimated simultaneously with population-level IC $\mathbf{s}_0$ instead of being reconstructed via post-ICA *ad-hoc* approaches. Therefore, all the subject ICs are aligned to the population ICs in our model specification and estimation, which eliminates the need to match

ICs across difference subjects. This is an advantage of our approach over single-subject ICA based analysis.

In fMRI analysis, researchers are often interested in thresholded IC maps to identify “significantly activated" voxels in each BFN. Following previous work (Guo, 2011), we propose a thresholding method based on the mixture distributions for this purpose (section 6 of the supplementary material).

Algorithm 1 The Exact EM Algorithm

Initial values: Start with initial values $\hat{\Theta}^{(0)}$ which can be obtained based on estimates from existing group ICA software.
repeat
E-step:
1. Determine $p[\mathbf{s}(v), \mathbf{z}(v) \mid \mathbf{y}(v); \hat{\Theta}^{(k)}]$ and its marginals using the proposed three-step approach:
1.a Evaluate the multivariate Gaussian $p[\mathbf{s}(v) \mid \mathbf{y}(v), \mathbf{z}(v); \hat{\Theta}^{(k)}]$;
1.b Evaluate $p[\mathbf{z}(v) \mid \mathbf{y}(v); \hat{\Theta}^{(k)}]$;
1.c $p[\mathbf{s}(v), \mathbf{z}(v) \mid \mathbf{y}(v), \hat{\Theta}^{(k)}] = p[\mathbf{s}(v) \mid \mathbf{y}(v), \mathbf{z}(v); \hat{\Theta}^{(k)}] \times p[\mathbf{z}(v) \mid \mathbf{y}(v); \hat{\Theta}^{(k)}]$;
$p[\mathbf{s}(v) \mid \mathbf{y}(v), \hat{\Theta}^{(k)}] = \sum_{\mathbf{z}(v) \in \mathcal{R}} p[\mathbf{s}(v), \mathbf{z}(v) \mid \mathbf{y}(v), \hat{\Theta}^{(k)}]$;
2. Evaluate conditional expectations in $Q(\Theta|\hat{\Theta}^{(k)})$.
M-step:
Update $\boldsymbol{\beta}(v)$, $\mathbf{A}_i$, $\pi_{\ell,j}$, $\mu_{\ell,j}$, $\sigma^2_{\ell,j}$;
Update the variance parameters $\mathbf{D}, \mathbf{E}$.
until $\frac{\|\hat{\Theta}^{(k+1)} - \hat{\Theta}^{(k)}\|}{\|\hat{\Theta}^{(k)}\|} < \epsilon$

The subspace-based approximate EM algorithm

One major limitation of the exact EM algorithm is that its complexity increases exponentially with regard to the number of ICs. Specifically, $\mathcal{O}(m^q)$ operations are required for the exact EM algorithm to complete. The main reason is that, at each voxel, the exact EM evaluates and sums the conditional distributions across the whole sample space $\mathcal{R}$ of the latent state variables $\mathbf{z}(v)$, which has a cardinality of m^q. A standard way to alleviate this issue is through mean field variational approximation. This method has been used by Attias (1999, 2000) for single subject ICA and by Guo (2011) for TC-GICA. However, the variational method cannot be easily generalized to other models such as hierarchical ICA because the

derivation of the variational approximate distributions depends heavily on the model specifications. In most cases, the estimates for the variational parameters do not have analytically tractable expressions and require extra numerical iterations, which sometimes causes convergence problems.

In this section, we propose a new approximate EM algorithm for solving MoG-based ICA models in fMRI studies. Compared with the exact EM that needs $\mathcal{O}(m^q)$ operations, this new EM algorithm only requires $\mathcal{O}(mq)$ operations. The key idea behind the approximate algorithm is that instead of considering the whole sample space $\mathcal{R}$ of the latent state vector $z(v)$, we only focus on a small subspace of $\mathcal{R}$ in the algorithm. Theorem 1 provides the definition for the subspace and shows that under certain conditions, the distribution of the latent state vectors is concentrated to the proposed subspace.

Theorem 1. *Define* $\mathcal{R} = \{z^r = [z_1^r, ..., z_q^r]' : z_\ell^r = j \text{ with } j \in \{1, ..., m\}, \ell = 1, ..., q\}$ *for* $r = 1, ..., m^q$*, which is the domain of* $z(v)$*. For all* $z(v) \in \mathcal{R}$*, suppose that* $p[z_\ell(v) = j] = \pi_{\ell,j}$ *and that* $p[z(v) = z^r] = \prod_{\ell=1}^q \pi_{\ell,z_\ell^r}$ *(i.e.,* $z(v)$ *has independent elements). Define* $\widetilde{\mathcal{R}}$ *as* $\widetilde{\mathcal{R}} = \mathcal{R}_0 \cup \mathcal{R}_1$ *where* $\mathcal{R}_0 = \{z^r \in \mathcal{R} : z_\ell^r = 1, \ell = 1, ..., q\}$ *and* $\mathcal{R}_1 = \{z^r \in \mathcal{R} : \exists$ *one and only one* ℓ, s.t., $z_\ell^r \neq 1\}$*. Then, for any* $0 < \epsilon < 1$*, if* $\pi_{\ell,1} > \frac{q}{q+\sqrt{\epsilon}}$ *for all* $\ell = 1, ..., q$*, we have* $p[z(v) \in \widetilde{\mathcal{R}}] > 1 - \epsilon$.

The proof of the Theorem is relegated to section 4 of the supplementary material. Based on the above theorem, when $\epsilon \approx 0$, i.e. $p[z_\ell(v) = 1] \approx 1$, we have $p[z(v) \in \widetilde{\mathcal{R}}] \approx 1$. For fMRI data, the latent state $j = 1$ in MOG model (2.4) corresponds to background fluctuation. Therefore, Theorem 1 implies that for each IC, if latent states at most voxels are background fluctuation, the probability distribution of the latent state vector $z(v)$ in our hc-ICA will be mostly restricted to the subspace $\widetilde{\mathcal{R}}$. The condition in Theorem 1, i.e. $p[z_\ell(v) = 1] \approx 1$, is supported by fMRI data because previous literature maintains that the fMRI spatial source signals are sparse across the brain (Mckeown et al., 1998; Daubechies et al., 2009). That is,

within a specific BFN, i.e. IC, most of the voxels exhibit background fluctuations with only a very small proportion of voxels being activated (or deactivated). The restriction of the latent states vector to the subspace $\widetilde{\mathcal{R}}$ implies that there is little chance for the same voxel to be activated in more than one ICs. Biologically, this means that there is little overlapping in the activated regions across different BFNs, which has been supported by findings in the existing neuroimaging literature.

Based on this result, we propose a subspace-based approximate EM for our ICA model. The approximate EM follows similar steps as the exact EM. The main difference is that we restrict the conditional distribution of the latent state vector $z(v)$ to the subspace $\widetilde{\mathcal{R}}$ in the E-step and M-step. That is, the conditional expectations in the E-step are evaluated with a subspace-based approximate distribution $\tilde{p}[\mathbf{z}(v) = z^r|\mathbf{y}(v); \hat{\Theta}^{(k)}] = p[\mathbf{z}(v) = z^r|\mathbf{y}(v); \hat{\Theta}^{(k)}]/\sum_{r\in\widetilde{\mathcal{R}}} p[\mathbf{z}(v) = z^r|\mathbf{y}(v); \hat{\Theta}^{(k)}]$ where $z^r \in \widetilde{\mathcal{R}}$ (see section 5 of the supplementary material for a detailed treatment). Since the subspace $\widetilde{\mathcal{R}}$ has a cardinality of $(m-1)q+1$, the approximate EM only requires $\mathcal{O}(mq)$ operations to complete. The concentration of measures to the subspace leads to the simplification in evaluating the conditional expectations in the E-step. For example,

$$\widetilde{E}[\mathbf{s}(v) \mid \mathbf{y}(v); \Theta] = \sum_{z(v)\in\widetilde{\mathcal{R}}} \tilde{p}[\mathbf{z}(v) \mid \mathbf{y}(v); \Theta] E[\mathbf{s}(v) \mid \mathbf{y}(v), \mathbf{z}(v); \Theta], \tag{2.10}$$

which implies that, instead of summing over m^q latent states in $\mathcal{R}$, we only need to perform $(m-1)q+1$ summations across the subspace of $\widetilde{\mathcal{R}}$. The subspace-based EM also reduces computation time in the M-step. Specifically, when updating the parameters for the MoG source distribution model, we now use approximate conditional marginal moments. For example, as compared with the exact results, we use the following approximate moment when updating parameters for the Gaussian mix-

tures,

$$\widetilde{E}[s_{0\ell}(\nu) \mid z_\ell(\nu) = j, \mathbf{y}(\nu); \Theta] = \frac{\sum_{z(\nu)\in\widetilde{\mathcal{R}}^{(\ell,j)}} \tilde{p}[z(\nu) \mid \mathbf{y}(\nu); \Theta] E[s_{0\ell}(\nu) \mid \mathbf{y}(\nu), z(\nu); \Theta]}{\sum_{z(\nu)\in\widetilde{\mathcal{R}}^{(\ell,j)}} \tilde{p}[z(\nu) \mid \mathbf{y}(\nu); \Theta]}, \tag{2.11}$$

where $\widetilde{\mathcal{R}}^{(\ell,j)} = \{z^r \in \widetilde{\mathcal{R}} : z_\ell^r = j\}$, whose cardinality equals $(m-1)(q-1)+1$ if $j = 1$ and 1 if $j \neq 1$. Comparing to its exact counterpart, $\mathcal{R}^{(\ell,j)} = \{z^r \in \mathcal{R} : z_\ell^r = j\}$, which has a cardinality of m^{q-1}, this can dramatically simplify the updating of $\pi_{\ell,j}, \mu_{\ell,j}$ and $\sigma^2_{\ell,j}$ in the M-step. We summarize the approximate EM algorithm as Algorithm 2.

Algorithm 2 The Subspace-based Approximate EM Algorithm

Initial values: Start with initial values $\hat{\Theta}^{(0)}$.
repeat
E-step:
1. Determine $\tilde{p}[\mathbf{s}(\nu) \mid \mathbf{y}(\nu); \hat{\Theta}^{(k)}]$ and its marginals as follows:
1.a Evaluate the multivariate Gaussian $p[\mathbf{s}(\nu) \mid \mathbf{y}(\nu), z(\nu); \hat{\Theta}^{(k)}]$;
1.b Evaluate $\tilde{p}[z(\nu) \mid \mathbf{y}(\nu); \hat{\Theta}^{(k)}]$ on the subset $\tilde{\mathcal{R}}$;
1.c $\tilde{p}[\mathbf{s}(\nu), z(\nu) \mid \mathbf{y}(\nu), \hat{\Theta}^{(k)}] = p[\mathbf{s}(\nu) \mid \mathbf{y}(\nu), z(\nu); \hat{\Theta}^{(k)}] \times \tilde{p}[z(\nu) \mid \mathbf{y}(\nu); \hat{\Theta}^{(k)}]$;
$p[\mathbf{s}(\nu) \mid \mathbf{y}(\nu), \hat{\Theta}^{(k)}] = \sum_{z(\nu)\in\tilde{\mathcal{R}}} \tilde{p}[\mathbf{s}(\nu), z(\nu) \mid \mathbf{y}(\nu), \hat{\Theta}^{(k)}]$;
2. Evaluate conditional expectations in $Q(\Theta|\hat{\Theta}^{(k)})$ with regard to $\tilde{p}[\mathbf{s}(\nu), z(\nu)|\mathbf{y}(\nu); \hat{\Theta}^{(k)}]$.
M-step:
Update $\beta(\nu)$, $A_i, \pi_{\ell,j}$, $\mu_{\ell,j}$; $\sigma^2_{\ell,j}$ with the modification of replacing the exact conditional moments with their counterparts based on $\tilde{p}[\mathbf{s}(\nu) \mid \mathbf{y}(\nu); \hat{\Theta}^{(k)}]$.
Update D, E with similar modifications of replacing the exact conditional moments with those based on $\tilde{p}[\mathbf{s}(\nu) \mid \mathbf{y}(\nu); \hat{\Theta}^{(k)}]$.
until $\frac{\|\hat{\Theta}^{(k+1)} - \hat{\Theta}^{(k)}\|}{\|\hat{\Theta}^{(k)}\|} < \epsilon$

2.1.5 Inference for covariate effects in hc-ICA model

Typically, statistical inference in maximum likelihood estimation is based on the inverse of the information matrix which is used to estimate the asymptotic variance-covariance matrix of the MLEs. Since Standard EM algorithms only provide parameter estimates, extensions to the EM algorithm have been developed to estimate the information matrix (Louis, 1982; Meilijson, 1989; Meng and Rubin, 1991). However, these methods are computationally expensive for the proposed hc-ICA model due to

the following reasons. First, the dimension of the information matrix for our model is huge due to the large number of parameters. Secondly, the ML estimates, $\hat{\boldsymbol{\beta}}(v), v = 1, ..., V$, are not independent across voxels because they rely on the estimates of the same set of parameters such as the mixing matrices. Consequently, the information matrix of the hc-ICA model is ultra-high dimensional and is not sparse, which makes it extremely challenging to invert.

In this section, we present a statistical inference procedure for covariate effects in hc-ICA model. The proposed method is developed based on the connection between the hc-ICA and standard linear models. Our method aims to provide an efficient approach to estimate the asymptotic standard errors of the covariate effects at each voxel, i.e., $\hat{\boldsymbol{\beta}}(v)(v = 1, \ldots, V)$, by directly using the output from our EM algorithms. Specifically, we first rewrite the hc-ICA model in a non-hierarchical form by collapsing the two-level models in (2.2) and (2.3) and then multiplying the orthogonal mixing matrix $\mathbf{A}_i$ on both sides:

$$\mathbf{A}_i' \mathbf{y}_i(v) = \mathbf{s}_0(v) + \mathbf{X}_i \text{vec}\left[\boldsymbol{\beta}(v)'\right] + \boldsymbol{\gamma}_i(v) + \mathbf{A}_i' \boldsymbol{e}_i(v), \tag{2.12}$$

where $\mathbf{X}_i = \mathbf{x}_i' \otimes \mathbf{I}_q$. (2.12) can be re-expressed as follows:

$$\mathbf{y}_i^*(v) = \mathbf{X}_i \text{vec}\left[\boldsymbol{\beta}(v)'\right] + \zeta_i(v), \tag{2.13}$$

where $\mathbf{y}_i^*(v) = \mathbf{A}_i' \mathbf{y}_i(v) - \mathbf{s}_0(v)$, and $\zeta_i(v) = \boldsymbol{\gamma}_i(v) + \mathbf{A}_i' \boldsymbol{e}_i(v)$ is a multivariate zero-mean Gaussian noise term. The model in (2.13) can be viewed as a general multivariate linear model at each voxel. The major distinction of (2.13) from the standard linear model is that the dependent variable $\mathbf{y}^*(v)$ not only depends on the observed data $\mathbf{y}(v)$ but also involves unknown parameters $\mathbf{A}_i$ and latent variables $\mathbf{s}_0(v)$. Given the similarity between hc-ICA and the standard linear model, we propose a variance estimator for vec $\left[\hat{\boldsymbol{\beta}}(v)'\right]$ following the linear model theory.

Note that, for a standard linear model, the asymptotic variance for $\text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]$ can be obtained by:

$$\text{Var}\left\{\text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]\right\} = \frac{1}{N}\left(\sum_{i=1}^{N} X_i' W(v)^{-1} X_i\right)^{-1}, \tag{2.14}$$

where $W(v)$ is the variance of the Gaussian noise in the linear model. Then, the variance of $\text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]$ can be estimated by plugging in an estimator for $W(v)$ in (2.14). Following this result, we consider a variance estimator for $\text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]$ based on (2.14) by plugging in the empirical variance estimator $\widehat{W}(v) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^*(v) - X_i \text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]\right)^{\otimes 2}$ (Seber and Lee, 2012). Because the dependent variable $y^*(v)$ in (2.13) is not directly observable, we estimate $y_i^*(v)$ using the ML estimates from our EM algorithm as $\widehat{y_i^*}(v) = \hat{A}_i' y_i(v) - \widehat{s_0}(v)$, where $\widehat{s_0}(v) = E[s_0(v)|y(v), \hat{\Theta}]$. That is, we modify the empirical variance estimator $\widehat{W}(v)$ as follows:

$$\widetilde{W}(v) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{A}_i' y_i(v) - E[s_0(v)|y(v), \hat{\Theta}] - X_i \text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]\right)^{\otimes 2}. \tag{2.15}$$

Thus, our final variance estimator is

$$\widehat{\text{Var}}\left\{\text{vec}\left[\hat{\boldsymbol{\beta}}(v)'\right]\right\} = \frac{1}{N}\left(\sum_{i=1}^{N} X_i' \widetilde{W}(v)^{-1} X_i\right)^{-1}.$$

Hypothesis testing on the covariate effects at each voxel can be performed by calculating the Z-statistics based on the proposed variance estimator and determine the corresponding p-values. Our method can test whether a certain covariate has significant effects on each of the BFNs at the voxel level. Based on the parametric Z-statistic maps, one can also apply standard multiple testing methods to control the family wise error rate (FWER) or the false discovery rate (FDR) when testing the covariate effects within a BFN.

We note that our variance estimator may underestimate the variabilities in $\hat{\boldsymbol{\beta}}(v)$, because it does not account for variabilities in estimating

$\mathbf{A}_i$ and $s_0(\nu)$. As a result, when performing hypothesis testing, the actual type-I errors of the proposed test statistic can be relatively higher than the nominal level. We evaluate via simulation studies the performance of this proposed inference procedure in Section 4.

2.2 Application to fMRI data from Grady PTSD study

We applied the proposed method to the fMRI data collected from the Grady PTSD study. In this study, 92 African American women were recruited as part of a larger study conducted by the Grady Health System in Atlanta, GA. The Structured Clinical Interview for DSM-IV (SCID) (First, 1995) was administered to all subjects and was used for diagnosis of PTSD. In addition, participants completed the Beck Depression Inventory (BDI) (Beck et al., 1996, 1988) for depression assessment. Out of the 92 subjects, 39 met a diagnosis of PTSD (PTSD+) and 53 did not meet the criteria for PTSD (PTSD−). The ages of these women at the time of study ranged from 20 to 62 (Mean $\pm$ SD: 35 ± 12 for PTSD+ group; 39 ± 12 for PTSD− group; between-group test $p = 0.1096$). The BDI depression scores were significantly higher in subjects with PTSD diagnosis (Mean $\pm$ SD: 16.6 ± 9.0 for PTSD+, 8.3 ± 7.8 for PTSD−, $p < 10^{-5}$).

2.2.1 Experimental design, image acquisition and pre-processing

MRI scans were obtained in a 3.0T Siemens scanner. Participants received task stimuli through an flexible mirror attached to the radiofrequency coil of the scanner. The mirror reflected a computer screen placed at the end of the MRI aperture. During all experiments, a white cross appeared on a black background for 500msec; it was replaced by an X or an O “Go" signal for 1000msec and followed by 750msec of blank screen. On a response pad, the subjects pressed 1 for X and 2 for O. The subjects were instructed to respond to each trial as fast as they could unless the “NoGo" signal appeared (i.e., the background changed to red), in which case they should not press either button. The task comprised four runs separated

by three 20s rest periods. Each run contained 26 “Go", 13 “NoGo", and 14 blank trials distributed randomly.

A T1-weighted high-resolution anatomical image was first acquired (176 sagittal slices, voxel size: $1 \times 1 \times 1$ mm). During task administration, a series of T2-weighted functional images (echo-planar, 26 axial slices, voxel size: $3.75 \times 3.75 \times 4$mm, TR=2.53s, TE=30ms) were acquired. The fMRI data were converted and preprocessed using Statistical Parametric Mapping, version 5 (SPM5, Wellcome Trust Centre for Neuroimaging, London, UK: `http://www.fil.ion.ucl.ac.uk/spm/`). Functional volumes were corrected for slice acquisition timing differences and subject movement. The anatomical image was registered to the mean of the corrected functional images and subsequently spatially normalized to the MNI standard brain space. These normalization parameters from MNI space were used for the functional images, which were smoothed with an 8mm FWHM Gaussian kernel. Prior to ICA analysis, we performed additional preprocessing steps, including centering, dimension reduction and whitening as described in section 2.1.1, on the fMRI data.

2.2.2 Analysis and findings

The preprocessed data from the 92 subjects were decomposed using the proposed hc-ICA model into 16 ICs (the number is chosen from the GIFT package: `http://mialab.mrn.org/software/gift/index.html`). To compare the networks between the two PTSD groups, we included PTSD diagnosis as the primary covariate of interest in the hc-ICA (PTSD$-$=0, PTSD$+$=1). We also included subject’s age and BDI score as covariates to control for potential confounding effects. We estimated the parameters in the hc-ICA model using the subspace-based EM algorithm implemented in MATLAB, which is available at the authors’ website. To ensure the validity of the results, we initialized the EM algorithm with 50 different starting values. The resulting estimates of the parameters were mostly close to each other. In this analysis, we reported the estimates corresponding to

the highest observed data likelihood. More details about this robustness check are included in the supplementary materials.

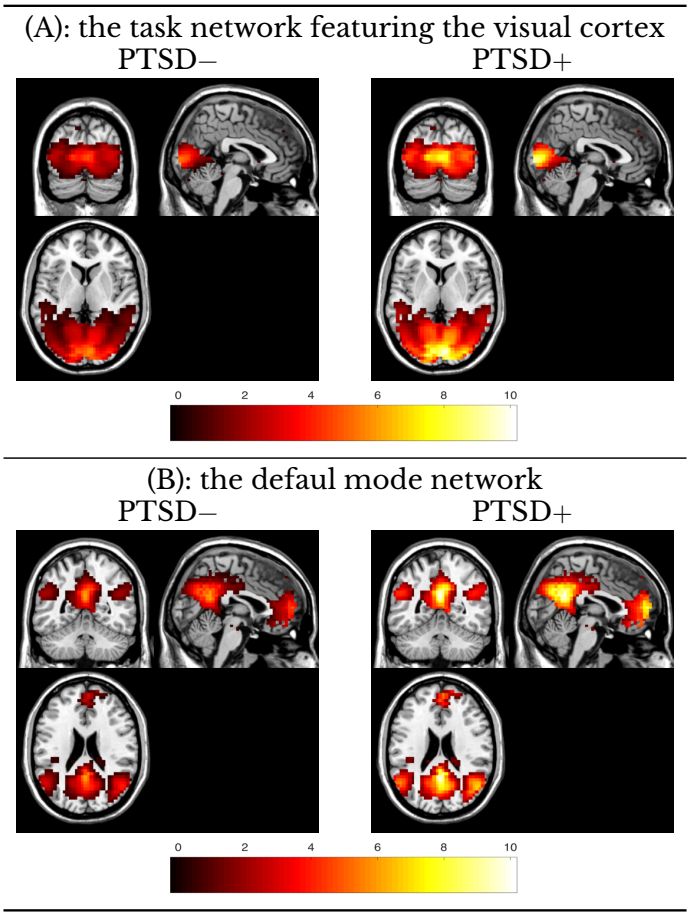


Figure 2.1: The estimated subpopulational maps for the PTSD− and PTSD+ women at the median age (36 year old) and the median depression score (BDI=10): Panel (A) shows the estimates for the network featuring the visual cortex, which has the highest positive correlation with the task time series. Panel (B) shows the estimates for the default mode network, which has the largest negative correlation with the task time series. All IC maps are thresholded at the posterior probability of activation above 0.9. PTSD+ women show stronger IC signals in both networks.

Among the extracted ICs, we identified two components of particular interest. The first network had the highest positive temporal correlation with the task time series, which were the task series convolved with the hemodynamic response function (HRF). The spatial pattern of this network features the visual cortex, which responded to the visual stimuli presented in the Go/NoGo task. In Figure 2.1(A), we present the hc-ICA

model-based estimates of the visual network for both the PTSD− and PTSD+ groups. The two subpopulation maps were estimated at the median age (36 year old) and the median BDI scores (BDI=10) to control for confounding effects. They were all thresholded based on the conditional probability of activation (section 6 of the supplementary material). According to Figure 2.1(A), the PTSD+ group demonstrated stronger spatial source signals in the visual network as compared to the PTSD− group with the same ages and BDI scores. It is worth noting that the existing group ICA methods cannot provide such model-based estimates of the brain networks for subpopulations defined by specific covariate patterns.

The second network of interest mainly includes the posterior cingulate cortex (PCC), the medial prefrontal cortex (mPFC) and the lateral parietal cortex (LPC). This network is known as the “default mode network", which shows increased activities during resting states and decreased activities during cognitive tasks (Raichle et al., 2001). Its temporal responses have the largest negative correlation with the task time series. Figure 2.1(B) presents the hc-ICA model-based estimates of this network for the two PTSD subpopulations (also adjusted at the median age and the median BDI score). Based on Figure 2.1(B), the default mode network of the PTSD+ women demonstrated stronger functional connectivity during the Go/NoGo tasks.

We then applied the proposed inference procedure to formally test the PTSD group differences in these two networks while controlling for the potential confounding effects from age and depression status. We also applied the method in Genovese et al. (2002) to calculate the FDR corrected p-values for the between-group tests. For comparison, we also used a TC-GICA based method, dual-regression ICA (Beckmann et al., 2009; Filippini et al., 2009), to examine the group differences. Dual-regression ICA is one of the most commonly used methods in the neuroimaging community for estimating subject-specific IC maps and performing between-group comparisons, see Smith et al. (2014); Reineberg

et al. (2015) for some examples of its application. It is also adopted as a standard analytical tool by the well-known Human Connectome Project (`http://www.humanconnectomeproject.org/`). The dual-regression procedure typically tests group differences via permutation tests which cannot adjust for any confounding factors. To provide a fair comparison between hc-ICA and dual-regression, we performed an additional regression analysis on the reconstructed subject IC maps from dual regression using the same set of covariates as in hc-ICA and then tested PTSD group differences with adjustment for age and BDI.

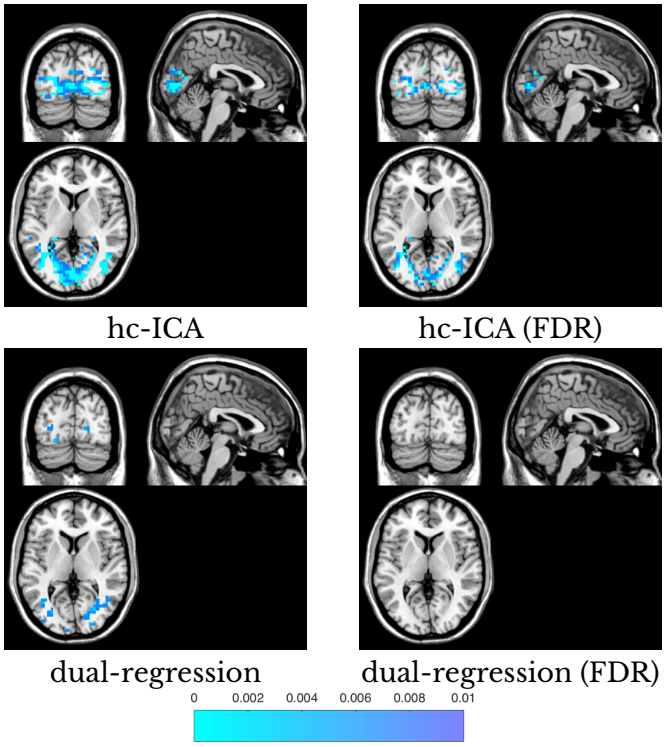


Figure 2.2: p-values, thresholded below 0.01, for comparing the adjusted PTSD group differences (PTSD− < PTSD+) in the task-related network: hc-ICA found increased spatial source signals at the central part of the visual cortex among PTSD+ women, which remained significant after FDR control; dual-regression found much less group differences in the network, all of which became insignificant with the FDR control.

The p-values for testing group differences in the task network, which features the visual cortex, are presented in Figure 2.2. Based on Figure 2.2, hc-ICA detected that PTSD+ women showed significantly stronger

spatial signals than PTSD− women in major parts of the visual network. This finding still held after FDR correction. This enhanced activities in visual cortex among PTSD subjects were previously reported in other fMRI studies involving visual stimuli (Hendler et al., 2003). The dual-regression analysis, however, found only a few differences in this network between the two groups and all of them became insignificant after FDR control.

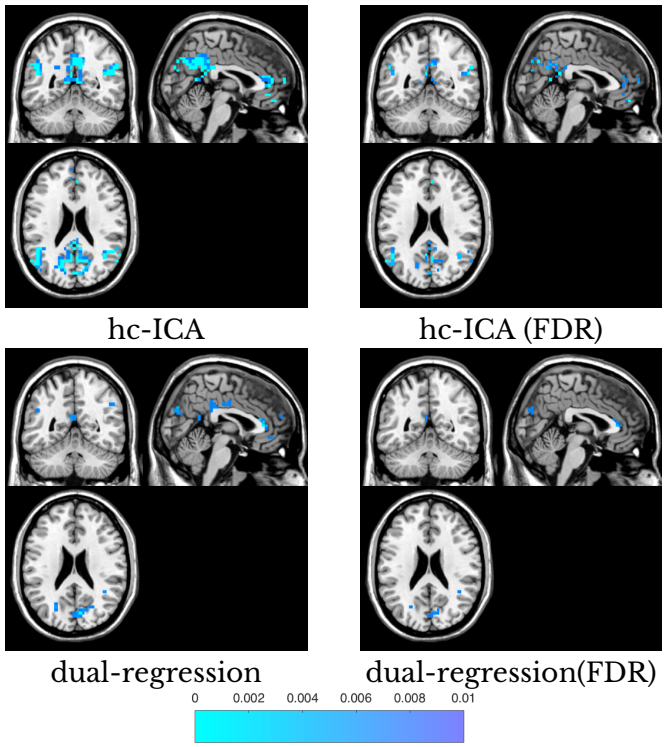


Figure 2.3: p-values, thresholded below 0.01, for comparing the adjusted PTSD group differences (PTSD− < PTSD+) in the default mode network: hc-ICA finds stronger network activities across all the major regions of this network for PTSD+ women. Many of these identified voxels still appear after FDR control; dual-regression findings only dicover a few differences in the PCC and mPFC regions.

Figure 2.3 shows the p-values regarding the group differences on the default mode network. Compared with PTSD− women, our method showed that the default mode network of PTSD+ women had significantly stronger source signals in all regions of the network as compared to the PTSD− women. This implies that functional connectivities among the brain re-

gions within this network were stronger for the PTSD+ women, after controlling for subjects' age and depression status. Our results are consistent with recent findings in neuroscience literature that report abnormally high functional connectivity within the default mode network during both resting states and tasks for patients with mental disorders such as schizophrenia, depression and PTSD (Greicius et al., 2007; Whitfield-Gabrieli et al., 2009; Daniels et al., 2011). In comparison, dual-regression only identified a few distinctions between the two groups in the PCC and mPFC regions but didn't detect any differences in the LPC part of the default mode network. After FDR correction, none of the findings based on dual-regression remained significant.

2.3 Simulation Study

We conducted three sets of simulation studies to 1) evaluate the performance of the proposed hc-ICA model as compared with the existing TC-GICA model, 2) to compare the accuracy of the subspace-based approximate EM algorithm vs. the exact EM, 3) and to evaluate the performance of the proposed inference method for testing covariate effects based on hc-ICA.

2.3.1 Simulation study I: performance of the hc-ICA v.s. TC-GICA

In the first simulation study, we evaluate the performance of the proposed hc-ICA model compared with dual-regression ICA. We simulated fMRI data from three underlying source signals, i.e., $q = 3$, and considered three sample sizes with the number of subjects of $N = 10, 20, 40$. For each source, we generated a 3D spatial map with the dimension of $25 \times 25 \times 4$ and the activated signals in each source is displayed in Figure 2.4(A). For spatial source signals, we first generated population-level spatial maps, i.e., $\{\mathbf{s}_0(v)\}$, as the activated signals plus Gaussian random variability of a variance of 0.5. We then generated two covariates for each subject with

one being categorical ($x_1 \overset{iid}{\sim}$ Bernoulli(0.5)) and the other being continuous ($x_2 \overset{iid}{\sim}$ Uniform$(-1, 1)$). The covariate effects maps, i.e., $\{\boldsymbol{\beta}(v)\}$, are presented in Figure 2.4(B1)-(B2) where the covariate effect parameters at each voxel took values from $\{0, 1.5, 1.8, 2.5, 3.0\}$. Additionally, we generated Gaussian subject-specific random effects, i.e., $\boldsymbol{\gamma}_i(v)$, and considered three levels of between-subject variability: low ($\mathbf{D} = \text{diag}(0.1, 0.3, 0.5)$), medium ($\mathbf{D} = \text{diag}(1.0, 1.2, 1.4)$) and high ($\mathbf{D} = \text{diag}(1.8, 2.0, 2.5)$). The subject-specific spatial source signals were then simulated as the linear combination of the population-level signals, covariate effects and subject-specific random effects. For temporal responses, each source signal had a time series of length of $T = 200$ that was generated based on time courses from real fMRI data and hence represented realistic fMRI temporal dynamics. We generated subject-specific time sources that had similar frequency features but different phase patterns (Guo, 2011), which represented temporal dynamics in resting-state fMRI signals. After simulating the spatial maps and time courses for the source signals, Gaussian background noise with a standard deviation of 1 ($E = I_q$) were added to generate observed fMRI data.

We applied both hc-ICA and dual-regression ICA to the simulated data. The computational time was about 10min (N=10), 16min ($N = 20$) and 25min ($N = 40$) for hc-ICA using the exact EM and around 45sec (approximately the same among all Ns) for dual-regression for each simulated dataset, using a desktop PC with an Intel i7 3.6GHz quad core CPU. Following previous work (Beckmann and Smith, 2005; Guo, 2011), we evaluate the performance of each method based on the correlations between the activation signals and estimated signals in both temporal and spatial domains. To compare the performance in estimating the covariate effects, we report the mean square errors (MSEs) of $\hat{\boldsymbol{\beta}}(v)$ defined by $\frac{1}{V}\sum_{v=1}^{V}\left\|\hat{\boldsymbol{\beta}}(v) - \boldsymbol{\beta}(v)\right\|_{\mathcal{F}}^2$ averaged across simulation runs. Here $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm for a matrix. Since ICA recovery is permutation invari-

ant, each estimated IC was matched with the original source with which it had the highest spatial correlation. We present the simulation results in Table 2.4. The results show that hc-ICA provides more accurate estimates for the source signals on both the population- and subject-level. It leads to smaller mean square errors in estimating the covariate effects. We also display the estimated population-level IC maps and the covariate effects maps from both methods in Figure 2.4. The hc-ICA shows, in Figure 2.4, much better performance in correctly detecting the true activation patterns and covariate effects for each IC. In comparison, the estimates of the population-level IC maps from dual-regression show clear “cross-talk" between the ICs. Furthermore, the estimated covariate effects maps based on dual-regression are noisier plus some mismatches across the ICs.

2.3.2 Simulation study II: performance of the approximate EM

In the second simulation study, we compare the performance of the exact EM algorithm with the approximate EM for the hc-ICA model. We simulated fMRI data for ten subjects and considered three model sizes with the number of source signals of $q = 3, 6, 10$. The fMRI data were generated using methods similar to that in Simulation Study I with 10 subjects and low between-subject variabilities. We then fitted the proposed hc-ICA model using both the exact EM and the approximate EM. Results from Table 2.2 show that the accuracy of the subspace-based EM is comparable with regard to that of the exact EM in both the spatial and temporal domains and on both population- and subject-level. The major advantage of the subspace-based EM is that it was much faster than the exact EM. This advantage becomes more clear with the increase of the number of ICs. For $q = 10$, the subspace-based EM only uses about 2% computation time of the exact EM.

The convergence rates are the same between the two EM algorithms. We note that as q increased to 10, the convergence rates slightly decrease

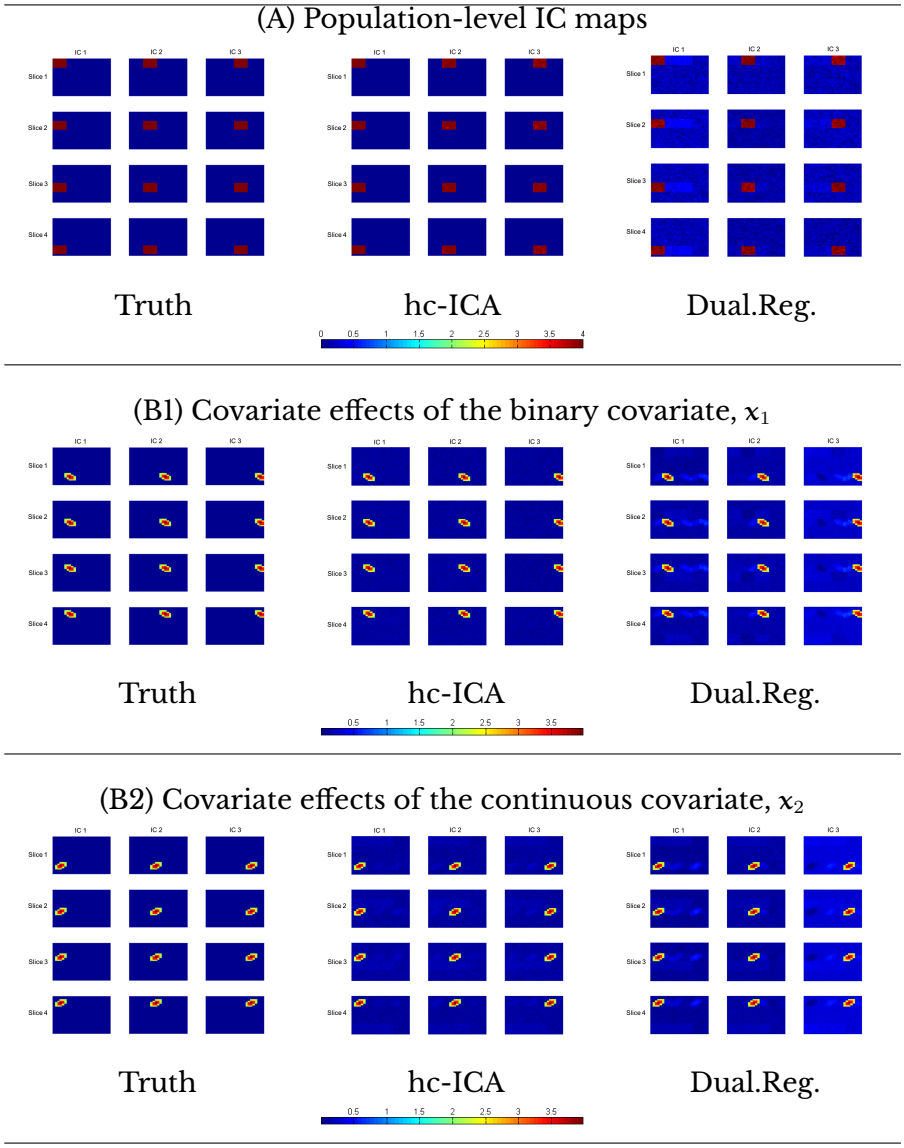


Figure 2.4: Comparison between our method and dual-regression ICA: truth, estimates from our model and estimates from dual-regression (N=10, between-subject variabilities are medium) are displayed based on 100 runs. All the images displayed are averaged across the 100 Monte Carlo datasets. Population-level spatial maps are shown in Figure 2.4(A). The results of dual-regression ICA are contaminated by the covariate effects. The results from our method are more accurate. Covariate effect estimates are shown in Figure 2.4(B1) and Figure 2.4(B2) respectively. The results of dual-regression show clear mismatching while our method provide accurate estimates.

Table 2.1: Simulation results for comparing our hc-ICA method against dual-regression ICA based on 100 runs. Values presented are mean and standard deviation of correlations between the true and estimated: subject-specific spatial maps, population-level spatial maps and subject-specific time courses. The mean and standard deviation of the MSE of the covariate estimates are also provided.

Btw-subj Var	Population-level spatial maps Corr.(SD)		Subject-specific spatial maps Corr.(SD)	
	hc-ICA	Dual.Reg.	hc-ICA	Dual.Reg.
Low				
N=10	0.982 (0.003)	0.956 (0.018)	0.984 (0.004)	0.945 (0.023)
N=20	0.990 (0.002)	0.968 (0.014)	0.996 (0.002)	0.949 (0.008)
N=40	0.992 (0.002)	0.976 (0.005)	0.996 (0.001)	0.956 (0.002)
Medium				
N=10	0.942 (0.017)	0.914 (0.048)	0.943 (0.011)	0.882 (0.030)
N=20	0.954 (0.002)	0.938 (0.034)	0.959 (0.004)	0.890 (0.016)
N=40	0.961 (0.002)	0.949 (0.020)	0.968 (0.003)	0.893 (0.009)
High				
N=10	0.833 (0.146)	0.740 (0.164)	0.894 (0.108)	0.689 (0.303)
N=20	0.850 (0.129)	0.795 (0.143)	0.909 (0.084)	0.695 (0.281)
N=40	0.871 (0.055)	0.809 (0.102)	0.928 (0.035)	0.705 (0.259)
Btw-subj Var.	Subject-specific time courses Corr.(SD)		Covariate Effects MSE(SD)	
	hc-ICA	Dual.Reg.	hc-ICA	Dual.Reg.
Low				
N=10	0.998 (0.001)	0.987 (0.010)	0.048 (0.019)	0.154 (0.055)
N=20	0.998 (0.001)	0.995 (0.004)	0.021 (0.003)	0.127 (0.044)
N=40	0.998 (0.001)	0.994 (0.004)	0.012 (0.001)	0.111 (0.030)
Medium				
N=10	0.993 (0.010)	0.970 (0.028)	0.273 (0.088)	0.485 (0.151)
N=20	0.998 (0.003)	0.976 (0.016)	0.117 (0.015)	0.285 (0.076)
N=40	0.998 (0.002)	0.991 (0.008)	0.064 (0.005)	0.187 (0.041)
High				
N=10	0.948 (0.021)	0.903 (0.045)	0.387 (0.157)	0.783 (0.325)
N=20	0.978 (0.018)	0.925 (0.029)	0.224 (0.075)	0.532 (0.271)
N=40	0.990 (0.015)	0.934 (0.022)	0.131 (0.056)	0.389 (0.198)

to 96%, which are lower than the EM algorithm for the TC-GICA model in (Guo, 2011). The main reason is that compared with the model in (Guo, 2011), which assumes common spatial maps across subjects, hc-ICA in-

volves a significantly larger number of parameters and latent variables by incorporating subject-specific IC maps and spatially varying covariate effects on each IC. The dramatic increase in the number of parameters for hc-ICA with larger q leads to the slightly decreased convergence rate. In practice, if the EM algorithm experiences convergence issues due to a large number of ICs, one can consider using existing group ICA software to first identify the uninteresting ICs, linearly remove them from the observed data and then perform hc-ICA on the new data with a smaller number of ICs. This technique has been commonly used in ICA applications to remove artifact-related components (Tohka et al., 2008; Griffanti et al., 2014).

Table 2.2: Simulation results for comparing the subspace-based approximate EM and the exact EM based on 50 runs. Mean and standard deviation of correlations between the true and estimated spatial maps and time courses are presented. The mean and standard deviation of the MSE of the covariate estimates are also provided.

	Population-level spatial maps Corr(SD)		Subject-specific spatial maps Corr(SD)	
# of IC	Exact EM	Approx. EM	Exact EM	Approx. EM
q=3	0.981(0.003)	0.981(0.001)	0.986(0.004)	0.981(0.002)
q=6	0.980(0.006)	0.980(0.006)	0.985(0.012)	0.981(0.011)
q=10	0.969(0.022)	0.963(0.020)	0.972(0.027)	0.970(0.022)
	Subject-specific time courses Corr(SD)		Covariate Effects MSE(SD)	
# of IC	Exact EM	Approx. EM	Exact EM	Approx. EM
q=3	0.998(0.001)	0.998(0.000)	0.048(0.020)	0.048(0.019)
q=6	0.997(0.003)	0.995(0.002)	0.069(0.024)	0.070(0.022)
q=10	0.992(0.016)	0.992(0.009)	0.105(0.033)	0.112(0.028)
	Time in miniute		Proportions of Convergence	
# of IC	Exact EM	Approx. EM	Exact EM	Approx. EM
q=3	9.91	5.22	100%	100%
q=6	71.05	9.09	100%	100%
q=10	860.10	19.02	96%	96%

2.3.3 Simulation study III: performance of the proposed inference procedures for covariate effects

We examine the performance of our inference procedures for $\hat{\boldsymbol{\beta}}(v)$ in the third simulation study. We simulated fMRI datasets with two source signals and considered sample sizes of $N = 20, 40, 80$. We generated two covariates in the same manner as in Simulation Study I. To facilitate computation, we generated images with the dimension of 20×20. The variance of between-subject random variabilities was 0.25 for both spatial source signals, and the within-subject variance was 0.4. We applied our hc-ICA method and dual-regression ICA for the simulated datasets and tested for the covariate effects using both methods. The hypotheses were $H_0 : \beta_{k\ell}(v) = 0$ versus $H_1 : \beta_{k\ell}(v) \neq 0$ at each voxel. Specifically, for hc-ICA, hypothesis tests were conducted for $\boldsymbol{\beta}(v)$ using the test proposed in section 2.5. In comparison, dual-regression method tested covariate effects by performing post-ICA regressions of the estimated subject-specific IC maps. We estimated the Type-I error rate with the empirical probabilities of not rejecting H_0 at voxels such that $\beta_{k\ell}(v) = 0$. We also estimated the power of the tests with the empirical probabilities of rejecting H_0 at voxels with non-zero values for the covariate effects, i.e., $\beta_{k\ell}(v) \in \{1.5, 1.8, 2.5, 3.0\}$. We report the average of the Type-I error rates at various significance levels, as well as and the powers with regard to different alternative hypothesis, in Table 2.3. According to Table 2.3, the type-I error rates from our inference method are always lower than those from dual-regression ICA. We do note that our Type-I error rates are slightly higher than the nominal level mainly due to the approximation in the inference procedure. From Table 2.3, we can also see that our method consistently demonstrate higher statistical power than dual-regression ICA. The results indicate that the proposed inference method based on hc-ICA provides more reliable and powerful inference about the covariate effects on the functional networks than the TC-GICA based dual-regression method.

Table 2.3: Simulation results for the inference of $\beta(v)$ based on 1000 runs. Type-I errors are averaged across all voxels with $\beta_{k\ell}(v) = 0$; powers are averaged across voxels having the same values of $\beta_{k\ell}(v) \neq 0$.

	N=20		N=40		N=80	
Type-I error analysis:						
size	hc-ICA	Dual.Reg.	hc-ICA	Dual.Reg	hc-ICA	Dual.Reg
0.01	0.014	0.029	0.012	0.025	0.012	0.018
0.05	0.062	0.084	0.056	0.076	0.055	0.062
0.10	0.129	0.205	0.118	0.190	0.112	0.149
0.50	0.522	0.580	0.516	0.565	0.514	0.557
0.80	0.835	0.872	0.820	0.856	0.810	0.840
Power analysis (test size: 0.05*):*						
$\beta(v)$	hc-ICA	Dual.Reg.	hc-ICA	Dual.Reg	hc-ICA	Dual.Reg
1.5	0.144	0.130	0.256	0.203	0.404	0.284
1.8	0.268	0.224	0.474	0.390	0.812	0.548
2.5	0.589	0.475	0.862	0.705	0.963	0.839
3.0	0.907	0.845	1.000	0.922	1.000	1.000

2.4 Discussion

We propose a hierarchical covariate-adjusted ICA (hc-ICA) model to formally quantify and test differences in brain functional networks related to subjects' demographic, clinical and biological characteristics. Our hc-ICA approach can be applied to study brain networks in both task-related and resting state fMRI studies. We develop a maximum likelihood estimation method based on EM algorithms for hc-ICA. We use an efficient approximate procedure to make inferences about covariate effects in our model. Simulation studies show that our methods provide more accurate estimation and inference for covariate effects on brain networks than the widely used dual-regression method. Application of hc-ICA to the Grady PTSD Study reveals important differences in brain functional networks between PTSD+ and PTSD− African American women, after adjusting for their ages and depression scores.

One of the main challenges in statistical modeling of brain imaging is the heavy computation load. In this chapter, we develop computa-

tionally efficient estimation and inference procedures for the proposed hc-ICA model. In particular, by exploiting sparsity in fMRI source signals, the subspace-based EM algorithm significantly reduces the computational time via concentration of probability masses on a subspace of the latent multinomial variables. We show theoretically that the subspace-based approximate method is supported by the characteristics of fMRI signals. We demonstrate empirically that the approximate EM provides highly accurate results. The definition of the subspace implies that there is little overlap in the spatial distributions of fMRI source signals. This is supported by findings in neuroscience literature which showed that brain functional networks are mostly separate (Beckmann et al., 2005; Smith et al., 2009). However, there are a few network hubs in the brain, consisted of a very small proportion of voxels, that may be involved in multiple networks. To investigate the performance of the subspace-based EM in this case, we have conducted additional simulation studies which generated data from overlapping source signals. Results show that the subspace EM still maintains good performance in recovering overlapping spatial signals.

Our hc-ICA model estimation is performed via a formal and unified maximum likelihood estimation which simultaneously estimates all parameters and latent variables in the model. By doing so, we improve the accuracy in estimating the brain networks on both population- and individual-level significantly; we also achieve higher statistical power in detecting differences in the networks. This holistic estimation approach does lead to heavier computation load compared with TC-GICA two-stage methods. The computation can be accelerated using several strategies. First, based on preliminary analysis of the data, we can identify ICs that are not of strong interest in a study, apply the standard procedure mentioned at the end of section 4.2 to remove them from the data, and then apply hc-ICA model to investigate group differences in the remaining ICs. Second, we can also apply standard multi-process/multi-

thread computing techniques to reduce computational time at a large scale, since most parts of our EM algorithm can be parallelized for each voxel (see supplementary material section 2-3 for details).

One potential extension to hc-ICA is to incorporate spatial dependence on modeling the spatially varying covariate effects $\beta(v)$. This can help increase the accuracy in detecting covariate-related network differences when they are spatially-correlated. Furthermore, we can accommodate spatial dependence in the residual terms in both the first and the second level of the hc-ICA model, which may help improve the accuracy and efficiency of the propose hc-ICA framework for investigating differences between functional networks.

2.5 Appendices

2.5.1 The Conditional Expectation Function in the E-step

The E-step of our EM algorithm evaluates the conditional expectation of the complete data log-likelihood which be expressed as

$$Q(\Theta|\hat{\Theta}^{(k)}) = Q_1(\Theta \mid \hat{\Theta}^{(k)}) + Q_2(\Theta \mid \hat{\Theta}^{(k)}) + Q_3(\Theta \mid \hat{\Theta}^{(k)}) + Q_4(\Theta \mid \hat{\Theta}^{(k)}),$$

where

$$\begin{aligned} Q_1(\Theta \mid \hat{\Theta}^{(k)}) = -\frac{NV}{2}\log|\mathbf{E}| - \frac{1}{2}\sum_{v=1}^{V}\sum_{i=1}^{N} \mathrm{tr}\Big\{ \mathbf{E}^{-1}\Big[\mathbf{y}_i(v)\mathbf{y}_i(v)' - 2\mathbf{A}_i E[\mathbf{s}_i(v)|\mathbf{y}(v); \hat{\Theta}^{(k)}]\mathbf{y}_i(v)' \\ + \mathbf{A}_i E[\mathbf{s}_i(v)\mathbf{s}_i(v)'|\mathbf{y}(v); \hat{\Theta}^{(k)}]\mathbf{A}_i' \Big] \Big\}, \end{aligned}$$

$$\begin{aligned} Q_2(\Theta \mid \hat{\Theta}^{(k)}) = -\frac{NV}{2}\log|\mathbf{D}| - \frac{1}{2}\sum_{v=1}^{V}\sum_{i=1}^{N} \mathrm{tr}\Big\{ \mathbf{D}^{-1}\Big[E[\mathbf{s}_i(v)\mathbf{s}_i(v)'|\mathbf{y}(v); \hat{\Theta}^{(k)}] \\ + E[\mathbf{s}_0(v)\mathbf{s}_0(v)'|\mathbf{y}(v); \hat{\Theta}^{(k)}] + \boldsymbol{\beta}(v)'\mathbf{x}_i\mathbf{x}_i'\boldsymbol{\beta}(v) - 2E[\mathbf{s}_i(v)\mathbf{s}_0(v)'|\mathbf{y}(v); \hat{\Theta}^{(k)}] \\ + 2E[\mathbf{s}_0(v)|\mathbf{y}(v); \hat{\Theta}^{(k)}]\mathbf{x}_i'\boldsymbol{\beta}(v) - 2E[\mathbf{s}_i(v)|\mathbf{y}(v); \hat{\Theta}^{(k)}]\mathbf{x}_i'\boldsymbol{\beta}(v) \Big] \Big\}, \end{aligned}$$

$$Q_3(\Theta \mid \hat{\Theta}^{(k)}) = -\frac{1}{2}\sum_{v=1}^{V}\sum_{\ell=1}^{q}\sum_{j=1}^{m} p[z_\ell(v)=j|\mathbf{y}(v);\hat{\Theta}^{(k)}]\Big\{\log\sigma^2_{\ell,j} + \frac{1}{\sigma^2_{\ell,j}}\Big[\mu^2_{\ell,j} + E[s_{0\ell}(v)^2|z_\ell(v)=j;\mathbf{y}(v),\hat{\Theta}^{(k)}] - 2\mu_{\ell,j}E[s_{0\ell}(v)|z_\ell(v)=j,\mathbf{y}(v);\hat{\Theta}^{(k)}]\Big]\Big\},$$

$$Q_4(\Theta \mid \hat{\Theta}^{(k)}) = \sum_{v=1}^{V}\sum_{\ell=1}^{q}\sum_{j=1}^{m} p[z_\ell(v)=j|\mathbf{y}(v);\hat{\Theta}^{(k)}]\log\pi_{\ell,j},$$

and $\mathbf{y}(v) = [\mathbf{y}_1(v)', ..., \mathbf{y}_N(v)']'$ contains all the observed data at voxel v (for all the N subjects). To evaluate the Q-functions, we need the joint conditional distribution, $p[\mathbf{s}(v), z(v) \mid \mathbf{y}(v);\Theta]$ where $\mathbf{s}(v) = [\mathbf{s}_1(v)', ..., \mathbf{s}_N(v)', \mathbf{s}_0(v)']'$.

2.5.2 The derivation of conditional probabilities in the E-step

In this section, we provide details of the E-step in our exact EM. We mainly focus on deriving $p[\mathbf{s}(v), z(v) \mid \mathbf{y}(v);\Theta]$ as well as its marginals. By collapsing our model across the N subjects as, for $v = 1, ..., V$,

$$\mathbf{A}'\mathbf{y}(v) = \mathbf{B}\mathbf{x} + \mathbf{U}\boldsymbol{\mu}_{z(v)} + \mathbb{R}\mathbf{r}_{z(v)} + \mathbf{e}(v), \tag{2.16}$$

where $\mathbf{r}_{z(v)} = [\boldsymbol{\gamma}_1(v)', ..., \boldsymbol{\gamma}_N(v)', \boldsymbol{\psi}'_{z(v)}]'$ concatenates error terms in the second and third level models, $\mathbf{e}(v) = [\boldsymbol{e}_1(v)', ..., \boldsymbol{e}_N(v)']'$ contains random errors for the first level model across all subjects, $\mathbf{x} = [\mathbf{x}'_1, ..., \mathbf{x}'_N]'$ represents all the covariate measurements, $\mathbf{B} = \mathbf{I}_N \otimes \boldsymbol{\beta}(v)'$, $\mathbf{U} = \mathbf{1}_N \otimes \mathbf{I}_q$, $\mathbf{R} = [\mathbf{I}_{Nq}, \mathbf{1}_N \otimes \mathbf{I}_q]$ and $\mathbf{A} = \text{blockdiag}(\mathbf{A}_1, ..., \mathbf{A}_N)$ is a combined mixing matrix with $\mathbf{A}_i$s as its block diagonal elements ($\mathbf{A}$ is also orthogonal). It is trivial to have that in (2.16), $\mathbf{e}(v) \sim N(\mathbf{0}, \boldsymbol{\Upsilon})$ and $\mathbf{r}_{z(v)} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_{z(v)})$ where $\boldsymbol{\Upsilon} = \mathbf{I}_N \otimes \mathbf{E}$ and $\boldsymbol{\Gamma}_{z(v)} = \text{blockdiag}(\mathbf{I}_N \otimes \mathbf{D}, \boldsymbol{\Sigma}_{z(v)})$. Thus (2.16) can be represent as

$$\mathbf{y}_0(v) \sim N(\mathbb{R}\mathbf{r}_{z(v)}, \boldsymbol{\Upsilon}), \quad \mathbf{r}_{z(v)} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_{z(v)})$$

where $\mathbf{y}_0(v) = \mathbf{A}'\mathbf{y}(v) - \mathbf{B}\mathbf{x} - \mathbf{U}\boldsymbol{\mu}_{z(v)}$. This representation is a canonical Bayesian general linear model given $z(v)$. Then given $z(v)$ and conditional

on $\mathbf{y}(\nu)$, $p[\mathbf{r}_{z(\nu)} \mid \mathbf{y}(\nu), z(\nu); \Theta] = g(\boldsymbol{\mu}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)}, \boldsymbol{\Sigma}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)})$ where

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)} &= \boldsymbol{\Sigma}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)}\mathbb{R}'\boldsymbol{\Upsilon}^{-1}[\mathbf{A}'\mathbf{y}(\nu) - \mathbf{B}\mathbf{x} - \mathbf{U}\boldsymbol{\mu}_{z(\nu)}],\\ \boldsymbol{\Sigma}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)} &= \left(\mathbb{R}'\boldsymbol{\Upsilon}^{-1}\mathbb{R} + \boldsymbol{\Gamma}_{z(\nu)}^{-1}\right)^{-1}.\end{aligned}$$

It is trivial to show that $\mathbf{s}(\nu) = \mathbf{P}\mathbf{r}_{z(\nu)} + \mathbf{Q}_{z(\nu)}$, where

$$\mathbf{P} = \begin{pmatrix} \mathbf{I}_{Nq}, & \mathbf{U} \\ \mathbf{0}, & \mathbf{I}_q \end{pmatrix}, \quad \mathbf{Q}_{z(\nu)} = \begin{pmatrix} \mathbf{B}\mathbf{x} + \mathbf{U}\boldsymbol{\mu}_{z(\nu)} \\ \boldsymbol{\mu}_{z(\nu)} \end{pmatrix},$$

we can easily have that:

$$p[\mathbf{s}(\nu) \mid \mathbf{y}(\nu), z(\nu); \Theta] = g(\mathbf{P}\boldsymbol{\mu}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)} + \mathbf{Q}_{z(\nu)}, \mathbf{P}\boldsymbol{\Sigma}_{\mathbf{r}(\nu)|\mathbf{y}(\nu)}\mathbf{P}'). \tag{2.17}$$

Next we need to find $p[z(\nu) \mid \mathbf{y}(\nu); \Theta]$. From (2.16), we have that $p[\mathbf{A}'\mathbf{y}(\nu) \mid z(\nu)] = g(\mathbf{B}\mathbf{x} + \mathbf{U}\boldsymbol{\mu}_{z(\nu)}, \mathbb{R}\boldsymbol{\Gamma}_{z(\nu)}\mathbb{R}' + \boldsymbol{\Upsilon})$. Notice that $p[z(\nu)] = \prod_{\ell=1}^{q} \pi_{\ell,z_\ell(\nu)}$ for all ν, by simply applying the Bayes' theorem,

$$p[z(\nu) \mid \mathbf{y}(\nu); \Theta] = \frac{\left[\prod_{\ell=1}^{q} \pi_{\ell,z_\ell(\nu)}\right] g(\mathbf{A}'\mathbf{y}(\nu); \mathbf{B}\mathbf{x} + \mathbf{U}\boldsymbol{\mu}_{z(\nu)}, \mathbb{R}\boldsymbol{\Gamma}_{z(\nu)}\mathbb{R}' + \boldsymbol{\Upsilon})}{\sum_{z(\nu)\in\mathcal{R}}\left[\prod_{\ell=1}^{q} \pi_{\ell,z_\ell(\nu)}\right] g(\mathbf{A}'\mathbf{y}(\nu); \mathbf{B}\mathbf{x} + \mathbf{U}\boldsymbol{\mu}_{z(\nu)}, \mathbb{R}\boldsymbol{\Gamma}_{z(\nu)}\mathbb{R}' + \boldsymbol{\Upsilon})}, \tag{2.18}$$

where $\mathcal{R}$ is the range of $z(\nu) = [z_1(\nu), ..., z_q(\nu)]', z_\ell(\nu) = 1, ..., m$, which contains m^q distinct vectors in $\mathbb{R}^q$.

Given this probability distributions, the moments in the Q-functions can be easily derived and they all have analytical forms.

2.5.3 Details of the M-step in the exact EM

In the M-step, we update the parameters within our model as follows:

- Update $\boldsymbol{\beta}(\nu)$: for $\nu = 1, ..., V$,

$$\hat{\boldsymbol{\beta}}(\nu)^{(k+1)} = \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \left\{\mathbf{x}_i \left(E[\mathbf{s}_i(\nu)'|\mathbf{y}(\nu); \hat{\Theta}^{(k)}] - E[\mathbf{s}_0(\nu)'|\mathbf{y}(\nu); \hat{\Theta}^{(k)}]\right)\right\}. \tag{2.19}$$

- Update $\mathbf{A}_i$: for $i = 1, ..., N$, we let

$$\check{\mathbf{A}}_i^{(k+1)} = \left\{ \sum_{v=1}^{V} \mathbf{y}_i(v) \mathrm{E}[\mathbf{s}_i(v) | \mathbf{y}(v); \hat{\Theta}^{(k)}] \right\} \left\{ \sum_{v=1}^{V} \mathrm{E}[\mathbf{s}_i(v) \mathbf{s}_i(v)' | \mathbf{y}(v); \hat{\Theta}^{(k)}] \right\}^{-1}, \tag{2.20}$$

and then update $\widehat{\mathbf{A}}_i^{(k+1)} = \mathcal{H}(\check{\mathbf{A}}_i^{(k+1)})$ where $\mathcal{H}(\cdot)$ is the orthogonalization transformation.

- Update $\mathbf{E} = \mathbf{I}_q \nu_0^2$ with:

$$\begin{aligned} \hat{\nu}_0^{2(k+1)} = & \frac{1}{TNV} \sum_{v=1}^{V} \sum_{i=1}^{N} \Big\{ \mathbf{y}_i(v)' \mathbf{y}_i(v) - 2 \mathbf{y}_i(v)' \widehat{\mathbf{A}}_i^{(k+1)} \mathrm{E}[\mathbf{s}_i(v) | \mathbf{y}(v); \hat{\Theta}^{(k)}] \\ & + \mathrm{tr} \left[\widehat{\mathbf{A}}_i^{(k+1)\prime} \widehat{\mathbf{A}}_i^{(k+1)} \mathrm{E}[\mathbf{s}_i(v) \mathbf{s}_i(v)' | \mathbf{y}(v); \hat{\Theta}^{(k)}] \right] \Big\}. \end{aligned} \tag{2.21}$$

- Update $\mathbf{D} = \mathrm{diag}(\nu_1^2, ..., \nu_q^2)$: for $\ell = 1, ..., q$,

$$\begin{aligned} \hat{\nu}_\ell^{2(k+1)} = & \frac{1}{NV} \sum_{v}^{V} \sum_{i=1}^{N} \Big\{ \mathrm{E}[s_{i\ell}(v)^2 | \mathbf{y}(v); \hat{\Theta}^{(k)}] + \mathrm{E}[s_{0\ell}(v)^2 | \mathbf{y}(v); \hat{\Theta}^{(k)}] \\ & - 2\mathrm{E}[s_{i\ell}(v) s_\ell(v) | \mathbf{y}(v); \hat{\Theta}^{(k)}] + \hat{\boldsymbol{\beta}}_\ell(v)^{(k+1)\prime} \mathbf{x}_i \mathbf{x}_i' \hat{\boldsymbol{\beta}}_\ell(v)^{(k+1)} \\ & + 2 \left(\mathrm{E}[s_{0\ell}(v) | \mathbf{y}(v); \hat{\Theta}^{(k)}] - \mathrm{E}[s_{i\ell}(v) | \mathbf{y}(v); \hat{\Theta}^{(k)}] \right) \mathbf{x}_i' \hat{\boldsymbol{\beta}}_\ell(v)^{(k+1)} \Big\}, \end{aligned} \tag{2.22}$$

where $\hat{\boldsymbol{\beta}}_\ell(v)^{(k+1)}$ is the ℓth column of $\hat{\boldsymbol{\beta}}(v)^{(k+1)}$.

- Update $\pi_{\ell,j}$:

$$\hat{\pi}_{\ell,j}^{(k+1)} = \frac{1}{V} \sum_{v=1}^{V} p[z_\ell(v) = j | \mathbf{y}(v); \hat{\Theta}^{(k)}]. \tag{2.23}$$

- Update $\mu_{\ell,j}$:

$$\hat{\mu}_{\ell,j}^{(k+1)} = \frac{\sum_{v=1}^{V} p[z_\ell(v) = j | \mathbf{y}(v); \hat{\Theta}^{(k)}] \mathrm{E}[s_{0\ell}(v) | z_\ell(v) = j, \mathbf{y}(v); \hat{\Theta}^{(k)}]}{V \hat{\pi}_{\ell,j}^{(k+1)}}. \tag{2.24}$$

- Update $\sigma^2_{\ell,j}$:

$$\hat{\sigma}^{2(k+1)}_{\ell,j} = \frac{\sum_{v=1}^{V} p[z_\ell(v) = j|\mathbf{y}(v); \hat{\Theta}^{(k)}]E[s_{0\ell}(v)^2|z_\ell(v) = j, \mathbf{y}(v); \hat{\Theta}^{(k)}]}{V\hat{\pi}^{(k+1)}_{\ell,j}} - [\hat{\mu}^{(k+1)}_{\ell,j}]^2. \tag{2.25}$$

Here, $E[s_{0\ell}(v) \mid z_\ell(v) = j, \mathbf{y}(v); \Theta]$, $E[s_{0\ell}(v)^2 \mid z_\ell(v) = j, \mathbf{y}(v); \Theta]$ and $p[z_\ell(v) = j \mid \mathbf{y}(v); \Theta]$ are the marginal conditional moments and probability related to the ℓth IC. They are derived by summing across all the possible states of the other $q - 1$ ICs as follows,

$$E[s_{0\ell}(v) \mid z_\ell(v) = j, \mathbf{y}(v); \Theta] = \frac{\sum_{z(v)\in\mathcal{R}^{(\ell,j)}} p[z(v) \mid \mathbf{y}(v); \Theta]E[s_{0\ell}(v) \mid \mathbf{y}(v), z(v); \Theta]}{p[z_\ell(v) = j \mid \mathbf{y}(v); \Theta]}, \tag{2.26}$$

$$p[z_\ell(v) = j \mid \mathbf{y}(v); \Theta] = \sum_{z(v)\in\mathcal{R}^{(\ell,j)}} p[z(v) \mid \mathbf{y}(v); \Theta]. \tag{2.27}$$

where $\mathcal{R}^{(\ell,j)}$ is defined as $\{z^r \in \mathcal{R} : z^r_\ell = j\}$ for all $\ell = 1, .., q, j = 1, ..., m$.

2.5.4 Proof of Theorem 1

We prove Theorem 1 by introducing a lemma.

Lemma 1. *If the elements of* $z(v) = [z_1(v), ..., z_q(v)]'$ *are independent with* $p[z_\ell(v) = j] = \pi_{\ell,j}$ *for* $j = 1, ..., m, \ell = 1, ..., q$, *then* $p[z(v) \in \mathcal{R}_0 \cup \mathcal{R}_1] = \mathcal{F}(\kappa)$ *where*

$$\mathcal{F}(\kappa) = \frac{1 + \sum_{\ell=1}^{q} \kappa_\ell}{\prod_{\ell=1}^{q}(1 + \kappa_\ell)}, \tag{2.28}$$

with $\kappa = [\kappa_1, ..., \kappa_q]'$ *and* $\kappa_\ell = p[z_\ell(v) \neq 1]/p[z_\ell(v) = 1]$ *for all* $\ell = 1, ..., q$.

The parameters $\kappa = [\kappa_1, ..., \kappa_q]'$ can be interpreted as the odds for a random voxel of being activated/deactivated versus exhibiting background fluctuation in IC ℓ. Lemma 1 indicates that the probability of interest, $p[z(v) \in \mathcal{R}_0 \cup \mathcal{R}_1]$, depends on $\{\pi_{\ell,j}\}$ only through the odds. The proof of Lemma 1 is provided as follows.

Proof. Let $\tau_{\ell,j} = \pi_{\ell,j}/\pi_{\ell,1}, j = 2, ..., m$, then $\kappa_\ell = \frac{p[z_\ell(v)\neq 1]}{p[z_\ell(v)=1]} = \sum_{j=2}^{m} \tau_{\ell,j}$. By definition $\mathcal{R}_0 \cup \mathcal{R}_1 = \emptyset$ and $p[z(v) \in \mathcal{R}_0] = \prod_{\ell=1}^{q} p[z_\ell(v) = 1] = \prod_{\ell=1}^{q} \pi_{\ell,1}$. For a

given $z(\nu) \in \mathcal{R}_1$, suppose $z_t(\nu) = j > 1$ for some $t \in \{1, ..., q\}$ and $z_{\ell \neq t}(\nu) = 1$, then $p[z(\nu)] = \tau_{t,j} \prod_{\ell=1}^{q} \pi_{\ell,1}$. This implies that

$$p[z(\nu) \in \mathcal{R}_1] = \left(\sum_{t=1}^{q} \sum_{j=2}^{m} \tau_{t,j} \right) \prod_{\ell=1}^{q} \pi_{\ell,1} = \left(\sum_{\ell=1}^{q} \kappa_\ell \right) \prod_{\ell=1}^{q} \pi_{\ell,1}.$$

Also we have that $\sum_{j=1}^{m} \pi_{\ell,j} = 1$ for all $\ell = 1, ..., q$, then $\pi_{\ell,1} + \pi_{\ell,1} \sum_{j=2}^{m} \tau_{\ell,j} = (1 + \kappa_\ell)\pi_{\ell,1} = 1$, which gives $\pi_{\ell,1} = 1/(1 + \kappa_\ell)$. Thus

$$\begin{aligned} p[z(\nu) \in \mathcal{R}_0 \cup \mathcal{R}_1] &= p[z(\nu) \in \mathcal{R}_0] + p[z(\nu) \in \mathcal{R}_1] \\ &= \left(1 + \sum_{\ell=1}^{q} \kappa_\ell \right) \prod_{\ell=1}^{q} \pi_{\ell,1} \\ &= \frac{1 + \sum_{\ell=1}^{q} \kappa_\ell}{\prod_{\ell=1}^{q} (1 + \kappa_\ell)} \end{aligned} \tag{2.29}$$

□

Based on Lemma 1, we prove Theorem 1 in the following.

Proof. We notice that

$$\kappa_\ell = \frac{p[z_\ell(\nu) \neq 1]}{p[z_\ell(\nu) = 1]} = \frac{1 - \pi_{\ell,1}}{\pi_{\ell,1}}. \tag{2.30}$$

For all $0 < \epsilon < 1$, let $\delta = \frac{\sqrt{\epsilon}}{\sqrt{\epsilon}+q} \in (0, 1)$. Then if $\pi_{\ell,1} > 1 - \delta$, i.e., $\pi_{\ell,1} > \frac{q}{\sqrt{\epsilon}+q}$, we have that $0 < \kappa_\ell < \frac{\delta}{1-\delta}$ for all $\ell = 1, ..., q$. Based on the Taylor expansion for $p[z(\nu) \in \mathcal{R}_0 \cup \mathcal{R}_1] = \mathcal{F}(\boldsymbol{\kappa})$ at $\boldsymbol{\kappa} = \mathbf{0}$, $\exists 0 < \kappa_\ell^0 < \kappa_\ell$ for all $\ell = 1, ..., q$, such that

$$\begin{aligned} p[z(\nu) \in \mathcal{R}_0 \cup \mathcal{R}_1] &= \mathcal{F}(\mathbf{0}) + \sum_{\ell=1}^{q} \left. \frac{\partial \mathcal{F}}{\partial \kappa_\ell} \right|_{\kappa_\ell = \kappa_\ell^0} \kappa_\ell \\ &= 1 - \sum_{\ell=1}^{q} \frac{\sum_{j \neq \ell} \kappa_j^0}{\prod_{j \neq \ell} (1 + \kappa_j^0)} \frac{1}{(1 + \kappa_\ell^0)^2} \kappa_\ell \\ &> 1 - \sum_{\ell=1}^{q} \sum_{j \neq \ell} \kappa_j \kappa_\ell \\ &> 1 - \left(\frac{q\delta}{1 - \delta} \right)^2 \\ &= 1 - \epsilon \end{aligned} \tag{2.31}$$

□

2.5.5 Remarks on the subspace-based approximate EM

In the approximate EM, the conditional distribution $z(v) \mid \mathbf{y}(v)$ is determined by the probability masses

$$\tilde{p}[z(v) \mid \mathbf{y}(v), \Theta] = \begin{cases} \frac{[\prod_{\ell=1}^{q} \pi_{\ell,z_\ell(v)}]\, g(\mathbf{A}'\mathbf{y}(v); \mathbf{B}\mathbf{x}+\mathbf{U}\boldsymbol{\mu}_{z(v)}, \boldsymbol{\Gamma}_{z(v)}\mathbb{R}'+\boldsymbol{\Upsilon})}{\sum_{z(v)\in\widetilde{\mathcal{R}}}[\prod_{\ell=1}^{q} \pi_{\ell,z_\ell(v)}]\, g(\mathbf{A}'\mathbf{y}(v); \mathbf{B}\mathbf{x}+\mathbf{U}\boldsymbol{\mu}_{z(v)}, \boldsymbol{\Gamma}_{z(v)}\mathbb{R}'+\boldsymbol{\Upsilon})}, & z(v) \in \widetilde{\mathcal{R}} \\ 0, & z(v) \in \mathcal{R}\backslash\widetilde{\mathcal{R}} \end{cases} \tag{2.32}$$

where $\widetilde{\mathcal{R}} = \mathcal{R}_0 \cup \mathcal{R}_1$. Thus we use a sparse vector of probability masses, with concentration of measures on the subset $\widetilde{\mathcal{R}} = \mathcal{R}_0 \cup \mathcal{R}_1$, to approximate the exact conditional distribution of $z(v)$ given $\mathbf{y}(v)$. The follow-up evaluations of the conditional moments in the E-step only involves $z(v) \in \widetilde{\mathcal{R}}$. And the corresponding definition of $\mathcal{R}^{(\ell,j)}$ is adapted to $\widetilde{\mathcal{R}}^{(\ell,j)} = \{z^r \in \widetilde{\mathcal{R}} : z_\ell^r = j\}$.

2.5.6 Thresholding the spatial maps based on the ML estimates for functional brain networks

We threshold the estimated spatial maps to identify the activated/deactivated regions of the brain in each functional network. This goal can be achieved naturally through our model estimation based on conditional probabilities. To be specific, if we assume that $z_\ell(v) = j$ indicates that the ℓth component is activated at voxel v, then we can calculate $p[z_\ell(v) = j \mid \mathbf{y}(v); \hat{\Theta}] = \sum_{z(v)\in\mathcal{R}^{(\ell,j)}} p[z(v) \mid \mathbf{y}(v); \hat{\Theta}]$, where $\mathcal{R}^{(\ell,j)}$ is defined as $\{z^r \in \mathcal{R} : z_\ell^r = j\}$ for all $\ell = 1, .., q, j = 1, ..., m$. This probability characterizes the state of voxel v within network ℓ. We can then obtain the spatial map for a functional network by thresholding $p[z_\ell(v) = j \mid \mathbf{y}(v); \hat{\Theta}]$ with a pre-specified probability.

2.5.7 Specifying the initial values for hc-ICA

The initialization of hc-ICA is specified as follows throughout this chapter:

- Concatenate fMRI data from each subject under the typical TC-GICA setting; perform group ICA based on FastICA with $g(u) = u^3$ non-linearity;

- Each subject's pre-processed fMRI data were regressed against the group spatial maps obtained via TC-GICA to obtain the initial values for $\widehat{\mathbf{A}}_i^{(0)}, i = 1, ..., n$, the subject-specific mixing matrices.

- Each subject's pre-processed fMRI data were then regressed against $\widehat{\mathbf{A}}_i^{(0)}$ to obtain initial estimates $\widehat{s_i}^{(0)}(v)$;

- $\widehat{s_i}^{(0)}(v)$ was then regressed against $x_i(v)$ to obtain the initial estimates $\widehat{\boldsymbol{\beta}}(v)^{(0)}$.

- Specific $m = 2$ for the MoG source distribution model. Specifying the following initial values for parameters in the MoG: $\hat{\boldsymbol{\pi}}^{(0)} = [0.9,\ 0.1]'$, $\hat{\boldsymbol{\mu}}^{(0)} = [0,\ 1]'$ and $\hat{\sigma}^{2(0)} = [1,\ 1]'$;

- set initial values for error variances: $\hat{\mathbf{D}}^{(0)} = \hat{\mathbf{E}}^{(0)} = \mathbf{I}_q$.

An alternative way to specifying initial values in the MoG and error variances is to fit MoG distributions on the estimated source signals $\widehat{s_i}^{(0)}(v)$ and use the estimated $\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$ as initial guesses. However, we found that our EM algorithms were rather robust against the specifications for these parameters throughout empirical analysis, thus we choose to fixed these initial values as ones given above.

2.5.8 Additional Simulation Studies

Comparison between hc-ICA and dual-regression with larger number of ICs

In this section, we conducted simulation studies under the same setting of Simulation I in the chapter but increased the number of ICs to $q = 10$. The goal is to evaluate the performance of hc-ICA as compared to a TC-GICA two-stage analysis, i.e. dual-regression method, when the number of ICs

is larger. We estimated parameters in hc-ICA using the approximate EM to speed up computation in the simulation.

Two levels of between-subject variabilities were considered: low ($\mathbf{D} = 0.5\mathbf{I}_{10}$) and high ($\mathbf{D} = 2\mathbf{I}_{10}$). Simulation results are summarized in Table 2.4. Based on Table 2.4, we confirmed that the advantage of hc-ICA over the dual-regression in terms of estimating the spatial maps (both population-level and subject-specific), time courses and covariate effects still holds when the number of ICs becomes larger.

Performance of the approximate EM for data that deviate from conditions in Theorem 1

We conducted simulations to evaluate the validity of the approximate EM algorithm when there are overlapping regions between the ICs. The simulation setting was similar to the one described in Simulation I except that we considered $40\times40\times1$ 2D planes with $q = 2$ ICs for simplicity. In this simulation, the 2 group ICs shared overlapping activated regions, representing two brain networks involving overlapping regions in the brain. We considered three sizes of overlapping regions: small (5.3% of the activated regions overlapped), medium (9.5% of the activated regions overlapped) and large (14.8% of the activated regions overlapped).

We applied hc-ICA to 50 replicates of the simulation data. Figure 3.3 presents the true group IC maps and the average estimates from both the exact EM algorithm and the approximate EM. The overlapping regions are marked by the red squares in the true signal maps (the top row). The activated regions are marked by the yellow color. For each panel, the two columns correspond to the two different ICs.

From Figure 3.3, we can see that even with moderate to high levels of overlapping ($9.5\% - 14.8\%$), the approximate EM algorithm can still generate highly comparable IC maps compared with the exact EM and correctly identified the activation of the overlapping regions in both ICs. However, we do observe attenuated signals from the approximate EM

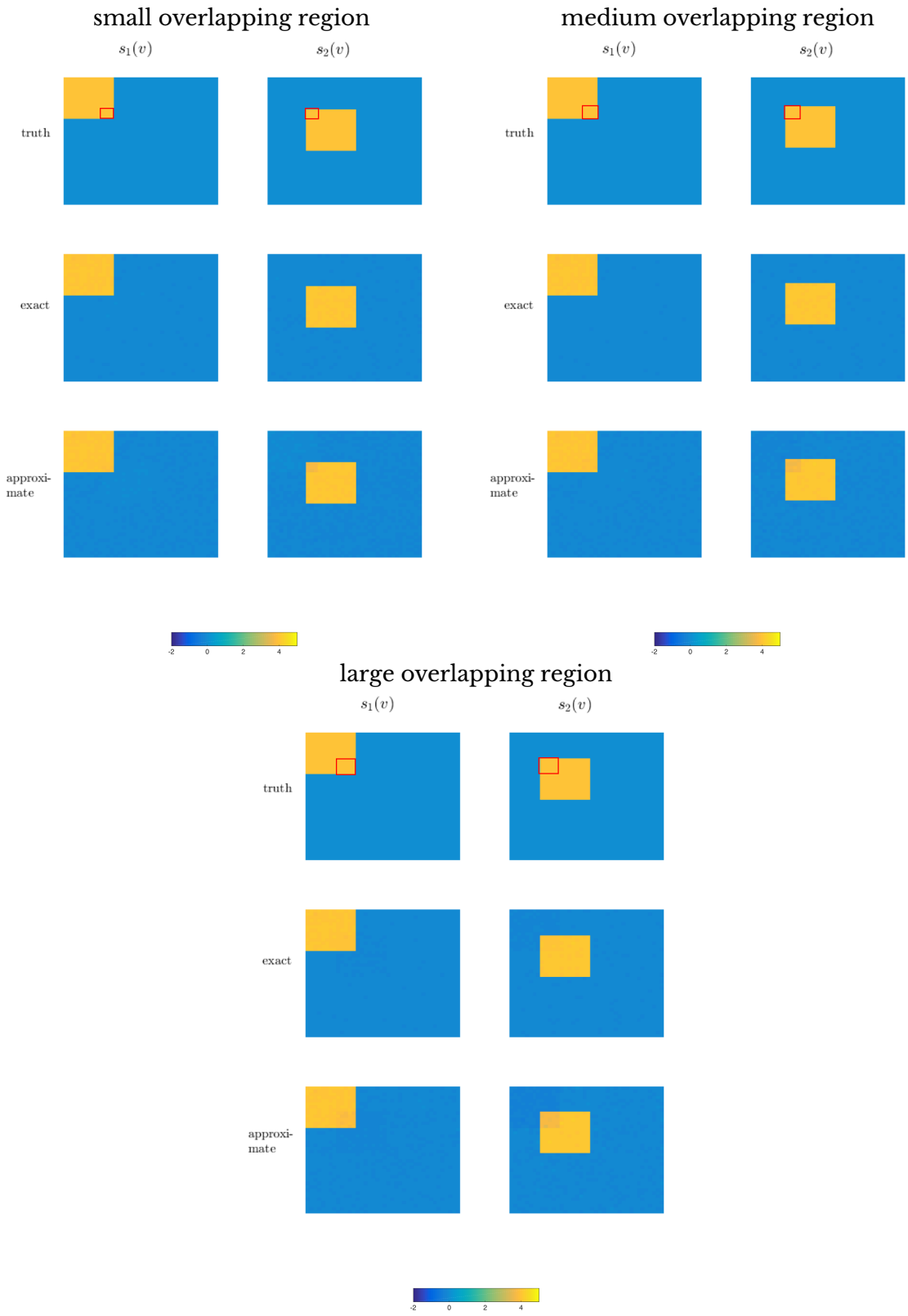


Figure 2.5: Evaluating the validity of subspace-based approximation EM for data generated from overlapping ICs: three panels show results from different level of overlapping; the two columns in each panel correspond to different ICs; the three rows in each panel represent true source signals, estimates based on the exact EM and estimates from the approximate EM

at pixels in the overlapping region due to the restriction to the subspace.

In the following, we present a toy example to provide theoretical explanation for why the subspace EM can capture overlapping source signals. We consider the case of two ICs, i.e. $\mathrm{q} = 2$. For the latent state variable $z_\ell(v), \ell = 1, 2$, $z_\ell(v) = 1$ representing voxel v demonstrates background fluctuation in IC ℓ and $z_\ell(v) = 2$ representing voxel v is activated in IC ℓ. Let $\mathbf{s}(v) = [\mathrm{s}_1(v), \mathrm{s}_2(v)]'$ represent the source signal at IC 1 and 2. The IC estimates $\hat{\mathbf{s}}$, which is based on the exact conditional expectation of $\mathbf{s}(v)$, is given as follows ((v) is dropped for the ease of exposition):

$$\begin{aligned}
\hat{\mathbf{s}} = \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \Theta] = \sum_{\boldsymbol{z} \in \mathcal{R}} \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z}, \Theta] \times \mathrm{p}[\boldsymbol{z} \mid \mathbf{y}, \Theta] \\
= & \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [1,\ 1]', \Theta] \times \mathrm{p}[\boldsymbol{z} = [1,\ 1]' \mid \mathbf{y}, \Theta] \\
& + \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [1,\ 2]', \Theta] \times \mathrm{p}[\boldsymbol{z} = [1,\ 2]' \mid \mathbf{y}, \Theta] \\
& + \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [2,\ 1]', \Theta] \times \mathrm{p}[\boldsymbol{z} = [2,\ 1]' \mid \mathbf{y}, \Theta] \\
& + \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [2,\ 2]', \Theta] \times \mathrm{p}[\boldsymbol{z} = [2,\ 2]' \mid \mathbf{y}, \Theta].
\end{aligned}$$

With the approximate EM, by restricting $\boldsymbol{z}$ within $\widetilde{\mathcal{R}}$, we set $\mathrm{p}[\boldsymbol{z} = [2,\ 2]' \mid \mathbf{y}, \Theta]$ to be zero and the approximate estimates $\hat{\tilde{\mathbf{s}}}$ is then

$$\begin{aligned}
\hat{\tilde{\mathbf{s}}} = \widetilde{\mathrm{E}}[\mathbf{s} \mid \mathbf{y}, \Theta] = & \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [1,\ 1]', \Theta] \times \frac{\mathrm{p}[\boldsymbol{z} = [1,\ 1]' \mid \mathbf{y}, \Theta]}{1 - \mathrm{p}[\boldsymbol{z} = [2,\ 2]' \mid \mathbf{y}(v), \Theta]} \\
& + \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [1,\ 2]', \Theta] \times \frac{\mathrm{p}[\boldsymbol{z} = [1,\ 2]' \mid \mathbf{y}, \Theta]}{1 - \mathrm{p}[\boldsymbol{z} = [2,\ 2]' \mid \mathbf{y}(v), \Theta]} \\
& + \mathrm{E}[\mathbf{s} \mid \mathbf{y}, \boldsymbol{z} = [2,\ 1]', \Theta] \times \frac{\mathrm{p}[\boldsymbol{z} = [2,\ 1]' \mid \mathbf{y}, \Theta]}{1 - \mathrm{p}[\boldsymbol{z} = [2,\ 2]' \mid \mathbf{y}(v), \Theta]}
\end{aligned}$$

For voxels that are activated in both IC 1 and IC2, although we do not have the state $[2, 2]$ in the restricted $\widetilde{\mathcal{R}}$, there will still be significant amount of probability masses assigned to both $\boldsymbol{z} = [2,\ 1]'$ and $\boldsymbol{z} = [1,\ 2]'$. Consequently, in $\hat{\tilde{\mathbf{s}}} = [\hat{\tilde{\mathrm{s}}}_1, \hat{\tilde{\mathrm{s}}}_2]'$, both $\hat{\tilde{\mathrm{s}}}_1$ and $\hat{\tilde{\mathrm{s}}}_2$ will still have values that are significantly different from zero indicating that the voxel is activated in both IC1 and IC2. The consequence of the restricting $\boldsymbol{z}$ to $\widetilde{\mathcal{R}}$ will be mainly reflected in the loss of some accuracy in the estimates $\hat{\tilde{\mathrm{s}}}_1$ and $\hat{\tilde{\mathrm{s}}}_2$. These

illustrative derivations are validated by the results from the above simulation studies (Figure 3.3).

2.5.9 Checking the stability of our EM algorithm for the PTSD data analysis

We repeated our real data analysis with 50 different initial values. For each of the 50 initializations, we simulated subject-specific noise matrices with random Gaussian elements. Then we added these noises to the initial guesses $\widehat{\mathbf{A}}_i^{(0)}$, computed from TC-GICA according to section 7. The noise contaminated mixing matrices were then orthogonalized and treated as our new initial guesses, $\widehat{\mathbf{A}}_i^{(0)}$. All other initial parameters were then derived based on this new $\widehat{\mathbf{A}}_i^{(0)}$ following the direction in section 7. The procedure above raises two sources of randomness: one from the TC-GICA estimates of $\mathbf{A}_i$ which mainly introduces sign changes and permutations to their columns; another one from the Gaussian noise matrices that we added to the TC-GICA estimates.

We ran our approximate EM algorithm on the PTSD data using the 50 different sets of initializations and reported the results from the one that corresponded to the highest observed data log-likelihood.

To check for robustness, we computed the correlations between all pairs of initializations for their resulting mixing matrices, group IC maps and covariate effects. We report in Table 2.5 the medians of these correlations, the inter-quantile ranges (IQR) as well as the 25th percentile (Q_1) and 75th percentile (Q_3). The high correlations between replicates from Table 2.5 indicates that our algorithm can generate stable estimates in general.

Table 2.4: Simulation results for comparing our hc-ICA method (approximate EM, $q = 10$) against the dual-regression ICA based on 100 runs. Values presented are mean and standard deviation of correlations between the true and estimated: subject-specific spatial maps, population-level spatial maps and subject-specific time courses. The mean and standard deviation of the MSE of the covariate estimates are also provided.

Btw-subj Var	Population-level spatial maps Corr.(SD)		Subject-specific spatial maps Corr.(SD)	
	hc-ICA	Dual.Reg.	hc-ICA	Dual.Reg.
Low				
N=10	0.965(0.012)	0.930(0.064)	0.968(0.014)	0.933 (0.072)
N=20	0.969(0.010)	0.945(0.045)	0.974(0.008)	0.952 (0.027)
N=40	0.974(0.006)	0.959(0.018)	0.990(0.003)	0.970 (0.008)
High				
N=10	0.814(0.162)	0.731(0.208)	0.869(0.187)	0.671(0.332)
N=20	0.820(0.147)	0.767(0.172)	0.879(0.136)	0.680(0.305)
N=40	0.842(0.096)	0.785(0.144)	0.908(0.049)	0.688(0.274)
Btw-subj Var.	Subject-specific time courses Corr.(SD)		Covariate Effects MSE(SD)	
	hc-ICA	Dual.Reg.	hc-ICA	Dual.Reg.
Low				
N=10	0.993 (0.003)	0.973 (0.027)	0.115 (0.054)	0.489 (0.160)
N=20	0.995 (0.001)	0.980 (0.010)	0.067 (0.011)	0.374 (0.112)
N=40	0.995 (0.001)	0.988 (0.009)	0.041 (0.005)	0.358 (0.107)
High				
N=10	0.917 (0.055)	0.868 (0.083)	1.013 (0.430)	1.998 (0.873)
N=20	0.935 (0.041)	0.887 (0.067)	0.656 (0.332)	1.504 (0.809)
N=40	0.974 (0.037)	0.940 (0.041)	0.408 (0.164)	1.142 (0.581)

Table 2.5: Checking the stability of EM algorithm in real data using the resulting correlations between pairs form 50 different initialization (IQR: interquantile range; Q_1: the 25th percentile; Q_3: the 75th percentile)

	Mixing matrices	Group IC maps	Covariate effects
Median	0.997	0.993	0.987
IQR	0.003	0.005	0.010
(Q_1, Q_3)	(0.996, 0.999)	(0.991, 0.996)	(0.982, 0.992)

Chapter 3

Bayesian Spatial Feature Selection for Massive Neuroimaging Data via Thresholded Gaussian Processes

This chapter is joint work with Dr. Jian Kang.

3.1 Feature selection within the spatially varying coefficient functions

We start with general notations and definitions. Denote by $\mathbb{R}^p$ a p-dimensional real Euclidean space for any $p \geqslant 1$. For any $\boldsymbol{\beta} \in \mathbb{R}^p$, write $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, define $\|\boldsymbol{\beta}\|_\infty = \max_{1 \leqslant k \leqslant p} |\beta_k|$, $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^p |\beta_k|$ and $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{k=1}^p \beta_k^2}$. Denote by $\mathcal{R} \subset \mathbb{R}^d$ a compact subset of the standard brain space ($d = 2$ or 3). Let $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{R}$ be a set of spatial locations where brain signals are measured. An empirical measure on $\mathcal{R}$ induced by $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is defined as $P_n(d\mathbf{s}) = \frac{1}{n} \sum_{i=1}^n I[\mathbf{s}_i \in d\mathbf{s}]$, where the indicator function $I[\mathcal{A}] = 1$ if event $\mathcal{A}$ occurs, $I[\mathcal{A}] = 0$, otherwise. For a scalar-valued function $\beta(\cdot) :$

$\mathcal{R} \mapsto \mathbb{R}$, define $\|\beta(\cdot)\|_\infty = \sup_{\mathbf{s}\in\mathcal{R}}|\beta(\mathbf{s})|$ and $\|\beta(\cdot)\|_1 = \int_{\mathbf{s}\in\mathcal{R}}|\beta(\mathbf{s})|P_n(d\mathbf{s})$. For a p-dimensional vector-valued function $\boldsymbol{\beta}(\mathbf{s}) = [\beta_1(\mathbf{s}), \ldots, \beta_p(\mathbf{s})]^T : \mathcal{R} \mapsto \mathbb{R}^p$, define $\|\boldsymbol{\beta}(\cdot)\|_{1,\infty} = \max_{1\leqslant k\leqslant p}\|\beta_k(\mathbf{s})\|_1$. Denote by $\mathcal{C}(\mathcal{R})$ a collection of all the continuous functions defined on $\mathcal{R}$. Let $D^{\boldsymbol{\alpha}}\beta$ be a partial derivative operator on function $\beta(\cdot)$ (given its existence) which is given by $\frac{\partial^{\|\boldsymbol{\alpha}\|_1}\beta}{\partial s_1^{\alpha_1}\cdots\partial s_d^{\alpha_d}}$ for $\boldsymbol{\alpha} \in \mathbb{R}^p$. Denote by $\mathcal{C}^\rho(\mathcal{R})$ a set of functions β defined on $\mathcal{R}$ with continuous partial derivatives $D^{\boldsymbol{\alpha}}\beta$ for all $\boldsymbol{\alpha}$ such that $\|\boldsymbol{\alpha}\|_1 \leqslant \rho$.

3.1.1 The spatially varying coefficient model for neuroimaging data

Suppose the data set consists of m subjects. For each subject j, let $y_j(\mathbf{s})$ be the brain signal measured from a certain imaging modality at location $\mathbf{s} \in \mathcal{R}$; and there are also p covariates are collected, denoted $\mathbf{x}_j = [x_{j1}, \ldots, x_{jp}]^T$, for $j = 1, \ldots, m$. The spatially varying coefficient model (SVCM) for neuroimaging data is given by

$$y_j(\mathbf{s}) = \mathbf{x}_j^T\boldsymbol{\beta}^0(\mathbf{s}) + e_j(\mathbf{s}), \tag{3.1}$$

in which elements of $\boldsymbol{\beta}^0(\mathbf{s}) = [\beta_1^0(\mathbf{s}), \ldots, \beta_p^0(\mathbf{s})]^T$ are the spatially varying coefficient functions (SVCFs) defined on the brain space $\mathcal{R}$. It characterizes associations between covariates and imaging outcomes: $\beta_k^0(\mathbf{s})$ $(k = 1, \ldots, p)$ quantifies the effects of the kth covariate at brain location $\mathbf{s}$. The zero-mean error process $e_j(\mathbf{s})$ is assumed to be spatially independent across the whole brain domain for each subject, conditional on a variance process $\sigma^2(\mathbf{s})$. Specifically in our model, we assume that $e_j(\mathbf{s}) \overset{\text{iid}}{\sim} N(0,\ \sigma^2(\mathbf{s}))$.

For neuroimaging data from commonly used imaging modalities, only outcomes at a number of locations, $\mathbf{s}_1$, ..., $\mathbf{s}_n$, are observed. At these locations, the SVCM proposed in (3.1) for the recorded neuroimaging data can be expressed as

$$[\mathbf{y}(\mathbf{s}_i) \mid \boldsymbol{\beta}^0(\mathbf{s}_i),\ \sigma^2(\mathbf{s}_i)] \sim N(\mathbf{X}\boldsymbol{\beta}(\mathbf{s}_i),\ \ \sigma^2(\mathbf{s}_i)\mathbf{I}_m) \tag{3.2}$$

independently for all $i = 1, \ldots, n$, in which $\mathbf{y}(\mathbf{s}_i) = [y_1(\mathbf{s}_i), \ldots, y_m(\mathbf{s}_i)]^T$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]^T$ and $\boldsymbol{e}(\mathbf{s}_i) = [e_1(\mathbf{s}_i), \ldots, e_m(\mathbf{s}_i)]^T$. For simplicity, we denote by $\mathcal{Y} = \{\mathbf{y}(\mathbf{s}_i)\}_{i=1}^n$ an $m \times n$ matrix recoding all the neuroimaging outcomes involved in the study.

In neuroimaging studies, there exists a natural region partition of the whole brain domain $\mathcal{R}$ into bounded connected sets $\mathcal{R}_1, ..., \mathcal{R}_G$ with non-empty interiors, such that $\mathcal{R} = \cup_{g=1}^G \mathcal{R}_g$, $\mathcal{R}_g \cap \mathcal{R}_{g'} = \emptyset$, $\forall g \neq g'$. In many cases, one can treat $\mathcal{R}_g$ as neuroanatomical regions from commonly used labeling systems such as the Automated Anatomical Labeling (AAL) (Tzourio-Mazoyer et al., 2002). For region of interest (ROI) based analysis, each ROI is one parcellated region. For seed-based region-level analysis, $\mathcal{R}_g$ can be regarded as the clusters showing strong functional connectivities based on preliminary results. In some voxelwise analysis with no regional information to be incorporated, we can simply consider each voxel (a 3D cubic) as a region and the centers of voxels as observed brain locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$.

To utilize the proposed SVCM for analyzing neuroimaging data and selecting features, we state our assumptions for the SVCF in model (3.1). Specifically, we work with region-wise smooth (ρ-times continuously differentiable to be accurate) functions with structured sparsity, which can be mathematically expressed as follows: a varying coefficient function in model (3.1), say $\beta^0(\mathbf{s})$ (omitting subscripts $k = 1, ..., p$ here for the ease of exposition), must satisfy that:

(C1) there exists an index set $\mathcal{I}_1 \subset \{1, ..., G\}$, such that $\beta^0(\mathbf{s}) \times I[\mathbf{s} \in \overline{\mathcal{R}}_g] \in \mathcal{C}^\rho(\overline{\mathcal{R}}_g)$ where $\overline{\mathcal{R}}_g$ is the closure of $\mathcal{R}_g$ for all $g \in \mathcal{I}_1$ with $\rho = \left[\frac{d}{2}\right] + 1$;

(C2) $\beta^0(\mathbf{s})$ is bounded away from zero on any $\mathcal{R}_g$ for all $g \in \mathcal{I}_1$, that is,

$$\lambda^0 = \inf_{\mathbf{s} \in \cup_{g \in \mathcal{I}_1} \mathcal{R}_g} |\beta^0(\mathbf{s})| > 0;$$

(C3) let $\mathcal{I}_0 = \{1, ..., G\} \backslash \mathcal{I}_1$, then $\beta^0(\mathbf{s}) = 0$ for all $\mathbf{s} \in \cup_{g \in \mathcal{I}_0} \mathcal{R}_g$.

Based on the definitions above, we introduce, for each brain region, an binary feature selection indicator function $r_g^0(\mathbf{s}) = I[\mathbf{s} \in \mathcal{R}_g] \times I[g \in \mathcal{I}_1]$ which reflects the existence of features within $\mathcal{R}_g$. We call the vector-valued function $\mathbf{r}^0(\mathbf{s}) = [r_1^0(\mathbf{s}), ..., r_G^0(\mathbf{s})]^T$, the range of which is $\{0,1\}^G$, the selection indicator function of $\beta(\mathbf{s})$.

The conditions on the SVCFs can be interpreted as follows: (C1) requires that the functions must be smooth within the closure of each brain region, which implies more homogeneous covariate effects at the regional level; (C2) indicates that jump discontinuities and sharp edge effects exist at the boundaries of brain regions demonstrating features of interest; (C3) introduces sparsity into each SVCF in model (3.1) and restricts the sparsity structure by a prespecified region partition.

Throughout the rest of this chapter, the notation $\mathcal{P}(\mathcal{R}_1, ..., \mathcal{R}_G)$ ($\mathcal{P}$ in short) stands for a set of functions defined on $\mathcal{R}$ with a known partition satisfying (C1)–(C3); in the same vein, $\mathbf{P}(\mathcal{R}_1, ..., \mathcal{R}_G)$ ($\mathbf{P}$ in short) represents a set of p-dimensional vector-valued functions, i.e., $\mathbf{P} = \{\boldsymbol{\beta}^0(\mathbf{s}) = [\beta_1^0(\mathbf{s}), \ldots, \beta_p^0(\mathbf{s})]^T : \beta_k^0(\mathbf{s}) \in \mathcal{P}, k = 1, \ldots, p\}$.

3.1.2 The thresholded Gaussian process priors

Construction of the prior

A Gaussian process (GP) can be regarded as a probabilistic measure on certain functional spaces, making it as popular prior models in Bayesian nonparametric data analysis. In general, the GP prior, denoted by $\mathcal{GP}[\mu(\cdot), C(\cdot,\cdot)]$, is determined by its mean function $\mu : \mathcal{R} \mapsto \mathbb{R}$ and the covariance kernel function $C : \mathcal{R} \times \mathcal{R} \mapsto \mathbb{R}$. A draw $\beta(\cdot) \sim \mathcal{GP}[\mu(\cdot), C(\cdot,\cdot)]$ is a function defined on $\mathcal{R}$ such that any finite collection of its function values are jointly multivariate Gaussian. To be specific, for any choices of $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{R}$, $[\beta(\mathbf{s}_1), \ldots, \beta(\mathbf{s}_n)]^T \sim N(\boldsymbol{\mu}, \mathbf{C})$ with $\boldsymbol{\mu} = [\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n)]^T$ and $\mathbf{C} = \{C(\mathbf{s}_i, \mathbf{s}_j)\}_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n}$. The boundedness and smoothness of random functions generated from GP are determined through the covariance kernel function. Typical choices for the covariance kernel functions include but not limited to the rational

quadratic kernel, Matérn class of kernels, the square exponential kernel (Rasmussen and Williams, 2006a).

To enable detailed feature selections within the SVCFs $\beta_k^0(\mathbf{s}) \in \mathcal{P}$ $(k = 1, \ldots, p)$ in model (3.1), we develop the thresholded Gaussian process (TGP) prior, which can be represented as follows: $\beta(\mathbf{s}) \sim \mathcal{TGP}[\tau^2,\ \theta^2,\ \lambda,\ \kappa(\cdot,\cdot)]$ implies that

$$\beta(\mathbf{s}) = \tilde{\beta}(\mathbf{s}) I_\lambda[\tilde{\beta}(\mathbf{s})], \tag{3.3}$$

$$\tilde{\beta}(\mathbf{s}) = \gamma(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{3.4}$$

$$\gamma(\mathbf{s}) \sim \mathcal{GP}[0, \tau^2 \kappa(\mathbf{s}, \mathbf{s}')], \tag{3.5}$$

$$\epsilon(\mathbf{s}) \sim \mathcal{GP}[0, \theta^2 \chi(\mathbf{s}, \mathbf{s}')] \tag{3.6}$$

for all $\mathbf{s} \in \mathcal{R}$, where

$$I_\lambda[\tilde{\beta}(\mathbf{s})] = \sum_{g=1}^{G} I\left[\mathbf{s} \in \mathcal{R}_g : \inf_{s \in \mathcal{R}_g} |\tilde{\beta}(\mathbf{s})| > \lambda\right] \tag{3.7}$$

is the thresholding function that generate region-wise sparse features; $\lambda > 0$ is the thresholding parameter; $\tau^2 > 0$ and $\theta^2 > 0$ are variance parameters in the GPs; $\kappa(\cdot,\cdot) : \mathcal{R} \times \mathcal{R} \mapsto \mathbb{R}$ is a kernel correlation function and $\chi(\mathbf{s}, \mathbf{s}')$ is constructed from κ as $\chi(\mathbf{s}, \mathbf{s}') = \sum_{g=1}^{G} \kappa(\mathbf{s}, \mathbf{s}') \times I[\mathbf{s}, \mathbf{s}' \in \mathcal{R}_g]$. The thresholding construction in (3.7) is motivated by condition (C2) defining the SVCF. It naturally leads to the definition of selection indicator processes: $\mathbf{r}_g(\mathbf{s}) = [r_1(\mathbf{s}), ..., r_G(\mathbf{s})]^T$ in which $r_g(\mathbf{s}) = I\left[\mathbf{s} \in \mathcal{R}_g : \inf_{s \in \mathcal{R}_g} |\tilde{\beta}(\mathbf{s})| > \lambda\right]$. As a result, (3.3) is equivalent to $\beta(\mathbf{s}) = \tilde{\beta}(\mathbf{s}) \sum_{g=1}^{G} r_g(\mathbf{s})$.

For voxel-wise analysis without regional information, the thresholding function in (3.7) can be simplified as $I_\lambda[\tilde{\beta}(\mathbf{s})] = I[|\tilde{\beta}(\mathbf{s})| > \lambda]$, where $\mathbf{s}$ denotes the center of a voxel. The GP $\tilde{\beta}(\mathbf{s})$ in our prior is a combination of one “global" GP, $\gamma(\mathbf{s})$, which captures the general dependence structures across the whole brain domain and one “local" GP, $\epsilon(\mathbf{s})$, reflecting the dependence and variabilities within each parcellated brain region. This construction can also naturally generate jumping discontinuities on

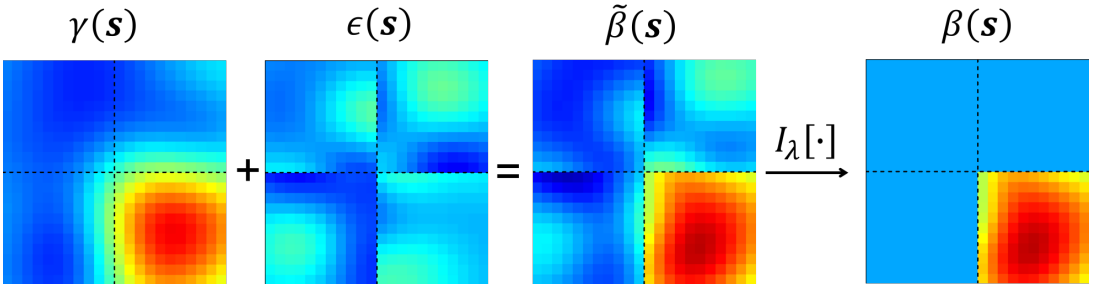


Figure 3.1: Sampling SVCFs from the TGP prior

the boundaries of the brain regions.

We illustrate the procedure to sample the SVCF from our TGP priors on a two-dimensional square region in Figure 3.1, where the dashed lines parcellate the whole region into four equally spaced sub-regions. $\gamma(s)$ is smooth over the whole region; $\epsilon(s)$ is smooth within each sub-region but has distinct jumps on the boundaries. The summation of these two GPs is thresholded by λ to generate SVCFs. We also summarize some key information regarding the SVCFs and our TGP prior in Table 3.1 for comparison.

3.2 Theoretical Results

We first introduce two sets of extra conditions in addition to the conditions (C1)-(C3) for the SVCFs. The design matrix X as defined in (3.2) satisfies:

(X1) Let d_{min} and d_{max} be the smallest and largest eigenvalues of $\frac{1}{m}X^TX$, then $0 < d_{min} < d_{max} < \infty$.

For the kernel correlation functions $\kappa(\cdot,\cdot)$ in our proposed TGP priors, we introduce the following condition:

(K1) $\kappa(s, s') = \prod_{j=1}^{d} K_j(\|s_j - s'_j\|)$ for some nowhere zero, continuous, symmetric density function (up to a normalization constant) K_j defined on $\mathbb{R}$.

(K2) $\kappa(s,\cdot)$ has continuous partial derivates up to order $2\rho + 2$ where $\rho = \lfloor \frac{d}{2} \rfloor + 1$.

Table 3.1: Summary of the true SVCFs and the TGP prior

$\beta^0(\mathbf{s})$	$\beta(\mathbf{s})$
SVCF $\beta^0(\mathbf{s})$ satisfying (C1)–(C3): a specific function in $\mathcal{P}$	**TGP prior $\beta(\mathbf{s}) \sim \mathcal{TGP}(\tau^2, \theta^2, \lambda, \kappa(\cdot,\cdot))$:** a stochastic process defined on $\mathcal{R}$
region-wise smoothness: $\beta^0(\mathbf{s})$ is smooth within each region	**region-wise dependence :** region-wise dependence structures imposed on $\beta(\mathbf{s})$
(multivariate) selection indicator function: $\mathbf{r}^0(\mathbf{s}) = [r_1^0(\mathbf{s}), ..., r_G^0(\mathbf{s})]^T \in \{0,1\}^G$, where $r_g^0(\mathbf{s}) = I[\mathbf{s} \in \mathcal{R}_g] \times I[g \in \mathcal{I}_1]$	**selection indicator processes:** $\mathbf{r}(\mathbf{s}) = [r_1(\mathbf{s}), ..., r_G(\mathbf{s})]^T \in \{0,1\}^G$, where $r_g(\mathbf{s}) = I\left[\mathbf{s} \in \mathcal{R}_g : \inf_{s \in \mathcal{R}_g} \lvert\tilde{\beta}(\mathbf{s})\rvert > \lambda\right]$
a positive real number $\lambda^0 > 0$: infimum of $\lvert\beta^0(\mathbf{s})\rvert$ on its support	**a parameter in the TGP prior $\lambda > 0$:** thresholding the GPs to control structured sparsity

Theorem 2. *Let* $\beta^0(\mathbf{s}) \in \mathcal{P}$ *be an arbitrary SVCF satisfying conditions (C1)-(C3) and suppose the partition number* $G < \infty$. *If* $0 < \lambda < \lambda^0,\ 0 < \tau^2, \theta^2 < \infty$ *and the kernel function* κ *satisfies (K1), then the proposed prior*

$$\beta(\mathbf{s}) \sim \mathcal{TGP}[\tau^2,\ \theta^2,\ \lambda,\ \kappa(\cdot,\cdot)],$$

given in (3.3)-(3.6) *satisfies that*

$$\Pi\left(\|\beta(\mathbf{s}) - \beta^0(\mathbf{s})\|_\infty < \varepsilon\right) > 0, \qquad \textit{for all}\ \ \varepsilon > 0.$$

Theorem 2 demonstrates that the proposed TGP prior assign positive measures to arbitrarily small neighborhoods of all elements within $\mathcal{P}$, the family of spatially varying coefficient functions defined in our model (3.1). This property is essential, especially for Bayesian nonparametric priors, since it is necessary for appropriate posterior behaviors and can not be guaranteed in many cases.

Theorem 3. *Let* $\varepsilon > 0$ *be an arbitrary positive number. Suppose that our observed data* $\mathbf{y}(\mathbf{s}_1), ..., \mathbf{y}(\mathbf{s}_n)$ *are independently generated from model* (3.2) *satisfying the following conditions*

(a) $p < m$ *and* $G < \infty$*;*

(b) all the SVCFs in model (3.2) *satisfies (C1)-(C3), i.e.,* $\boldsymbol{\beta}^0(\mathbf{s}) \in \boldsymbol{P}$*;*

(c) the design matrix $\mathbf{X}$ *satisfies condition (X1);*

(d) the variance function $\sigma^2(\mathbf{s})$ *satisfies* $\inf_{\mathbf{s}\in\mathcal{R}} \sigma^2(\mathbf{s}) > 0$ *and* $\frac{1}{m}\sup_{\mathbf{s}\in\mathcal{R}} \sigma^2(\mathbf{s}) < \frac{\varepsilon^2 d_{\min}}{8\log 2}$.

If we assign a distinct TGP prior independently for each SVCF in model (3.2) *with kernel functions satisfying both (K1), (K2) and other conditions in Theorem 2, then the posterior distribution of* $\beta(\mathbf{s})$ *satisfies that*

$$\Pi\left[\ U_\varepsilon \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\ \right] \to 1$$

as $n \to \infty$ *in* $\mathbb{P}^n_{\beta^0}$ *probability, where* $U_\varepsilon = \{\beta(s) \in \boldsymbol{P} : \|\beta(s) - \beta^0(s)\|_{1,\infty} < \varepsilon\}$.

Theorem 3 justifies the posterior consistency of the proposed TGP prior given model (3.1) under the infill asymptotic framework. It implies that, if a ground truth of the SVCFs exists and the data is generated accordingly, then the posterior distribution of $\beta(s)$ can be concentrated to an arbitrarily small $\|\cdot\|_{1,\infty}$ neighborhood around the truth as the number of spatial locations goes to infinity. The conditions of this theory also imply that a small ratio between the number of subjects and the variance of pure noise, i.e., $\frac{\sup_{s \in \mathcal{R}} \sigma^2(s)}{m}$, is also important to guarantee a good performance of our method. One limitation of Theorem 3 is that it does not apply to the voxel level analysis where $G = n \to \infty$. However, this type of analysis generally works well empirically.

Although the $\|\cdot\|_{1,\infty}$ norm is not common in Bayesian asymptotic literatures, a direct result based on Theorem 3 is the element-wise posterior consistency for $\beta(s)$ under the commonly used empirical $\|\cdot\|_1$ norm, e.g., Ghosal et al. (2006), for a fixed design of spatial locations s_i. This results from the independent TGP prior assignment for each functional element of $\beta(s)$.

Corollary 1. *Let* $\varepsilon > 0$ *be an arbitrary positive number. Under the same assumptions and conditions in Theorem 3, the posterior distributions satisfies that, for all* $k = 1, \ldots, p$*,*

$$\Pi[\, U_{\varepsilon,k} \mid y(s_1), \ldots, y(s_n)\,] \to 1$$

as $n \to \infty$ *in* $\mathbb{P}^n_{\beta^0}$ *probability, where* $U_{\varepsilon,k} = \{\beta(s) \in \mathcal{P} : \|\beta(s) - \beta^0_k(s)\|_1 < \varepsilon\}$.

Of note, in neuroimaging studies, s_i are usually fixed 3D grid points, thus we do not consider the $\|\cdot\|_1$ norms with regard to random measures for s when proving posterior consistency. For a fixed design within a finite domain $\mathcal{R}$ (the volume of brain is limited), another useful direction is to show posterior consistency under the $\|\cdot\|_1$ norm with regard to the Lebesgue measure. We see this as a potential extension to the theory de-

velopment in the future work.

Theorem 4. *We assume the conditions in Theorem 3 hold. For any specific SVCF* $\beta^0(s) \in \mathcal{P}$ *(dropping the subscript* k*) in model* (3.2)*, let* $\mathbf{r}^0(s)$ *be its selection indicator function; for the prior* $\beta(s) \sim \mathcal{TGP}[\tau^2,\ \theta^2,\ \lambda,\ \kappa(\cdot,\cdot)]$*, let* $\mathbf{r}(s)$ *be the resulting selection indicator processes, then*

$$\Pi\left[\ \mathbf{r}(s) = \mathbf{r}^0(s) \mid \mathbf{y}(s_1), \ldots, \mathbf{y}(s_n)\ \right] \to 1$$

as $n \to \infty$ *in* $\mathbb{P}^n_{\beta^0}$ *probability, where* $\mathbf{r}(s) = \mathbf{r}^0(s)$ *means* $r^0_g(s) = r_g(s)$ *for all* $g = 1, ..., G$.

Theorem 4 establishes the posterior feature selection consistency, at the regional level, for assigning TGP priors on the SVCFs defined by (C1)-(C3) in our working model. This provides theoretical justifications for selecting important brain regions of interest (ROIs) using our method.

3.3 Posterior Inferences

3.3.1 Model Representation

Now consider the SVCM defined in (3.1). For the p-dimensional multivariate spatially varying coefficient function, $\boldsymbol{\beta}(s) = [\beta_1(s), ..., \beta_k(s)]^T$, we assume that

$$\beta_k(s) \sim \mathcal{TGP}[\tau_k^2, \theta^2, \lambda_k, \kappa(\cdot,\cdot)],$$

with $\kappa(\cdot,\cdot)$ being a smooth kernel function. This specification implies that the global processes (3.5) have distinct flexible variance parameters τ_k^2, while the local fluctuation processes (3.6) have a small fixed marginal variance θ^2.

Based on the prior specification for $\boldsymbol{\beta}(s)$, we have that $\beta_k(s) = \tilde{\beta}_k(s) I_{\lambda_k}[\tilde{\beta}_k(s)]$ for $k = 1, \ldots, p$, where

$$\tilde{\beta}_k(s) = \gamma_k(s) + \epsilon_k(s), \tag{3.8}$$

where $\gamma_k(s) \sim \mathcal{GP}[0, \tau_k^2 \kappa(s, s')]$ and $\epsilon_k(s) \sim \mathcal{GP}\left(0, \sum_{g=1}^G \theta^2 \kappa(s, s') \times I[s, s' \in R_g]\right)$.

For global GPs: $\gamma_k(\mathbf{s})$ in (3.8), its Karhunen-Loève (KL) expansion can be expressed as

$$\gamma_k(\mathbf{s}) = \sum_{l=1}^{\infty} \varphi_l(\mathbf{s}) u_{kl}, \tag{3.9}$$

where $u_{kl} \sim N(0, \tau_k^2 \zeta_l)$ independently with $\zeta_l > 0$ such that $\sum_{l=1}^{\infty} \zeta_l \varphi_l(\mathbf{s}) \varphi_l(\mathbf{s}') = \kappa(\mathbf{s}, \mathbf{s}')$ and that $\int \varphi_l(\mathbf{s}) \varphi_{l'}(\mathbf{s}) d\mathbf{s} = 0, \forall l \neq l'$ based on the Mercer's Theorem (Rasmussen and Williams, 2006a). In practice, we truncate the infinite sum in (3.9) into L terms such that $\sum_{l=1}^{L} \zeta_l / \sum_{l=1}^{\infty} \zeta_l$ is close to 1.

The decomposition of global GP in (3.9) implies a model representation of the proposed SVCM along with the TGP prior specifications. To be more specific, the neuroimaging signal $y_j(\mathbf{s})$ on locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ can be modeled through latent mGPs: $\tilde{\boldsymbol{\beta}}(\mathbf{s}) = [\tilde{\beta}_1(\mathbf{s}), \ldots, \tilde{\beta}_p(\mathbf{s})]^T$ and the truncated KL expansion coefficients $\{\mathbf{u}_k\}_{k=1}^p$ with $\mathbf{u}_k = [u_{k1}, ..., u_{kL}]^T$ by integrating out the local GPs $\epsilon_k(\mathbf{s})$ in (3.8), which is given by

$$[y_j(\mathbf{s}_i) \mid \tilde{\boldsymbol{\beta}}(\mathbf{s}_i),\ \sigma^2(\mathbf{s}_i)] \sim N\left(\mathbf{x}_j^T g_{\boldsymbol{\lambda}}[\tilde{\boldsymbol{\beta}}(\mathbf{s}_i)],\ \sigma^2(\mathbf{s}_i)\right), \tag{3.10}$$

$$[\{\tilde{\beta}_k(\mathbf{s}_i)\}_{\mathbf{s}_i \in \mathcal{R}_g} \mid \mathbf{u}_k] \sim N\left(\boldsymbol{\varphi}_g \mathbf{u}_k,\ \theta^2 K_g\right), \tag{3.11}$$

$$u_{kl} \sim N(0, \zeta_l \tau_k^2), \tag{3.12}$$

for $j = 1, \ldots, m,\ i = 1, \ldots, n,\ k = 1, \ldots, p,\ g = 1, \ldots, G$ and $l = 1, \ldots, L$, where $N(\mu, \sigma^2)$ represents a normal distribution with mean μ and variance σ^2, $\boldsymbol{\varphi}_g = \{\boldsymbol{\varphi}(\mathbf{s}_i)^T\}_{\mathbf{s}_i \in \mathcal{R}_g}$ with $\boldsymbol{\varphi}(\mathbf{s}_i)$ representing $[\varphi_1(\mathbf{s}_i), \ldots, \varphi_L(\mathbf{s}_i)]^T$ and $K_g = \{\kappa(\mathbf{s}_i, \mathbf{s}_{i'})\}_{\mathbf{s}_i, \mathbf{s}_{i'} \in \mathcal{R}_g}$ being a correlation matrix. The p-dimensional vector value functional operator $g_{\boldsymbol{\lambda}}(\cdot)$ is defined on the domain of all functions in $\mathcal{P}$, which is given by

$$g_{\boldsymbol{\lambda}}[\tilde{\boldsymbol{\beta}}(\mathbf{s})] = \left[\tilde{\beta}_1(\mathbf{s}) I_{\lambda_1}[\tilde{\beta}_1(\mathbf{s})], \ldots, \tilde{\beta}_p(\mathbf{s}) I_{\lambda_p}[\tilde{\beta}_p(\mathbf{s})]\right]^T,$$

with $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]^T$.

3.3.2 Hyper Prior Specifications

We assign conjugate priors to variance parameters τ_k^2, i.e., $\tau_k^2 \overset{iid}{\sim}$ Inv-Ga(ν, w), in which Inv-Ga(ν, w) represents an inverse gamma distribution with shape ν and rate w. We fix $\theta^2 = 1$ to restrict local deviations to a relatively small scale, especially for the case that a small number of observations are recorded in each region. For the variance process $\sigma^2(\mathbf{s})$, we assign a log-Gaussian process prior to capture the its stochastic volatility and spatial correlations by assuming: $\log\left[\sigma^2(\mathbf{s})\right] \sim \mathcal{GP}(0, \xi^2\kappa_\sigma(\cdot,\cdot))$.

We develop a data-driven method to specify the prior of thresholding parameters $\lambda_k, k = 1, \ldots, p$. We consider the log full conditional of λ_k which is given by

$$\log \pi[\lambda_k \mid \mathcal{Y}, \tilde{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_{-k}, \sigma^2] = \ell[\lambda_k; \mathcal{Y}, \tilde{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_{-k}, \sigma^2] + C,$$

where C is a constant, $\tilde{\boldsymbol{\beta}}_k = [\tilde{\beta}_k(\mathbf{s}_1), \ldots, \tilde{\beta}_k(\mathbf{s}_n)]^T$, $\boldsymbol{\beta}_k = [\beta_k(\mathbf{s}_1), \ldots, \beta_k(\mathbf{s}_n)]^T$, $\boldsymbol{\beta}_{-k} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}_{k+1}, \ldots, \boldsymbol{\beta}_p]^T$, and

$$\ell(\lambda_k) := \ell[\lambda_k; \mathcal{Y}, \tilde{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_{-k}, \sigma^2] = \sum_{i=1}^{n} \omega_k(\mathbf{s}_i) I_{\lambda_k}[\tilde{\beta}_k(\mathbf{s}_i)]/\sigma^2(\mathbf{s}_i), \tag{3.13}$$

with $\omega_k(\mathbf{s}) = \sum_{j=1}^{m} \tilde{\beta}_k(\mathbf{s}) x_{jk} \left[2y_{j,-k}(\mathbf{s}) - \tilde{\beta}_k(\mathbf{s}) x_{jk}\right]$ and $y_{j,-k}(\mathbf{s}) = y_j(\mathbf{s}) - \sum_{j' \neq k} x_{jj'} \beta_{j'}(\mathbf{s})$. The function $\ell(\lambda_k)$ is flat when λ_k is around zero and dramatically decreases when λ_k is greater a certain value, to which we refer as a "turning point". It should be close to the true threshold. Figure 3.2 shows the profiles of $\ell(\lambda_k)$ for a model with three SVCFs on a space of 900 locations from 50 simulated datasets. The true thresholds $\lambda_k = k + 1$ for $k = 1, 2, 3$. The turning points in the profiles of $\ell(\lambda_k)$ are all around the true thresholds.

Based on the profile of $\ell(\lambda_k)$, we can specify the priors of λ_k according to $\ell(\lambda_k)$. In practice, we need to provide rough estimates of $\tilde{\boldsymbol{\beta}}_k$ and $\boldsymbol{\beta}_{-k}$ in order to evaluate $\ell(\lambda_k)$ before posterior inferences. We consider an SVCM with smoothed SVCFs approximated by the truncated K-L expansion, where we compute the ordinary least squares (OLS) of the coef-

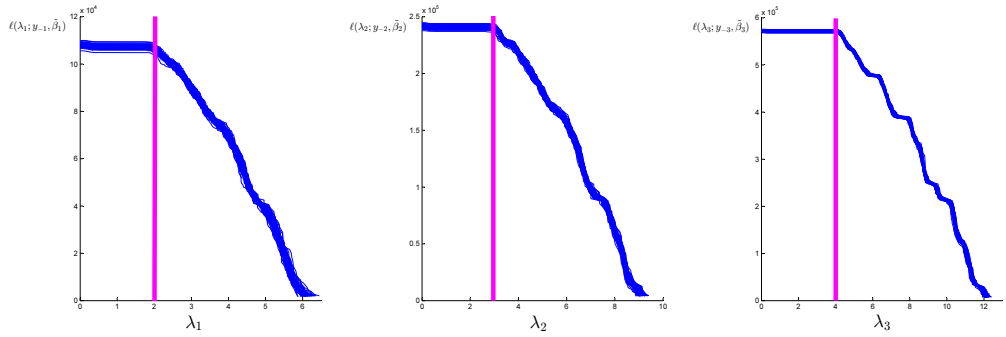


Figure 3.2: Simulated $\widehat{\ell}(\lambda_k)$ from 50 synthetic datasets: ground truth ($\lambda_1 = 2, \lambda_2 = 3, \lambda_3 = 4$) are marked in the figures.

ficients, i.e.

$$\{\widehat{w}_{lk}\}_{l=1\,k=1}^{L\quad p} = \arg\min_{\{w_{lk}\}} \sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_j(\mathbf{s}_i) - \sum_{k=1}^{p}\sum_{l=1}^{L} x_{jk}\varphi_l(\mathbf{s}_i)w_{lk}\right)^2.$$

Then both $\tilde{\beta}_k(\mathbf{s})$ and $\beta_k(\mathbf{s})$ can be approximated by $\widehat{\beta}_k(\mathbf{s}) = \sum_{l=1}^{L}\varphi_l(\mathbf{s})\widehat{w}_{lk}$; the variances $\sigma^2(\mathbf{s}_i)$ can be approximated by

$$\widehat{\sigma}^2(\mathbf{s}_i) = \frac{1}{n}\sum_{i=1}^{n}\left(y_j(\mathbf{s}_i) - \sum_{k=1}^{p}\sum_{l=1}^{L} x_{jk}\widehat{\beta}_k(\mathbf{s}_i)\right)^2.$$

Thus, we replace $\boldsymbol{\beta}_k$ and $\tilde{\boldsymbol{\beta}}_{-k}$ by $\widehat{\boldsymbol{\beta}}_k$ and $\widehat{\boldsymbol{\beta}}_{-k}$, respectively, in (3.13). Write $\widehat{\ell}(\lambda_k) = \ell(\lambda_k; y, \widehat{\boldsymbol{\beta}}_k, \widehat{\boldsymbol{\beta}}_{-k}, \widehat{\sigma}^2)$.

We propose to assign uniform priors to λ_k, i.e. $\lambda_k \sim \mathrm{Unif}(c_k - h_k, c_k + h_k)$, where the half range h_k and center c_k can be determined based on the profile of $\widehat{\ell}(\lambda_k)$. More specifically, we evaluate $\widehat{\ell}(\lambda_k)$ on a set of grid points $\{\lambda_k^{(1)}, \ldots, \lambda_k^{(G)}\}$, denoted $\{\ell_k^{(1)}, \ldots, \ell_k^{(G)}\}$. Given an interval (a, b), define the sample correlation between λ_k and $\widehat{\ell}(\lambda_k)$ within (a, b) as

$$\widehat{\rho}(a, b) = \frac{\sum_{\lambda_k^{(g)}\in(a,b)}(\lambda_k^{(g)} - \overline{\lambda}_k)(\ell_k^{(g)} - \overline{\ell}_k)}{\sqrt{\sum_{\lambda_k^{(g)}\in(a,b)}(\lambda_k^{(g)} - \overline{\lambda}_k)^2}\sqrt{\sum_{\lambda_k^{(g)}\in(a,b)}(\ell_k^{(g)} - \overline{\ell}_k)^2}}$$

with $\overline{\lambda}_k = \sum_{\lambda_k^{(g)}\in(a,b)}\lambda_k^{(g)}/M(a, b)$, $\overline{\ell}_k = \sum_{\lambda_k^{(g)}\in(a,b)}\ell_k^{(g)}/M(a, b)$ and $M(a, b) =$

$\sum_{g=1}^{G} I[\lambda_k^{(g)} \in (a, b)]$. And define

$$\widetilde{c}_k(h) = \min\{\lambda_k^{(g)} : |\widehat{\rho}(\lambda_k^{(g)} - h, \lambda_k^{(g)} + h)| > \zeta_k\},$$

where ζ_k is determined by the rejection region of Pearson correlation test. Given $h > 0$, $\widetilde{c}_k(h)$ represents the location that $\widehat{\ell}(\lambda_k)$ and λ_k have no significant correlation. Then we specify

$$h_k = \min\{h : |\widehat{\rho}(\widetilde{c}_k(h) - h, \widetilde{c}_k(h) + h)| > \zeta_k\}, \quad \text{and} \quad c_k = \widetilde{c}_k(h_k).$$

This leads to an informative prior range $[h_k - c_k, h_k + c_k]$ for λ_k with a high probability to cover the turning point of $\ell(\lambda_k)$.

3.3.3 Kernel Expansion for Massive Data Analysis

Theoretically, the KL expansion for the GP with kernel function $\kappa(\cdot, \cdot)$ relies on solving the integral equation $\int \kappa(\mathbf{s}, \mathbf{s}')\varphi(\mathbf{s})d\mathbf{s} = \zeta_l \varphi_(\mathbf{s}')$, which might not admit analytical solutions. Empirically, the expansion is often achieved by calculating the eigenvalues and eigenvectors of the $n \times n$ correlation matrix $K_n = \{\kappa(\mathbf{s}_i, \mathbf{s}_{i'})\}_{1 \leqslant i, i' \leqslant n}$ on a set of pre-specified locations. However, in the analysis of massive neuroimaging data that can involve a very large number (n can be hundreds of thousands) of brain locations, it is computationally infeasible to perform eigen decompositions on K_n. To solve this issue, we introduce the modified square exponential kernel

$$\kappa(\mathbf{s}, \mathbf{s}') = \exp\{-a\|\mathbf{s}\|_2^2 - a\|\mathbf{s}'\|_2^2 - b\|\mathbf{s} - \mathbf{s}'\|_2^2\}, \ a, b > 0 \tag{3.14}$$

with a relatively small value for a as a numerical approximation to the square exponential kernel when dealing with massive neuroimaging data. The major benefit of this kernel function is that it has analytically tractable expansion. The detailed properties of this kernel is summarized in Proposition 2, which is a direct extension from the one-dimensional case derived in Zhu et al. (1998).

Proposition 1. *For a specific* $l \in \{1, ..., \infty\}$*, define series* $\{k_i\}_{i=0}^d$*,* $\{l_i\}_{i=0}^d$ *and* $\{m_i\}_{i=1}^d$ *as follows*

$$k_i = \left\{ k_i \in \mathbb{N}^0 : \binom{k_i + d - i - 1}{d - i} \leqslant l_i \leqslant \binom{k_i + d - i}{d - i} - 1 \right\},\ 0 \leqslant i \leqslant d-1,\ k_d = 0,$$

$$l_0 = l - 1,\ l_i = l_{i-1} - \binom{k_{i-1} + d - i}{d - i + 1},\ i \geqslant 1,$$

$$m_i = k_{i-1} - k_i,\ i \geqslant 1,$$

where $\mathbb{N}^0$ *is the set of nonnegative integers;* $\binom{n}{k} = 0$ *if* $k > n$*. Define* $L = \binom{m+d}{d} = \sum_{k=0}^{m} \binom{k+d-1}{d-1}$*. For* $\mathbf{s} = [s_1, ..., s_d]^T \in \mathbb{R}^d$*, let* $\varphi_l(\mathbf{s})$ *and* ζ_l *be the* l*th (ranking from large* ζ_l *to smale) eigenfunctions and eigenvalues for the modified square exponential kernel* $\kappa(\mathbf{s}, \mathbf{s}')$ *as defined in* (4.15)*, then*

$$\zeta_l = \left(\frac{\pi}{A}\right)^d B^{k_0}, \quad \frac{\sum_{l=1}^{L} \zeta_l}{\sum_{l=1}^{\infty} \zeta_l} = (1 - B)^d \sum_{k=0}^{m} \binom{k + d - 1}{d - 1} B^k,$$

$$\varphi_l(\mathbf{s}) = (2c)^{\frac{d}{4}} \exp\left(-c\|\mathbf{s}\|_2^2\right) \prod_{i=1}^{d} H_{m_i}(\sqrt{2c} s_i),$$

where $c = \sqrt{a^2 + 2ab}$*,* $A = a + b + c$ *and* $B = b/A$*;* $H_k(\cdot)$ *is the* k*th (*$k \in \mathbb{N}^0$*) order normalized hermit polynomial.*

3.3.4 A Markov chain Monte Carlo Algorithm

We developed a efficient MCMC sampling algorithm for posterior inference about $[\{\tilde{\boldsymbol{\beta}}(\mathbf{s}_i)\}_{i=1}^n, \{\mathbf{u}_k\}_{k=1}^p, \{\sigma^2(\mathbf{s}_i)\}_{i=1}^n, \{\tau_k^2\}_{k=1}^p, \lambda \mid \mathcal{Y}]$ based on the representation and approximation for our model with the TGP priors (3.10)-(3.12).

Updating $\tilde{\boldsymbol{\beta}}(\mathbf{s}_i),\ i = 1, \ldots, n$, is an essential step in the MCMC algorithm. The Metropolis-Hasting (M-H) algorithm is employed with a block updating scheme separately for $\{\tilde{\beta}_k(\mathbf{s}_i)\}_{\mathbf{s}_i \in \mathcal{R}_g},\ g = 1, \ldots, G$ to facilitate efficient chain mixing. Under the scenario where region partition structure $\mathcal{R}_1, \ldots, \mathcal{R}_G$ is available and reliable, we can directly use this partition information. For voxel level analysis or analysis where no prior knowledge

about the regional information are adopted, we first fit voxel-wise GLMs and then use certain clustering algorithms to cluster the resulting spatially varying coefficient values. This initial clustering results for the brain locations (usually centers of voxels) are used for block updating. More details about the MCMC algorithm are available in the supplementary material.

3.3.5 Posterior Inference on SVCFs

With the recorded MCMC samples $\tilde{\beta}_k^{(t)}(\mathbf{s}_i)$, $\mathbf{u}_k$ and $\lambda_k^{(t)}$, $t = 1, \ldots, T$, we can achieve two major goals: 1) selecting neuroimaging features; 2) estimating covariate effects at the selected brain regions.

To select the important imaging features at the regional level, we estimate the selection probability of every region, $g = 1, \ldots, G$, according to the definition (C1)-(C3), using the MCMC samples as

$$\widehat{\mathbb{P}}(g \in \mathcal{I}_1) = \widehat{\mathbb{P}}\left(\inf_{1 \leqslant i \leqslant n, \mathbf{s}_i \in \mathcal{R}_g} |\tilde{\beta}_k(\mathbf{s}_i)| > \lambda_k \mid \mathcal{Y} \right) \approx \frac{1}{T} \sum_{t=1}^{T} I\left[\inf_{1 \leqslant i \leqslant n, \mathbf{s}_i \in \mathcal{R}_g} |\tilde{\beta}_k^{(t)}(\mathbf{s}_i)| > \lambda_k^{(t)} \right],$$

then we estimate $\boldsymbol{\beta}(\mathbf{s}_i)$ as follows, if $\mathbf{s}_i \in \mathcal{R}_g$,

$$\hat{\beta}_k(\mathbf{s}_i) = \begin{cases} \widehat{\mathbb{E}}[\tilde{\beta}_k(\mathbf{s}_i) \mid \inf_{1 \leqslant j \leqslant n, \mathbf{s}_j \in \mathcal{R}_g} |\tilde{\beta}_k(\mathbf{s}_j)| > \lambda_k, \mathcal{Y}], & \widehat{\mathbb{P}}(g \in \mathcal{I}_1) > q \\ 0, & \widehat{\mathbb{P}}(g \in \mathcal{I}_1) \leqslant q \end{cases}, \quad (3.15)$$

for all $k = 1, ..., p$, where $0.5 < q < 1$ is a threshold for the posterior probabilities of being nonzero at certain brain locations. We use $q = 0.90$ throughout the rest of our analysis. Estimates for the posterior conditional expectations in (3.15) can be easily calculated based on the posterior samples.

As a special case, to conduct voxel level selection (i.e., each voxel is a

region with voxel centers being $s_1, \ldots, s_n$), we can simply adapt (3.15) to

$$\hat{\beta}_k(s_i) = \begin{cases} \widehat{\mathbb{E}}[\tilde{\beta}_k(s_i) \mid |\tilde{\beta}_k(s_i)| > \lambda_k, y], & \widehat{\mathbb{P}}(|\tilde{\beta}_k(s_i)| > \lambda_k \mid y) > q \\ 0, & \widehat{\mathbb{P}}(|\tilde{\beta}_k(s_i)| > \lambda_k \mid y) \leqslant q \end{cases}, \tag{3.16}$$

where $\widehat{\mathbb{P}}(|\tilde{\beta}_k(s_i)| > \lambda_k \mid y) \approx \frac{1}{T}\sum_{t=1}^{T} I\left[|\tilde{\beta}_k^{(t)}(s_i)| > \lambda_k^{(t)}\right]$ can be regarded as the posterior probability of activation for each voxel.

3.4 Numerical Examples

3.4.1 Simulation Study: Synthetic Imaging Data

In this simulation, three covariate functions, $\beta_1(s), \beta_2(s)$ and $\beta_3(s)$, were created on $\mathcal{R}$ as shown in Figure 3.3 (the "true signal" column); a 2D variance process, $\sigma^2(s)$, which we assigned the $\log$-Gaussian process prior to, is generated by exponentiating Gaussian processes. We considered $n = 30 \times 30$ and 50×50 spatial locations. The data was simulated from

$$y_j(s_i) = \beta_1(s_i)x_{j1} + \beta_2(s_i)x_{j2} + \beta_3(s_i)x_{j3} + e_j(s_i)\sqrt{\sigma^2(s)}, \tag{3.17}$$

in which $x_{j1} \overset{iid}{\sim} N(0, 4), x_{j2} \overset{iid}{\sim} \text{Unif}(-1, 1), x_{j3} \overset{iid}{\sim} \text{Bernoulli}(0.5)$ and $e_j(s_i) \overset{iid}{\sim} N(0, 1)$.

We considered scenarios with sample size $m = 100$ and 200 and the variance process $\sigma^2(s)$ ranging from 8 to 10 or from 16 to 20. Given one combination of $\{n,\ m,\ \sigma^2(s)\}$, a total of 50 datasets were independently generated and separately analyzed by the proposed method. Modified SE kernel was chosen used for the TGP priors with $a = 0.25$ and b chosen as the marginal posterior mode. The priors for the thresholding parameters were fixed as $\text{Unif}(0.3, 1.25)$. All pixels are divided into four groups for block updating. The MCMC algorithm was run $10,000$ iterations with $5,000$ burn-in.

We compared our method with the standard voxelwise GLM methods. The features estimated from the GLM method are thresholded based on

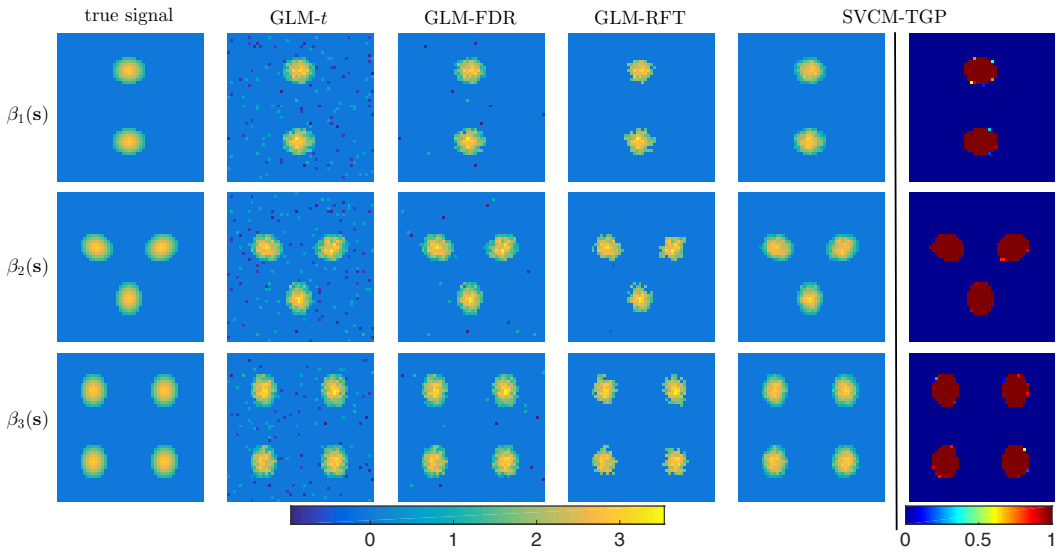


Figure 3.3: Column 1-5: true and estimated spatial covariate effects from GLM-t, GLM-FDR, GLM-RFT and SVCM-TGP; Column 6: the selection probability estimated from SVCM-TGP. The result is generated from one simulated dataset with $m = 200$ subjects, $n = 2500$ pixels and noise level $\sigma^2(\mathbf{s}) \in [16, 20]$.

the p-values for testing null hypothesis $\beta_k(\mathbf{s}_i) = 0$. In particular, we considered three thresholding approaches: the direct thresholding through naïve t-test (GLM-t) at a significant level of 0.05, thresholding using the FDR adjusted p-values (GLM-FDR) (Benjamini and Yekutieli, 2001) and thresholding by controlling FWER based on standard random field theory (GLM-RFT) (Nichols and Hayasaka, 2003). Figure 3.3 presents the estimated covariate effect functions from GLM-t, GLM-FDR, GLM-RFT and our method (SVCM-TGP) based on one simulated dataset from our experiments. According to Figure 3.3, our method provides more accurate feature selection results by eliminating false positive signals as well as maintaining high sensitivity. SVCM-TGP also provides a more accurate estimate on the SVCFs taking advantages of incorporating spatial smoothness in the model. In addition, the Bayesian SVCM-TGP can assess the uncertainty of feature selection by using the posterior probability of SVCF not being zero at each location, as shown in the last column in Figure 3.3.

For all the scenarios, Table 3.2 summerizes the relative mean square errors (ReMSE) with regard to the GLM estimates, defined as

$$\text{ReMSE} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{p}\left[\hat{\beta}_k(\mathbf{s}_i) - \beta_k(\mathbf{s}_i)\right]^2}{\sum_{i=1}^{n}\sum_{k=1}^{p}\left[\hat{\beta}^*_k(\mathbf{s}_i) - \beta_k(\mathbf{s}_i)\right]^2},$$

where $\hat{\beta}_k(\mathbf{s}_i)$ are the estimates from a certain method, $\hat{\beta}^*_k(\mathbf{s}_i)$ are the voxel-wise GLM estimates without any thresholding and $\beta_k(\mathbf{s}_i)$ represent the true values. The false discovery rates (FDRs) and the false negative rates (FNRs) are also reported in Table 3.2, specified as:

$$\text{FDR} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{p} I[\hat{\beta}_k(\mathbf{s}_i) \neq 0] \times I[\beta_k(\mathbf{s}_i) = 0]}{\sum_{i=1}^{n}\sum_{k=1}^{p} I[\hat{\beta}_k(\mathbf{s}_i) \neq 0]}$$

$$\text{FNR} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{p} I[\hat{\beta}_k(\mathbf{s}_i) = 0] \times I[\beta_k(\mathbf{s}_i) \neq 0]}{\sum_{i=1}^{n}\sum_{k=1}^{p} I[\beta_k(\mathbf{s}_i) \neq 0]}.$$

From Table 3.2, our method performs well in terms of both feature selection (small FDRs and FNRs) as well as estimation (small ReMSE), especially when the noise level is high or the number of subjects is small. GLM-RFT performs well at low noise level but deteriorates notably as noise level increases due to low sensitivity. GLM-FDR is also relatively robust but consistently generates false positive signals. The performance of the proposed SVCM-TGP method also becomes better as the number of spatial locations increases within a fixed domain, which agrees with our posterior consistency theory based on infill asymptotics.

To evaluate the robustness of our method when the error distribution is misspecified, we also generated data from (3.17) with errors from three types of distributions other than $\mathrm{N}(0,1)$: a skewed distribution, $e_j(\mathbf{s}) \overset{\text{iid}}{\sim} \{\chi^2(3) - 3\}/\sqrt{6}$; a heavy-tail distribution, $e_j(\mathbf{s}) \overset{\text{iid}}{\sim} \mathrm{DE}(1)/\sqrt{2}$ in which DE(λ) stands for the double exponential distribution with scale parameter λ; and a dual-mode distribution, $e_j(\mathbf{s}) \overset{\text{iid}}{\sim} \frac{1}{2}\mathrm{N}(-\frac{1}{2}, \frac{\sqrt{3}}{2}) + \frac{1}{2}\mathrm{N}(\frac{1}{2}, \frac{\sqrt{3}}{2})$. All of them have zero mean and unit variance. Table 3.3 summarizes the results when $n = 2500$ for these three error distributions in a similar setup as Table 3.2.

Table 3.2: Quantitative comparison of SVCM-TGP to voxel-wise GLM fitting results with various thresholdings. All results reported are the means and standard errors based on 50 independently simulated datasets with $e_j(\mathbf{s}) \sim N(0,1)$.

n		ReMSE	FDR(%)	FNR(%)	ReMSE	FDR(%)	FNR(%)
900		$\sigma^2(\mathbf{s}) \in [8,10],\ m = 100$			$\sigma^2(\mathbf{s}) \in [8,10],\ m = 200$		
	GLM-t	0.43(0.12)	20.8(2.3)	1.1(0.2)	0.41(0.23)	22.6(5.3)	0.0(0.0)
	GLM-FDR	0.32(0.08)	6.3(1.0)	3.1(0.4)	0.25(0.06)	6.9(1.5)	0.0(0.0)
	GLM-RFT	0.94(0.13)	0.6(0.2)	26.6(1.5)	0.44(0.09)	0.5(0.0)	6.7(0.6)
	SVCM-TGP	0.19(0.08)	1.5(0.4)	2.2(0.2)	0.16(0.02)	1.1(0.1)	0.4(0.0)
900		$\sigma^2(\mathbf{s}) \in [16,20],\ m = 100$			$\sigma^2(\mathbf{s}) \in [16,20],\ m = 200$		
	GLM-t	0.46(0.13)	23.3(3.5)	6.5(1.1)	0.41(0.11)	19.5(4.9)	1.0(0.1)
	GLM-FDR	0.42(0.10)	6.3(0.5)	15.7(3.1)	0.30(0.06)	4.9(0.9)	3.1(0.6)
	GLM-RFT	1.28(0.02)	1.0(0.0)	58.2(6.3)	0.95(0.14)	0.0(0.0)	26.4(2.2)
	SVCM-TGP	0.22(0.05)	3.7(0.4)	10.2(1.3)	0.18(0.03)	2.4(0.2)	2.6(0.5)
2500		$\sigma^2(\mathbf{s}) \in [8,10],\ m = 100$			$\sigma^2(\mathbf{s}) \in [8,10],\ m = 200$		
	GLM-t	0.36(0.10)	31.1(5.0)	0.5(0.1)	0.35(0.08)	31.3(3.4)	0.0(0.0)
	GLM-FDR	0.21(0.02)	7.2(1.3)	3.9(0.1)	0.16(0.02)	6.7(1.0)	0.1(0.0)
	GLM-RFT	0.63(0.21)	0.0(0.0)	30.1(4.9)	0.30(0.08)	0.3(0.0)	7.5(1.8)
	SVCM-TGP	0.08(0.06)	1.1(0.2)	1.2(0.4)	0.06(0.03)	0.6(0.1)	0.9(0.2)
2500		$\sigma^2(\mathbf{s}) \in [16,20],\ m = 100$			$\sigma^2(\mathbf{s}) \in [16,20],\ m = 200$		
	GLM-t	0.39(0.14)	32.9(6.9)	8.1(2.0)	0.35(0.08)	31.6(4.2)	1.2(0.2)
	GLM-FDR	0.31(0.06)	5.7(0.6)	20.1(2.4)	0.18(0.04)	3.5(1.0)	4.0(0.9)
	GLM-RFT	0.87(0.25)	0.3(0.0)	61.4(7.5)	0.66(0.14)	0.6(0.1)	30.8(1.3)
	SVCM-TGP	0.13(0.04)	3.6(1.0)	3.0(1.2)	0.07(0.02)	1.0(0.2)	1.6(0.3)

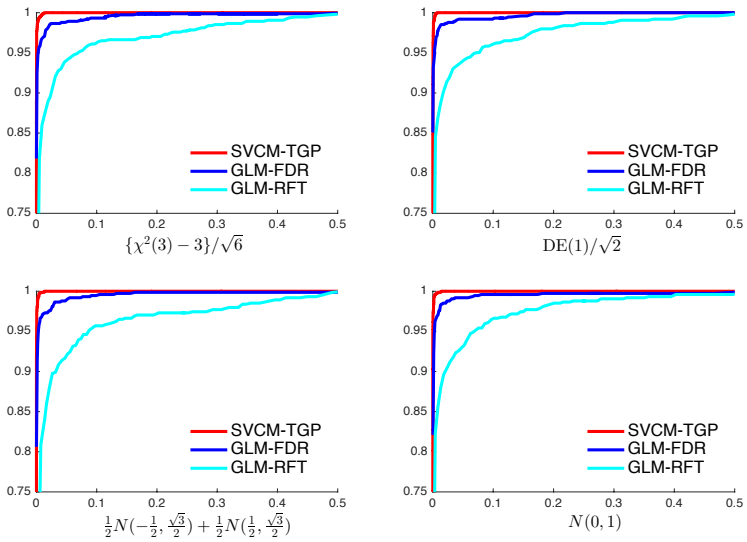


Figure 3.4: The ROC curves for our method (red curves) compared with GLM fittings using the FDR control (blue curves) or the FWER correction based on the random field theory (cyan curves) under four different distribution assumptions ($n = 2500,\ m = 200,\ \sigma^2(\mathbf{s}) \in [16, 20]$).

It shows the superiority of the proposed method in terms of estimation and feature selection for SVCFs even the error distribution is misspecified.

To comprehensively compare the performance of TGP priors in feature selection with other common methods, we also conduct the receiver operating characteristic (ROC) analysis. Since our original method will automatically generate the optimal thresholding values, in this ROC analysis, we fix λ at different values and rerun the MCMC simulation to alternate the specificities. Figure 3.4 shows the ROC curves of our method, GLM-FDR and GLM-RFT with $\sigma^2(\mathbf{s}) \in [16, 20]$ and $m = 200,\ n = 2500$ for the three alternative error specifications along with Gaussian errors. Under all four settings, SVCM-TGP achieves the best performance with the largest area under the curve.

Table 3.3: Quantitative comparison of SVCM-TGP to voxel-wise GLM fitting results with various thresholdings. All results reported are the means and standard errors based on 50 independently simulated datasets with $n = 2500$, three alternative distributions.

$e_j(\mathbf{s}_i)$		ReMSE	FDR(%)	FNR(%)	ReMSE	FDR(%)	FNR(%)
$\{\chi^2(3)-3\}/\sqrt{6}$		$\sigma^2(\mathbf{s}) \in [8,10],\ m = 100$			$\sigma^2(\mathbf{s}) \in [8,10],\ m = 200$		
	GLM-t	0.36(0.17)	32.0(8.1)	1.1(0.3)	0.35(0.11)	30.8(6.3)	0.1(0.0)
	GLM-FDR	0.19(0.06)	4.2(1.6)	4.2(0.8)	0.16(0.05)	4.3(1.3)	0.7(0.2)
	GLM-RFT	0.71(0.30)	0.0(0.0)	33.5(5.8)	0.35(0.13)	0.0(0.0)	9.3(2.2)
	SVCM-TGP	0.08(0.08)	1.9(0.5)	2.8(0.9)	0.07(0.04)	0.9(0.3)	0.9(0.2)
$\{\chi^2(3)-3\}/\sqrt{6}$		$\sigma^2(\mathbf{s}) \in [16,20],\ m = 100$			$\sigma^2(\mathbf{s}) \in [16,20],\ m = 200$		
	GLM-t	0.41(0.13)	33.9(8.5)	7.5(2.9)	0.41(0.11)	19.5(4.9)	1.0(0.1)
	GLM-FDR	0.34(0.10)	5.9(2.1)	23.3(4.3)	0.21(0.05)	5.6(1.6)	5.1(1.2)
	GLM-RFT	0.81(0.17)	0.0(0.0)	59.8(8.6)	0.65(0.11)	0.4(0.1)	31.0(6.9)
	SVCM-TGP	0.13(0.04)	2.8(0.4)	5.5(1.0)	0.07(0.03)	1.6(0.3)	2.4(0.4)
$\mathrm{DE}(1)/\sqrt{2}$		$\sigma^2(\mathbf{s}) \in [8,10],\ m = 100$			$\sigma^2(\mathbf{s}) \in [8,10],\ m = 200$		
	GLM-t	0.37(0.15)	31.8(5.0)	1.2(0.0)	0.37(0.12)	31.4(4.5)	0.1(0.0)
	GLM-FDR	0.23(0.0)	5.1(1.2)	5.7(1.3)	0.17(0.08)	4.2(1.0)	0.1(0.0)
	GLM-RFT	0.62(0.22)	0.0(0.0)	29.7.3(6.1)	0.33(0.15)	0.1(0.0)	8.1(2.5)
	SVCM-TGP	0.08(0.06)	1.9(0.4)	1.5(0.4)	0.05(0.02)	0.8(0.2)	0.8(0.0)
$\mathrm{DE}(1)/\sqrt{2}$		$\sigma^2(\mathbf{s}) \in [16,20],\ m = 100$			$\sigma^2(\mathbf{s}) \in [16,20],\ m = 200$		
	GLM-t	0.36(0.19)	31.5(5.2)	5.5(1.4)	0.35(0.12)	29.2(5.0)	0.0(0.0)
	GLM-FDR	0.31(0.16)	5.9(0.9)	21.3(5.8)	0.21(0.09)	5.4(1.3)	5.2(1.9)
	GLM-RFT	0.85(0.35)	0.3(0.0)	59.5(8.4)	0.59(0.22)	0.0(0.0)	28.9(4.9)
	SVCM-TGP	0.11(0.05)	2.9(1.2)	2.3(0.7)	0.09(0.01)	2.1(0.2)	2.1(0.4)
$\frac{1}{2}\mathrm{N}\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$		$\sigma^2(\mathbf{s}) \in [8,10],\ m = 100$			$\sigma^2(\mathbf{s}) \in [8,10],\ m = 200$		
$+\frac{1}{2}\mathrm{N}\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$	GLM-t	0.37(0.17)	33.3(5.2)	1.5(0.5)	0.35(0.14)	31.8(4.3)	0.0(0.0)
	GLM-FDR	0.20(0.06)	6.0(2.1)	4.8(1.3)	0.15(0.04)	5.6(1.1)	0.3(0.0)
	GLM-RFT	0.60(0.21)	0.2(0.0)	29.5(5.2)	0.26(0.12)	0.1(0.0)	6.1(1.6)
	SVCM-TGP	0.08(0.03)	1.5(0.6)	2.1(0.8)	0.07(0.04)	1.2(0.4)	2.0(0.6)
$\frac{1}{2}\mathrm{N}\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$		$\sigma^2(\mathbf{s}) \in [16,20],\ m = 100$			$\sigma^2(\mathbf{s}) \in [16,20],\ m = 200$		
$+\frac{1}{2}\mathrm{N}\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$	GLM-t	0.38(0.16)	32.5(5.4)	7.7(2.0)	0.35(0.13)	31.2(4.0)	1.1(0.3)
	GLM-FDR	0.30(0.11)	5.3(1.0)	21.5(6.3)	0.19(0.08)	5.8(2.0)	3.4(1.3)
	GLM-RFT	0.87(0.26)	0.4(0.1)	62.2(9.5)	0.60(0.07)	0.4(0.1)	29.7(5.3)
	SVCM-TGP	0.15(0.02)	6.6(2.2)	3.2(1.0)	0.10(0.03)	1.1(0.3)	3.1(0.9)

3.4.2 Real Data Application: The Autism Brain Imaging Data Exchange (ABIDE)

We apply our method to the data from ABIDE, which is a consortium collecting and sharing resting-state fMRI data from 1,112 subjects. Covariate information such as age at scan, sex, IQ, handedness and diagnostic information are also available from ABIDE studies. Among the subjects, 539 individuals have Autism spectrum disorders (ASD), which are characterized by symptoms such as social difficulties, communication deficits, stereotyped behaviors and cognitive delays. The remaining subjects are the age-matched normal controls (NC). All the fMRI images are preprocessed through slice-timing, motion correction, nuisance signal regression and temporal filtering. The resulting fMRI data, which are $91 \times 109 \times 91$ 3D matrices, are normalized and registered to Montreal Neurological Institute (MNI) 152 stereotactic space. We aim to investigate the voxel-wise measures of latent functional architecture of the brains through fractional amplitude of low-frequency fluctuations (fALFF)(Zou et al., 2008). fALFF is a metric reflecting the percentage of power spectrum within low-frequency domain ($0.01 - 0.1$Hz) which characterizes the intensity of spontaneous brain activities. We calculate the fALFF for each subject at every voxel. Since the fALFF is restricted to $(0, 1)$, we perform the following monotone transformation

$$y_j(\mathbf{s}) = \log\left(\frac{f_j(\mathbf{s})}{1 - f_j(\mathbf{s})}\right), \tag{3.18}$$

where $f_j(\mathbf{s})$ represents the fALFF for subject $j = 1, \ldots, 1112$ at brain location $\mathbf{s}$ and treat the transformed data as our outcomes.

The covariates we choose for fitting model (3.1) are [1, group, age, gender, group$\times$ age, group$\times$ gender]. We use all the voxels at the gray matter as the observed spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and all the anatomical parcellation based on MNI templates as our brain regions $\mathcal{R}_1, \ldots, \mathcal{R}_G$ ($n = 177,743$

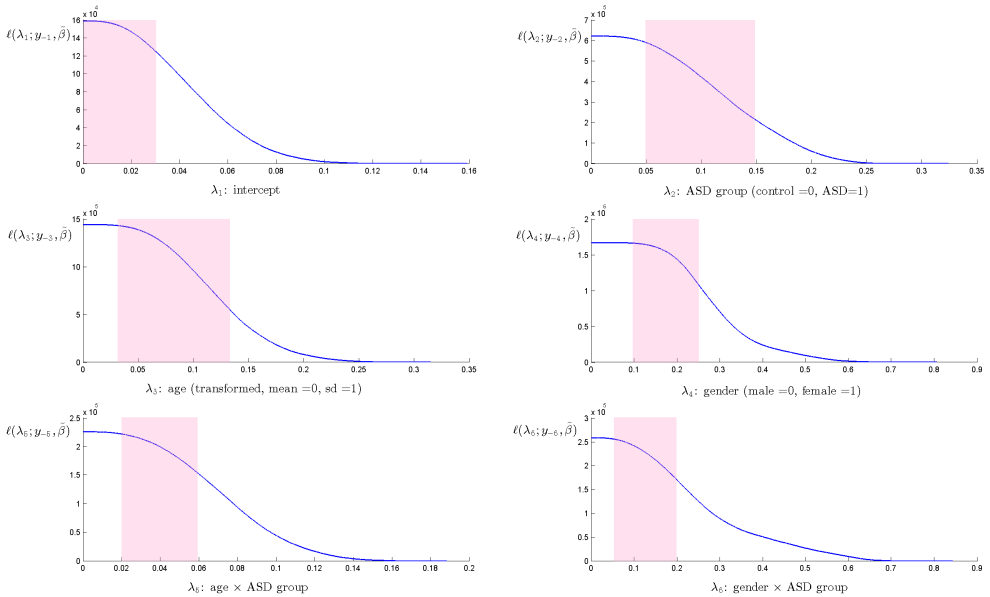
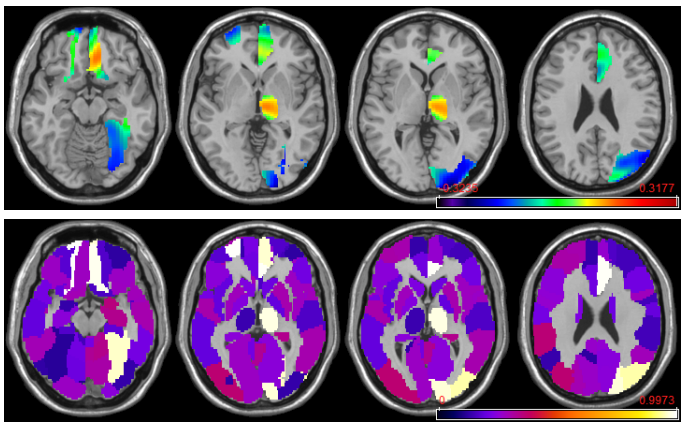


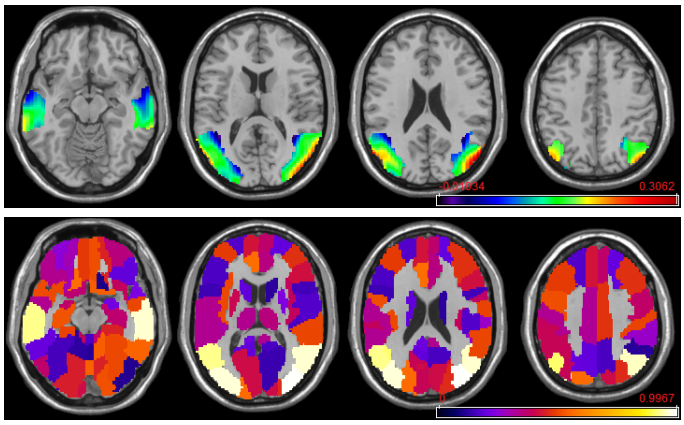
Figure 3.5: $\ell(\lambda_k)$ for specifying λ priors in the analysis of ABIDE data. The colored shades mark the intervals we choose as the range of the uniform priors for λ.

and $G = 116$). The imaging outcomes are centered across all subjects at each voxel. The group variable equals to 1 for the ASD subjects; the ages are all centered and scaled with zero mean and unit variance; the gender variable equals to 1 for female subjects. The priors for the thresholding parameters are determined through the method described in subsection 3.3.2. The profiles of $\widehat{\ell}(\lambda_k)$ are shown in Figure 3.5. The priors for the six thresholding parameters are Unif$(0, 0.03)$, Unif$(0.05, 0.15)$, Unif$(0.03, 0.13)$, Unif$(0.1, 0.25)$, Unif$(0.02, 0.06)$ and Unif$(0.05, 0.2)$ according to the plots. The Gaussian kernel we use is the modified square exponential kernel with a fixed as 0.25. b equals to 95, which is determined by maximizing the marginal posterior. To achieve 90% recovery rate of the KL expansion, we set $L = 1,140$ eigenfunctions. The MCMC algorithm runs $60,000$ iterations with $25,000$ burn-in.

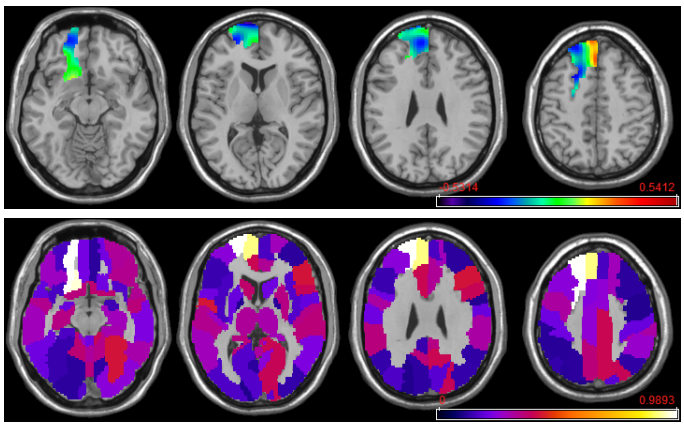
Based on our results, the ASD subjects tend to show lower fALFF outcomes at the median and superior part of the right occipital lobe, which is the visual processing centers of human brains (the visual cortex). We ob-



(A) Covariate effects for the ASD group versus the control



(B) Covariate effects for the age



(C) Covariate effects for group and gender interaction

Figure 3.6: Estimated SVCFs (top row in each subplot) and regional selection probabilities (bottom row in each subplot) based on posterior samples from our MCMC algorithm for “ASD group", “age" and “ASD group× gender"

serve significantly higher activities at the right fusiform gyrus, which has been reported to be related to Autism in (Hadjikhani et al., 2004). Sim-

ilar findings are observed at the right median orbitofrontal cortex, the region involved in most human cognition processes, especially decision-making, indicating more spontaneous brain cognition activities among the ASD subjects. From the axial view, Figure 3.6(A) shows the information discussed above. Some other regions that are selected includes the right thalamus and the right anterior cingulum, which we do not discuss here in detail.

Another major findings are the age effect on the fALFFs. We identified three brain regions that show higher fALFF outcomes as the age increases: the median occipital lobe, the median temporal lobe and the angular gyrus. These regions are generally involved in brain functions such as spatial temporal cognition, language, memory, attention and visual processing. Figure 3.6(B) shows the findings above in brain slices from the axial view.

Although no specific regions of interest are observed for the “gender" variable, certain brain regions demonstrate different activation patterns its interaction with the disease group. Specifically, female ASD subjects have higher fALFFs as compared with male ASD effects at the left median and superior part of the orbital gyrus but lower fALFFs at the left frontal lobe gyrus and the left rectus. Figure 3.6(C) shows these findings in three views for the ease of demonstration. Beyond these findings, we also note that the right inferior temporal gyrus displays smaller effects among the female ASD subjects as compared with the male autistics.

3.5 Discussion

In this chapter, we introduced a new family of prior, the TGP prior, for feature selections within spatially varying coefficient functions and its applications to massive neuroimaging data analysis. We demonstrate the prior large support properties of the TGP prior and its posterior consistency under the spatially varying coefficient models under the spatial infill asymptotics. Simulation studies show that the TGP prior is espe-

cially useful for imaging feature selections with relatively large noise or small sample sizes.

In most spatial statistics literatures such as Diggle et al. (1998); Gelfand et al. (2003), a spatial process are decomposed into three parts: a deterministic trend process or, in other words, mean process; a zero-mean variance process with continuous sample path and a zero-mean white noise process, i.e., the nugget effect. Under this general framework, Zhu et al. (2014) considered a more complex model as compared with our model (3.1), which could be expressed using our notations as

$$y_j(\mathbf{s}) = \mathbf{x}_j^T \boldsymbol{\beta}(\mathbf{s}) + \eta_j(\mathbf{s}) + e_j(\mathbf{s}), \tag{3.19}$$

for all subjects $j = 1, ..., m$. They estimated the additional term $\eta_j(\mathbf{s})$ for every subject through standard local linear regression techniques, which do not require a pre-specified spatial covariance structure $\boldsymbol{\Sigma}_\eta$, or equivalently, its Karhunen-Loève basis. Since in this chapter, our primary focus is on the GLM framework for imaging data, we did not apply our TGP prior under the setting of model (3.19). To enable a similar analysis using the TGP prior for $\boldsymbol{\beta}(\mathbf{s})$ in (3.19) under the Bayesian framework, there are four major tasks. First, a proper prior specification for $\eta_j(\mathbf{s})$ needs to be introduced which should be flexible enough to capture various spatially smooth dynamics. Second, a computationally efficient algorithm for estimating the additional parameters brought by $\eta_j(\mathbf{s})$ is required since this set of parameters scale with the number of subjects. Third, since for each subject, we will have a subject specific random effect term, we need to carefully monitor the model fitting procedures to avoid potential overfitting issues. Fourth, the theoretical analysis for posterior consistency in Theorem 3 needs to be adapted to the more challenging model structure.

In addition to applying the TGP prior to model (3.19), our study can be extended to some other directions. With a focus on neuroimaging studies using model (3.1), we can extend the prior constructions to enable

self-guided parcellation while conducting feature selection tasks. This can help relax the region based sparsity assumptions for the SVCFs. We can also explore the Bayesian asymptotic theories when the number of brain region partitions diverges. With a focus on general Bayesian analysis, we can extend TGP for modeling high-dimensional multivariate binary processes or selecting features for scalar-on-image models such as neuropsychiatric disease predictions.

Appendices

3.5.1 Proof of Theorem 1

Based on the conditions (C1)–(C3) for $\beta^0(\mathbf{s})$, let $\mathcal{I}_1$ be the index set of brain regions with features and $\mathcal{I}_0 = \{1, ..., G\} \backslash \mathcal{I}_1$ which implies $\beta^0(\mathbf{s}) = 0$ for all $\mathbf{s} \in \cup_{g \in \mathcal{I}_0} \mathcal{R}_g$. Then

$$\Pi(\|\beta(\mathbf{s}) - \beta^0(\mathbf{s})\|_\infty < \varepsilon) \geqslant \Pi\left(\sup_{\mathbf{s} \in \cup_{g \in \mathcal{I}_1} \mathcal{R}_g} |\tilde{\beta}(\mathbf{s}) - \beta^0(\mathbf{s})| < \varepsilon, \inf_{\mathbf{s} \in \cup_{g \in \mathcal{I}_1} \mathcal{R}_g} |\tilde{\beta}(\mathbf{s})| > \lambda, \sup_{\mathbf{s} \in \cup_{g \in \mathcal{I}_0} \mathcal{R}_g} |\tilde{\beta}(\mathbf{s})| \leqslant \lambda \right). \tag{3.20}$$

Without loss of generality, we only consider $0 < \varepsilon < \lambda^0 - \lambda$. Note that for all $\mathbf{s} \in \cup_{g \in \mathcal{I}_1} \mathcal{R}_g$, $|\tilde{\beta}(\mathbf{s}) - \beta^0(\mathbf{s})| < \varepsilon$ and $|\beta^0(\mathbf{s})| \geqslant \lambda^0$ implies that $|\tilde{\beta}(\mathbf{s})| \geqslant \lambda^0 - \varepsilon > \lambda$, then (3.20) leads to

$$\Pi\left(\|\beta(\mathbf{s}) - \beta^0(\mathbf{s})\|_\infty < \varepsilon\right) \geqslant \Pi\left(\sup_{\mathbf{s} \in \cup_{g \in \mathcal{I}_1} \mathcal{R}_g} |\tilde{\beta}(\mathbf{s}) - \beta^0(\mathbf{s})| < \varepsilon, \sup_{\mathbf{s} \in \cup_{g \in \mathcal{I}_0} \mathcal{R}_g} |\tilde{\beta}(\mathbf{s})| \leqslant \lambda \right).$$

Let $\phi_l(\mathbf{s})$ and $\zeta_l, l = 1, ..., \infty$, be the normalized eigenfunctions and eigenvalues of the kernel function $\kappa(\cdot, \cdot)$, then the Karhunen-Loève expansions of $\gamma(\mathbf{s})$ and $\epsilon(\mathbf{s})$ can be expressed as $\gamma(\mathbf{s}) = \sum_{l=1}^{\infty} u_l \phi_l(\mathbf{s})$, $\forall \mathbf{s} \in \mathcal{R}$, and $\epsilon(\mathbf{s}) = \sum_{l=1}^{\infty} v_{lg} \phi_l(\mathbf{s})$, $\mathbf{s} \in \mathcal{R}_g, g = 1, ..., G$, such that $u_l \overset{\text{iid}}{\sim} N(0, \zeta_l \tau^2)$, $v_{lg} \overset{\text{iid}}{\sim} N(0, \zeta_l \theta^2)$ and $\{u_l, v_{lg}, l = 1, ..., L, g = 1, ..., G\}$ are mutually independent. Since the RKHS of $\kappa(\cdot, \cdot)$ satisfying (K1) is $\mathcal{C}(\mathcal{R})$ (Tokdar and Ghosh, 2007), for all $\mathbf{s} \in \mathcal{R}$, $\beta^0(\mathbf{s})$ can be represented by $\sum_{l=1}^{\infty} w_{lg} \phi_l(\mathbf{s})$, where

$\sum_{l=1}^{\infty} w_{lg}^2/\zeta_l < \infty$.

For $\mathbf{s} \in \mathcal{R}_g$ with $g \in \mathcal{I}_1$

$$\sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}(\mathbf{s}) - \beta^0(\mathbf{s})| \leqslant \sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}_{L,g}(\mathbf{s}) - \beta^0_{L,g}(\mathbf{s})| + \sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}^*_{L,g}(\mathbf{s})| + \sup_{\mathbf{s}\in\mathcal{R}_g} |\beta^{0*}_{L,g}(\mathbf{s})|, \quad (3.21)$$

where $\tilde{\beta}_{L,g}(\mathbf{s}) = \sum_{l=1}^{L}(u_l+v_{lg})\phi_l(\mathbf{s})$, $\beta^0_{L,g}(\mathbf{s}) = \sum_{l=1}^{L} w_{lg}\phi_l(\mathbf{s})$, $\tilde{\beta}^*_{L,g}(\mathbf{s}) = \tilde{\beta}(\mathbf{s}) - \tilde{\beta}_{L,g}(\mathbf{s})$ and $\beta^{0*}_{L,g}(\mathbf{s}) = \beta^0(\mathbf{s}) - \beta^0_{L,g}(\mathbf{s})$. Since the RKHS of $\kappa(\cdot,\cdot)$ is $\mathcal{C}(\mathcal{R})$, $\tilde{\beta}(\mathbf{s})$ is uniformly continuous on $\overline{\mathcal{R}}_g$ with probability 1. Based on this property, Theorem 3.1.2 of Adler and Taylor (2009) guarantees that

$$\lim_{L\to\infty} \sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}^*_{L,g}(\mathbf{s})| = 0$$

with probability 1. By the uniform convergence of $\sum_{l=1}^{L} w_{lg}\phi_l(\mathbf{s})$ to $\beta^0(\mathbf{s})$ as $L \to \infty$ on $\overline{\mathcal{R}}_g$,

$$\lim_{L\to\infty} \sup_{\mathbf{s}\in\mathcal{R}_g} |\beta^{0*}_{L,g}(\mathbf{s})| = 0.$$

Then we can find a finite integer L_g such that for all $L \geqslant L_g$,

$$\sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}^*_{L,g}(\mathbf{s})| < \frac{\varepsilon}{3} \text{ with probability 1}, \quad \sup_{\mathbf{s}\in\mathcal{R}_g} |\beta^{0*}_{L,g}(\mathbf{s})| < \frac{\varepsilon}{3}.$$

Since $\phi_l(\mathbf{s}), l = 1, ..., L_g$ are all continuous on $\mathcal{R}$, we have that $\max_{1\leqslant l\leqslant L_g} \|\phi_l(\mathbf{s})\|_\infty < M_{\phi,L_g}$ where M_{ϕ,L_g} is a certain constant. Let $|u_l + v_{lg} - w_{lg}| < \frac{\varepsilon}{3L_g M_{\phi,L_g}}$ for all $l = 1, ..., L_g$ and consider $L = L_g$ in (3.21), we have that

$$\sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}_{L,g}(\mathbf{s}) - \beta^0_{L,g}(\mathbf{s})| < \frac{\varepsilon}{3}.$$

Thus, the condition $|u_l + v_{lg} - w_{lg}| < \frac{\varepsilon}{3L_g M_{\phi,L_g}}, l = 1, ..., L_g$ can guarantee that $\sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}(\mathbf{s}) - \beta^0(\mathbf{s})| < \varepsilon$ with probability 1 when $g \in \mathcal{I}_1$.

For $\mathbf{s} \in \mathcal{R}_g$ with $g \in \mathcal{I}_0$, similar to (3.21) and the definitions above, we have

$$\sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}(\mathbf{s})| \leqslant \sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}_{L,g}(\mathbf{s})| + \sup_{\mathbf{s}\in\mathcal{R}_g} |\tilde{\beta}^*_{L,g}(\mathbf{s})|. \quad (3.22)$$

Then, in a similar fashion, we can find L_g and M_{ϕ,L_g} such that $|u_l + v_{lg}| \leqslant$

$\frac{\lambda}{2L_g M_{\phi,L_g}}, l = 1, ..., L_g$ guarantees that $\sup_{\mathbf{s}\in\mathcal{R}_g}|\tilde{\beta}(\mathbf{s})|\leqslant \lambda$ with probability 1 if $g \in \mathcal{I}_0$.

Based on the discussion above,

$$\Pi\left(\|\beta(\mathbf{s}) - \beta^0(\mathbf{s})\|_\infty < \varepsilon\right) \geqslant \Pi\Bigg(\left\{|u_l + v_{lg} - w_{lg}| < \tfrac{\varepsilon}{3L_g M_{\phi,L_g}} : l = 1, ..., L_g, g \in \mathcal{I}_1\right\} \cup \left\{|u_l + v_{lg}| \leqslant \tfrac{\lambda}{2L_g M_{\phi,L_g}} : l = 1, ..., L_g, g \in \mathcal{I}_0\right\}\Bigg) > 0,$$

due to the positive measures assigned on arbitrary nonempty sets by the $\left(\sum_{g=1}^{G} L_g + L_{max}\right)$-dimensional multivariate Gaussian distribution:

$$(u_1, ..., u_{L_{max}}, v_{11}, ..., v_{L_1 1}, ..., v_{1G}, ..., v_{L_G G}),$$

where $L_{max} = \max_{g=1,...,G} L_g$.

3.5.2 Proof of Theorem 2

For the ease of exposition, let $\sigma_i^2 = \sigma^2(\mathbf{s}_i)$, $\sigma^2_{min} = \inf_{\mathbf{s}\in\mathcal{R}} \sigma^2(\mathbf{s})$ and $\sigma^2_{max} = \sup_{\mathbf{s}\in\mathcal{R}} \sigma^2(\mathbf{s})$ from here on.

KL neighborhood conditions for noniid outcomes

Lemma 2. *Consider for our observed data* $\mathbf{y}(\mathbf{s}_i) \in \mathbb{R}^m$, $\mathbf{y}(\mathbf{s}_i) \sim f_{i,\boldsymbol{\beta}^0}(\mathbf{y})$ *where*

$$f_{i,\boldsymbol{\beta}^0}(\mathbf{y}) = \left\{2\pi\sigma_i^2\right\}^{-m/2} \exp\left\{-\frac{1}{2\sigma_i^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^0(\mathbf{s}_i)\|_2^2\right\}.$$

Define

$$D_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \log\frac{f_{i,\boldsymbol{\beta}_0}}{f_{i,\boldsymbol{\beta}}}, \quad K_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \mathbb{E}_{f_{i,\boldsymbol{\beta}_0}}[D_i(\boldsymbol{\beta}_0, \boldsymbol{\beta})], \quad V_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \mathit{Var}_{f_{i,\boldsymbol{\beta}_0}}[D_i(\boldsymbol{\beta}_0, \boldsymbol{\beta})].$$

If we a assign an independent TGP prior for each dimension of $\boldsymbol{\beta}$ *in* $f_{i,\boldsymbol{\beta}}(\mathbf{y})$*, i.e.,*

$$\beta_k(\mathbf{s}) \sim \mathcal{TGP}(\tau_k^2, \theta_k^2, \lambda_k, \kappa(\cdot,\cdot)),$$

and $\boldsymbol{\beta}^0 \in \boldsymbol{P}$ *then we have that* $\exists B, \Pi(B) > 0$ *such that*

$$\liminf_{n\to\infty} \Pi\left(\left\{\boldsymbol{\beta} \in B : \frac{1}{n}\sum_{i=1}^{n} K_i(\boldsymbol{\beta}^0, \boldsymbol{\beta}) < \varepsilon\right\}\right) > 0,$$

$$\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} V_i(\boldsymbol{\beta}^0, \boldsymbol{\beta}) \to 0, \forall \boldsymbol{\beta} \in B.$$

Proof. The Gaussian residuals implies that

$$D_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \frac{1}{\sigma_i^2}(\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i))^T \left[\mathbf{X}^T\mathbf{y} - \frac{1}{2}\mathbf{X}\mathbf{X}^T(\boldsymbol{\beta}^0(\mathbf{s}_i) + \boldsymbol{\beta}(\mathbf{s}_i))\right].$$

Then the we can evaluate the quantities defined above as,

$$K_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \frac{1}{2\sigma_i^2}(\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i))^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i)) \leqslant \frac{m d_{\max}}{2\sigma_i^2}\|\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i)\|_2^2,$$

$$V_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \frac{1}{\sigma_i^2}(\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i))^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i)) \leqslant \frac{m d_{\max}}{\sigma_i^2}\|\boldsymbol{\beta}^0(\mathbf{s}_i) - \boldsymbol{\beta}(\mathbf{s}_i)\|_2^2.$$

because $\mathbb{E}_{f_{i,\beta_0}}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}^0(\mathbf{s}_i), Var_{f_{i,\beta_0}}[\mathbf{y}] = \sigma_i^2\mathbf{I}_m$.

Now consider

$$B_k = \left\{\beta_k(\mathbf{s}) : \|\beta_k(\mathbf{s}) - \beta^0(\mathbf{s})\|_\infty < \sqrt{\frac{2\sigma_{\min}^2\varepsilon}{mpd_{\max}}}\right\}$$

for $k = 1, ..., p$ and let $B = B_1 \times \cdots \times B_p$. Since the priors for $\beta_k(\mathbf{s}), k = 1, ..., p$ are independent and $\Pi(B_k) > 0$ due to Theorem 1, $\Pi(B) = \prod_{k=1}^{p}\Pi(B_k) > 0$.

For all $\boldsymbol{\beta} \in B$, $K_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \leqslant \frac{m d_{\max}}{2\sigma_i^2}\sum_{k=1}^{p}\|\beta_k^0(\mathbf{s}) - \beta_k(\mathbf{s})\|_\infty^2 < \varepsilon$ and similarly $V_i(\boldsymbol{\beta}_0, \boldsymbol{\beta}) < 2\varepsilon$. Then $\liminf_{n\to\infty}\Pi\left(\left\{\boldsymbol{\beta} \in B : \frac{1}{n}\sum_{i=1}^{n}K_i(\boldsymbol{\beta}^0, \boldsymbol{\beta}) < \varepsilon\right\}\right) \geqslant \Pi(B) > 0$ and $\frac{1}{n^2}\sum_{i=1}^{n}V_i(\boldsymbol{\beta}^0, \boldsymbol{\beta}) < \frac{2\varepsilon}{n} \to 0$ for all $\boldsymbol{\beta} \in B$ as $n \to \infty$. □

Sieve constructions

Define the set of functions

$$\mathcal{P}_n = \left\{\beta(\mathbf{s}) \in \mathcal{P} : \|\beta(\mathbf{s})\|_\infty < \sqrt{n},\ \sup_{\mathbf{s}\in\mathcal{R}_g}|D^\alpha\beta(\mathbf{s})| < \sqrt{n},\ g \in \mathcal{I}_1,\ 1 \leqslant \|\alpha\|_1 \leqslant \rho\right\},$$

as our sieve.

Lemma 3. *If* $G < \infty$*, the* ε*-covering number under the sup-norm for* $\mathcal{P}_n$ *satisfies*

$$\log N(\varepsilon, \mathcal{P}_n, \|\cdot\|_\infty) < Cn^{\frac{d}{2\rho}}\varepsilon^{-d},$$

for some finite constant C*.*

Proof. Define

$$\mathcal{P}_{n,g} = \left\{ \beta(\mathbf{s}) \in \mathcal{C}^\rho(\overline{\mathcal{R}}_g) : \sup_{\mathbf{s}\in\mathcal{R}_g} |D^\alpha \beta(\mathbf{s})| < \sqrt{n},\ 0 \leqslant \|\alpha\|_1 \leqslant \rho \right\},$$

for all $g = 1, ..., G$. Theorem 2.7.1 of Van Der Vaart and Wellner (1996) implies that

$$\log N(\varepsilon, \mathcal{P}_{n,g}, \|\cdot\|_\infty) \leqslant C_g n^{\frac{d}{2\rho}}\varepsilon^{-d},$$

for some constants $C_g < \infty$. Then by the definition of $\mathcal{P}_n$, we have that

$$N(\varepsilon, \mathcal{P}_n, \|\cdot\|_\infty) \leqslant \prod_{g=1}^{G} N(\varepsilon, \mathcal{P}_{n,g}, \|\cdot\|_\infty) \leqslant \exp\left\{ Cn^{\frac{d}{2\rho}}\varepsilon^{-d} \right\},$$

where $C = \sum_{g=1}^{G} C_g < \infty$. □

Lemma 4. *Consider the TGP prior for* $\beta(\mathbf{s})$ *with kernel function satisfying condition (K1)(K2), then* $\Pi(\mathcal{P} \cap \mathcal{P}_n^c) \leqslant De^{-bn}$ *for some constant* $D, b > 0$*.*

Proof. The construction of the TGP prior implies that

$$\Pi(\mathcal{P}_n) \geqslant \prod_{g=1}^{G} \Pi\left(\sup_{s\in\mathcal{R}_g} |D^\alpha \tilde{\beta}_g(\mathbf{s})| > \sqrt{n}, 0 \leqslant \|\alpha\|_1 \leqslant \rho \right),$$

where $\tilde{\beta}_g(\mathbf{s}) \overset{iid}{\sim} \mathcal{GP}(0, (\theta^2+\tau^2)\kappa(\mathbf{s}, \mathbf{s}'))$ for all $g = 1, ..., G$. By applying Theorem 5 of Ghosal et al. (2006), we have that $\Pi\left(\sup_{s\in\mathcal{R}_g} |D^\alpha \tilde{\beta}_g(\mathbf{s})| > \sqrt{n}, 0 \leqslant \|\boldsymbol{\alpha}\|_1 \leqslant \rho\right) \geqslant 1 - Ae^{-bn}$ for some $A, b > 0$, given that $\kappa(\mathbf{s}, \cdot)$ has continuous partial derivatives of order $2\rho + 2$ on the compact set $\mathcal{R}$. Then we have that $\Pi(\mathcal{P} \cap \mathcal{P}_n^c) \leqslant 1 - (1 - Ae^{-bn})^G \leqslant De^{-dn}$ where $D = AG$ due to the fact that $(1-x)^G \geqslant 1 - Gx$ for all $0 < x < 1$ and $G = 1, 2, ...$ □

Now we define $\mathbf{P}_n = \{\boldsymbol{\beta}(\mathbf{s}) = [\beta_1(\mathbf{s}), ..., \beta_p(\mathbf{s})]^T : \beta_k(\mathbf{s}) \in \mathcal{P}_n, k = 1, ..., p\}$ here and below. Then we can easily get that

$$N(\varepsilon, \mathbf{P}_n, \|\cdot\|_\infty) < \exp\{Cpn^{\frac{d}{2\rho}}\varepsilon^{-d}\}, \tag{3.23}$$

and if assign TGP priors independently for all elements in $\boldsymbol{\beta}(\mathbf{s})$ then

$$\Pi\left(\mathbf{P} \cap \mathbf{P}_n^c\right) \leqslant Dp\exp\{-bn\}. \tag{3.24}$$

Test Constructions

Lemma 5. *Consider* $y \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_m)$*, a standard linear model with sample size* m *where* $y = [y_1, ..., y_m]$*;* $\mathbf{X}$ *is an* $m \times p$ *design matrix satisfying assumption (X1). Consider the test function* $\Phi = I\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 > \frac{\varepsilon\sqrt{p}}{2}\right)$ *for testing* $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^0$ *versus* $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}^1$*, where* $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ *and* $\boldsymbol{\beta}^1 \in \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_\infty \geqslant \varepsilon\}$*;* $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}y$ *is the ordinary least square estimator. Then for* $m > \frac{8(\log 2)\sigma^2}{\varepsilon^2 d_{\min}}$*, we have that*

$$\mathbb{E}_{P_0}[\Phi] \leqslant \exp\{-\Omega_m mp\}, \quad \mathbb{E}_{P_1}[1 - \Phi] \leqslant \exp\{-\Omega_m mp\},$$

for some $\Omega_m > 0$ *depending on* m*, where* P_0 *and* P_1 *represents the probability distributions under* H_0 *and* H_1*.*

Proof. Note that for $t = 0, 1$, under H_t, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t\|_2^2 d_{\min} m/\sigma^2 \leqslant (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t)^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^t)/\sigma^2 \sim \chi_p^2$, then we have that

$$\mathbb{E}_{P_0}[\Phi] = P_0\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 > \frac{\varepsilon^2 p}{4}\right) \leqslant P_0\left(\chi_p^2 > \frac{\varepsilon^2 p d_{\min} m}{4\sigma^2}\right) \leqslant \exp\left\{-\left(\frac{\varepsilon^2 d_{\min}}{16\sigma^2} - \frac{\log 2}{2m}\right) mp\right\}, \tag{3.25}$$

where the last inequality is simply due to the f act that $P(\chi_p^2 > x) \leqslant (1 -$

$2t)^{-p/2}\exp\{-tx\}, \forall 0 < t < 1/2$ by letting $t = 1/4$. Similarly,

$$\begin{aligned}
\mathbb{E}_{P_1}[1-\Phi] &= P_1\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \leqslant \frac{\varepsilon\sqrt{p}}{2}\right) \\
&\leqslant P_1\left(\left|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^1\|_2 - \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^1\|_2\right| \leqslant \frac{\varepsilon\sqrt{p}}{2}\right) \\
&\leqslant P_1\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^1\|_2 \geqslant -\frac{\varepsilon\sqrt{p}}{2} + \|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^1\|_2\right) \\
&\leqslant P_1\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^1\|_2 \geqslant \frac{\varepsilon\sqrt{p}}{2}\right) \\
&\leqslant \exp\left\{-\left(\frac{\varepsilon^2 d_{\min}}{16\sigma^2} - \frac{\log 2}{2m}\right) mp\right\}. \qquad (3.26)
\end{aligned}$$

Define $\Omega_m = \frac{\varepsilon^2 d_{\min}}{16\sigma^2} - \frac{\log 2}{2m}$ here and below. Notice that $m > \frac{8(\log 2)\sigma^2}{\varepsilon^2 d_{\min}}$ is equivalent to $\Omega_m > 0$, we complete the proof. □

Lemma 6. *We consider n_0 locations $\mathbf{s}_1, ..., \mathbf{s}_{n_0}$ and define $\boldsymbol{\beta}_i = [\beta_1(\mathbf{s}_i), ..., \beta_p(\mathbf{s}_i)]^T$ ($\boldsymbol{\beta}_i^0$ and $\boldsymbol{\beta}_i^1$ can be defined accordingly). Suppose that we have observed the data $\mathbf{y}_i = [y_1(\mathbf{s}_i), ..., y_m(\mathbf{s}_i)]^T$ generated from $\mathbf{y}_i \sim N(\mathbf{X}\boldsymbol{\beta}_i, \sigma_i^2\mathbf{I}_m)$ independently. Consider testing $H_0 : \boldsymbol{\beta}_i = \boldsymbol{\beta}_i^0, i = 1, ..., n_0$ versus $H_1 : \boldsymbol{\beta}_i = \boldsymbol{\beta}_i^1, i = 1, ..., n_0$ where $\|\boldsymbol{\beta}_i^1 - \boldsymbol{\beta}_i^0\|_\infty \geqslant \varepsilon$ for all $i = 1, ..., n_0$. Define $\Phi_i = I\left(\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\|_2 > \frac{\varepsilon\sqrt{p}}{2}\right)$ with $\hat{\boldsymbol{\beta}}_i = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}_i$. Then for the test function $\tilde{\Phi} = I\left(\sum_{i=1}^{n_0}\Phi_i > \frac{n_0}{2}\right)$, we have that*

$$\mathbb{E}_{P_0}[\tilde{\Phi}] \leqslant \exp\{-Cn_0\}, \quad \mathbb{E}_{P_1}[1-\tilde{\Phi}] \leqslant \exp\{-Cn_0\}$$

Proof. By the results from Lemma 5, $\mathbb{E}_{P_0}[\Phi_i] \leqslant e^{-\Omega_m^i mp}$ and $\mathbb{E}_{P_1}[1-\Phi_i] \leqslant e^{-\Omega_m^i mp}$ for all $i = 1, ..., n_0$ where $\Omega_m^i = \frac{\varepsilon^2 d_{\min}}{16\sigma_i^2} - \frac{\log 2}{2m} \geqslant \frac{\varepsilon^2 d_{\min}}{16\sigma_{\max}^2} - \frac{\log 2}{2m} = \Omega_m^0 > 0$. Then

$$\mathbb{E}_{P_0}[\tilde{\Phi}] \leqslant P_0\left(\sum_{i=1}^{n_0}\Phi_i - \sum_{i=1}^{n_0}\mathbb{E}_{P_0}[\Phi_i] \geqslant \frac{n_0}{2} - n_0 e^{-\Omega_m^0 mp}\right), \qquad (3.27)$$

$$\mathbb{E}_{P_1}[1-\tilde{\Phi}] \leqslant P_1\left(\sum_{i=1}^{n_0}(1-\Phi_i) - \sum_{i=1}^{n_0}E[1-\Phi_i] \geqslant \frac{n_0}{2} - n_0 e^{-\Omega_m^0 mp}\right). \qquad (3.28)$$

By the Hoeffding inequality (Hoeffding, 1963), the right hand side of both (3.27) and (3.28) are bounded by $\exp\left\{-\frac{2}{n_0}\frac{n_0^2(1-2e^{-\Omega_m^0 mp})^2}{4}\right\}$ when $1-2e^{-\Omega_m^0 mp} >$

0, which always holds true as long as $\Omega_m^0 > 0$. That is, $\mathbb{E}_{P_0}[\tilde{\Phi}] \leqslant \exp\{-Cn_0\}$, $\mathbb{E}_{P_1}[1-\tilde{\Phi}] \leqslant \exp\{-Cn_0\}$ where $C = \frac{(1-2e^{-\Omega_m^0 mp})^2}{2}$. □

Lemma 7. *Let* $P_n(ds) = \frac{1}{n}\sum_{i=1}^{n} I[s_i \in ds]$ *be the empirical measure on* $\mathcal{R}$*. For two functions* $\beta^0(s), \beta^1(s) \in \mathcal{P}$*, if* $\|\beta^0(s)-\beta^1(s)\|_1 = \int_{s\in\mathcal{R}}|\beta^0(s)-\beta^1(s)|P_n(ds) \geqslant \varepsilon$*, we have that* $P_n(|\beta^0(s)-\beta^1(s)| \geqslant \frac{\varepsilon}{2}) \geqslant c'$ *where* $0 < c' \leqslant 1$ *is a constant. That is, the set* $\{s \in \{s_1, ..., s_n\} : |\beta^0(s)-\beta^1(s)| \geqslant \frac{\varepsilon}{2}\}$ *has* $n_0 \geqslant c'n - 1$ *elements.*

Proof. Let $S = \{s \in \mathcal{R} : |\beta^0(s)-\beta^1(s)| \geqslant \frac{\varepsilon}{2}\}$, then

$$\begin{aligned}
\varepsilon &\leqslant \int_{s\in\mathcal{R}} |\beta^0(s)-\beta^1(s)|P_n(ds) \\
&= \int_{s\in S} |\beta^0(s)-\beta^1(s)|P_n(ds) + \int_{s\in\mathcal{R}\setminus S} |\beta^0(s)-\beta^1(s)|P_n(ds) \\
&\leqslant (M_0+M_1)P_n\left(|\beta^0(s)-\beta^1(s)| \geqslant \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2}P_n(\mathcal{R}),
\end{aligned}$$

where $\mathbb{P}_n(\mathcal{R}) = 1$; $M_0 = \|\beta^0(s)\|_\infty$ and $M_1 = \|\beta^1(s)\|_\infty$ are finite constants due to absolute continuity. Thus $P_n\left(|\beta^0(s)-\beta^1(s)| \geqslant \frac{\varepsilon}{2}\right) \geqslant c'$ by letting $c' = \frac{\varepsilon}{2(M_0+M_1)}$. □

Lemma 8. *There exists a test* Φ_{β^1,β^0} *for testing* $H_0 : \beta(s) = \beta^0(s)$ *against* $H_1 : \beta(s) = \beta^1(s)$ *where* $\|\beta^1(s)-\beta^0(s)\|_{1,\infty} \geqslant \varepsilon$ *in our proposed SVCM, such that*

$$\mathbb{E}_{P_0}[\Phi_{\beta^1,\beta^0}] \leqslant \exp\{-Cn\}, \quad \mathbb{E}_{P_1}[1-\Phi_{\beta^1,\beta^0}] \leqslant \exp\{-Cn\},$$

for some constant C *with* P_0 *and* P_1 *corresponding to the probability distributions under* H_0 *and* H_1.

Proof. For two vector-valued functions $\beta^t(s) = [\beta_1^t(s), ..., \beta_p^t(s)]^T, t = 0, 1$, if $\|\beta^1(s)-\beta^0(s)\|_{1,\infty} \geqslant \varepsilon$, we must have at least one $k \in \{1, ..., p\}$, such that $\|\beta_k^1(s)-\beta_k^0(s)\|_1 \geqslant \varepsilon$, then due to Lemma 7, we can find $n_0 \geqslant c'n-1$ elements in $\{s_1, ..., s_n\}$ such that $|\beta_k^1(s)-\beta_k^0(s)| \geqslant \frac{\varepsilon}{2}$. Without loss of generality, we denote these points as $s_1, ..., s_{n_0}$. Then for all $s_i, i = 1, ..., n_0$, we have that $\|\beta^1(s_i)-\beta^0(s_i)\|_\infty \geqslant \frac{\varepsilon}{2}$.

Now define the set $S_{\beta^1,\beta^0} = \{s \in \{s_1, ..., s_n\} : \|\beta^1(s_i)-\beta^0(s_i)\|_\infty \geqslant \frac{\varepsilon}{2}\}$. Then

$n_0 = |\mathcal{S}_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}| \geqslant c'n - 1$. Define the test function

$$\Phi_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0} = I\left(\sum_{\mathbf{s}\in\mathcal{S}_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}} \Phi(\mathbf{s}) > \frac{n_0}{2}\right),$$

where $\Phi(\mathbf{s}) = I\left(\|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}(\mathbf{s}) - \boldsymbol{\beta}^0(\mathbf{s})\|_2 > \frac{\varepsilon\sqrt{p}}{2}\right)$. Then by Lemma 6 (replacing ε by $\varepsilon/2$) we have

$$\mathbb{E}_{P_0}[\Phi_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}] \leqslant \exp\{-C_0 n_0\}, \quad \mathbb{E}_{P_1}[\Phi_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}] \leqslant \exp\{-C_0 n_0\},$$

where $C_0 > 0$ is a constant. Since $n_0 \geqslant cn - 1$ for a positive constant c', we have that $\mathbb{E}_{P_0}[\Phi_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}] \leqslant \exp\{-Cn\}$ and $\mathbb{E}_{P_1}[\Phi_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}] \leqslant \exp\{-Cn\}$. □

Lemma 9. *There exists a test* Ψ *for testing* $H_0 : \boldsymbol{\beta}(\mathbf{s}) = \boldsymbol{\beta}^0(\mathbf{s})$ *against* $H_1 : \boldsymbol{\beta}(\mathbf{s}) \in U^c_{\varepsilon,n} = U^c_\varepsilon \cap \boldsymbol{P} = \{\boldsymbol{\beta}(\mathbf{s}) \in \boldsymbol{P}_n : \|\boldsymbol{\beta}(\mathbf{s}) - \boldsymbol{\beta}^0(\mathbf{s})\| \geqslant \varepsilon\}$ *in our proposed SVCM, such that*

$$\mathbb{E}_{P_0}[\Psi] \leqslant \exp\{-d_0 n\}, \quad \mathbb{E}_{P_1}[1 - \Psi] \leqslant \exp\{-d_1 n\},$$

for some constant d_0, d_1 *with* P_0 *and* P_1 *corresponding to the probability distributions under* H_0 *and* H_1.

Proof. Let $\mathcal{N} = N(\frac{\varepsilon}{2}, \mathbf{P}_n, \|\cdot\|_\infty)$ be the covering number of $\mathbf{P}_n$ by $\varepsilon/2$-balls under the supreme norm. Then for all $\boldsymbol{\beta}(\mathbf{s}) \in U^c_{\varepsilon,n}$, we can find $\boldsymbol{\beta}^j(\mathbf{s}), j \in \{1, ..., \mathcal{N}\}$ such that $\|\boldsymbol{\beta}^j(\mathbf{s}) - \boldsymbol{\beta}(\mathbf{s})\|_\infty \leqslant \frac{\varepsilon}{2}$, which implies that $\|\boldsymbol{\beta}^j(\mathbf{s}) - \boldsymbol{\beta}^0(\mathbf{s})\|_\infty \geqslant \|\boldsymbol{\beta}^0(\mathbf{s}) - \boldsymbol{\beta}(\mathbf{s})\|_\infty - \|\boldsymbol{\beta}^j(\mathbf{s}) - \boldsymbol{\beta}(\mathbf{s})\|_\infty \geqslant \frac{\varepsilon}{2}$ for all $j = 1, ..., \mathcal{N}$. Following the notations and results in Lemma 8 with regard to $\varepsilon/2$, we have that the tests $\Phi_{\boldsymbol{\beta}^j,\boldsymbol{\beta}^0}$ all satisfy that $\mathbb{E}_{\boldsymbol{\beta}^0}[\Phi_{\boldsymbol{\beta}^j,\boldsymbol{\beta}^0}] \leqslant \exp\{-d_1 n\}$ and $\mathbb{E}_{\boldsymbol{\beta}^j}[\Phi_{\boldsymbol{\beta}^j,\boldsymbol{\beta}^0}] \leqslant \exp\{-d_1 n\}$ for some constant d_1. Now for the test function

$$\Psi = \max_{j=1,...,\mathcal{N}} \Phi_{\boldsymbol{\beta}^j,\boldsymbol{\beta}^0},$$

which only depend on the set $\mathbf{P}_n$ instead of a specific $\boldsymbol{\beta}(\mathbf{s})$ in the alterna-

tive hypothesis,

$$\mathbb{E}_{P_0}[\Psi] \leqslant \sum_{j=1} \mathbb{E}_{\boldsymbol{\beta}^j}[\Phi_{\boldsymbol{\beta}^j,\boldsymbol{\beta}^0}] \leqslant \mathcal{N}\exp\{-d_1 n\} < \exp\{Cpn^{\frac{d}{2\rho}}\varepsilon^{-d} - d_1 n\} \leqslant \exp\{-d_0 n\},$$

due to the fact that $n^{\frac{d}{2\rho}} = o(n)$ for some constant d_0 based on (3.23). At the same time

$$\mathbb{E}_{P_1}[1-\Psi] \leqslant \mathbb{E}_{\boldsymbol{\beta}^1}[1-\Phi_{\boldsymbol{\beta}^1,\boldsymbol{\beta}^0}] \leqslant \exp\{-d_1 n\},$$

which complete our proof. □

Now based on Lemma 2, equation (3.24) and Lemma 9, Theorem 2 follows from a direct application of Theorem A.1. of Choudhuri et al. (2004).

3.5.3 Proof of Theorem 3

Without loss of generality, we consider $0 < \varepsilon < \min\{\lambda, \lambda^0\}$ where $\lambda^0 = \inf_{\mathbf{s}\in\mathcal{R}}\{\beta^0(\mathbf{s})I\left[|\beta^0(\mathbf{s})|> 0\right]\}$ (same to the definition in condition (C2)) is a positive number. Then we have

$$\begin{aligned}
&\Pi\left[\|\beta(\mathbf{s}) - \beta_k^0(\mathbf{s})\|_1 \geqslant \varepsilon \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right] \\
\geqslant &\Pi\left[\min\{\lambda, \lambda^0\} \times \sum_{g=1}^{G} I[r_g(\mathbf{s}) \neq r_g^0(\mathbf{s})] \geqslant \varepsilon \;\middle|\; \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right] \\
= &\Pi\left[\sum_{g=1}^{G} I[r_g(\mathbf{s}) \neq r_g^0(\mathbf{s})] \geqslant \varepsilon' \;\middle|\; \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right],
\end{aligned}$$

where $\varepsilon' = \frac{\varepsilon}{\min\{\lambda,\lambda^0\}} \in (0,1)$. Due to Corollary 1,

$$\Pi\left[\|\beta(\mathbf{s}) - \beta_k^0(\mathbf{s})\|_1 \geqslant \varepsilon \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right] \to 0$$

as $n \to \infty$ in $\mathbb{P}_{\boldsymbol{\beta}^0}^n$ probability, then

$$\Pi\left[\sum_{g=1}^{G} I[r_g(\mathbf{s}) \neq r_g^0(\mathbf{s})] \geqslant \varepsilon' \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right] \to 0$$

as $n \to \infty$ in $\mathbb{P}^n_{\beta^0}$ probability. Considering the fact that if $\mathbf{r}(\mathbf{s}) \neq \mathbf{r}^0(\mathbf{s})$, $\sum_{g=1}^{G} I[r_g(\mathbf{s}) \neq r_g^0(\mathbf{s})] \geqslant 1 > \varepsilon'$, we have

$$\Pi\left[\sum_{g=1}^{G} I[r_g(\mathbf{s}) \neq r_g^0(\mathbf{s})] \geqslant \varepsilon' \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right] > \Pi\left[\mathbf{r}(\mathbf{s}) \neq \mathbf{r}^0(\mathbf{s}) \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right],$$

which implies that

$$\Pi\left[\mathbf{r}(\mathbf{s}) = \mathbf{r}^0(\mathbf{s}) \mid \mathbf{y}(\mathbf{s}_1), \ldots, \mathbf{y}(\mathbf{s}_n)\right] \to 1,$$

as $n \to \infty$ in $\mathbb{P}^n_{\beta^0}$ probability.

3.5.4 Details about the MCMC algorithm

We list the details about our MCMC algorithm here. Denote by $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the density function of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We normally fix $\nu = w = 0.001$ in the inverse-gamma priors, fix $\theta^2 = 1$ within the local GPs and fix $\xi^2 = 1$ for the log-Gaussian variance process.

- Updating $\tilde{\boldsymbol{\beta}}(\mathbf{s}_i)$, $i = 1, ..., n$: given the block structures of $\tilde{\boldsymbol{\beta}}(\mathbf{s}_i)$, we update $\tilde{\boldsymbol{\beta}}_k^g = \{\tilde{\beta}_k(\mathbf{s}_i)\}_{\mathbf{s}_i \in \mathcal{R}_g}, k = 1, ..., p,\ g = 1, ..., G$ separately with $p \times G$ M-H steps. Specifically, the full conditional $[\tilde{\boldsymbol{\beta}}_k^g \mid \tilde{\boldsymbol{\beta}}, \mathbf{u}, \sigma^2,\ \boldsymbol{\lambda},\ \mathcal{Y}]$ is proportional to

$$h(\tilde{\boldsymbol{\beta}}_k^g) = \left[\prod_{i:\ \mathbf{s}_i \in \mathcal{R}_g} \prod_{j=1}^{m} \phi\left(y_{j,-k}(\mathbf{s}_i); x_{jk}\tilde{\beta}_k(\mathbf{s}_i) I_{\lambda_k}[\tilde{\beta}_k(\mathbf{s}_i)], \sigma^2(\mathbf{s}_i)\right)\right] \phi\left(\tilde{\boldsymbol{\beta}}_k^g; \boldsymbol{\varphi}_g \mathbf{u}_k, \theta^2 K_g\right),$$

  where $y_{j,-k}(\mathbf{s}_i) = y_j(\mathbf{s}_i) - \sum_{t \neq k} x_{jt}\tilde{\beta}_t(\mathbf{s}_i) I_{\lambda_k}[\tilde{\beta}_k(\mathbf{s}_i)]$. We adopt a Metropolis-Hasting (M-H) algorithm to update $\tilde{\boldsymbol{\beta}}_k^g$ by first generating a proposal, $\tilde{\boldsymbol{\beta}}_k^g + \Delta\tilde{\boldsymbol{\beta}}_k^g$ with a zero mean Gaussian fluctuation $\Delta\tilde{\boldsymbol{\beta}}_k^g$. Then we set $\tilde{\boldsymbol{\beta}}_k^g \leftarrow \tilde{\boldsymbol{\beta}}_k^g + \Delta\tilde{\boldsymbol{\beta}}_k^g$ with probability: $\min\left\{1, \frac{h(\tilde{\boldsymbol{\beta}}_k^g + \Delta\tilde{\boldsymbol{\beta}}_k^g)}{h(\tilde{\boldsymbol{\beta}}_k^g)}\right\}$.

- Updating $\sigma^2(\mathbf{s}_i)$, $i = 1, ..., n$: assume that the eigen-decomposition pairs for $\kappa_\sigma(\mathbf{s}, \mathbf{s}')$ are $\{\psi_w(\mathbf{s}), \eta_w\}_{w=1}^{\infty}$, the approximate KL expansion $\sigma^2(\mathbf{s}) \approx \exp\left(\sum_{w=1}^{W} \psi_w(\mathbf{s})\nu_w\right)$, $\nu_w \sim N(0, \eta_w\xi^2)$ implies that we can update $\sigma^2(\mathbf{s}_i)$ by updating $\boldsymbol{\nu} = [\nu_1, ..., \nu_W]^T$. The posterior of $\boldsymbol{\nu}$ given all

other parameters is

$$\pi[\boldsymbol{\nu} \mid \tilde{\boldsymbol{\beta}}, y] = \prod_{i=1}^{n}\prod_{j=1}^{m} \phi\left(y_j(\mathbf{s}_i); \mathbf{x}_j^{T} g_{\lambda}[\tilde{\boldsymbol{\beta}}(\mathbf{s})],\ \exp\left(\sum_{w=1}^{W} \psi_w(\mathbf{s}_i)\nu_w\right)\right) \times \prod_{w=1}^{W} \phi(\nu_w; 0, \eta_w \xi^2).$$

We use the M-H algorithm with a Langevin-type proposal (Roberts and Tweedie, 1996) to update $\boldsymbol{\nu}$ by generating a proposal $\boldsymbol{\nu}^*$ from $N(\boldsymbol{\nu} + {}^2\nabla \log \pi[\boldsymbol{\nu} \mid \tilde{\boldsymbol{\beta}}, y],\ 2^2 \mathbf{I}_W)$ in which ^{2} is a small positive constant determining the diffusion step length. Then we accept $\boldsymbol{\nu}^*$ with probability $\min\left\{1, \frac{\pi[\boldsymbol{\nu}^* \mid \tilde{\boldsymbol{\beta}}, y] q(\boldsymbol{\nu}, \boldsymbol{\nu}^*)}{\pi[\boldsymbol{\nu} \mid \tilde{\boldsymbol{\beta}}, y] q(\boldsymbol{\nu}^*, \boldsymbol{\nu})}\right\}$ in which the transition kernel function $q(\cdot, \cdot)$ is defined as $q(\boldsymbol{\nu}^1, \boldsymbol{\nu}^2) = \phi(\boldsymbol{\nu}^1; \boldsymbol{\nu}^2 + {}^2\nabla \log \pi[\boldsymbol{\nu} \mid \tilde{\boldsymbol{\beta}}, y],\ 2^2 \mathbf{I}_W)$. A detail look at $\nabla \log \pi[\boldsymbol{\nu} \mid \tilde{\boldsymbol{\beta}}, y]$ gives us that its wth element is

$$\frac{\partial \log \pi[\boldsymbol{\nu} \mid \tilde{\boldsymbol{\beta}}, y]}{\partial \nu_w} = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\left[y_j(\mathbf{s}_i) - \mathbf{x}_j^{T} g_{\lambda}[\tilde{\boldsymbol{\beta}}(\mathbf{s})]\right]^2 \psi_w(\mathbf{s}_i)}{2\exp\left(\boldsymbol{\psi}(\mathbf{s}_i)^{T}\boldsymbol{\nu}\right)} - \frac{\psi_w(\mathbf{s}_i)}{2} - \frac{\nu_w}{2\eta_w \xi^2}$$

- Updating $\boldsymbol{\lambda}$: we sequentially update $\lambda_1, ..., \lambda_p$ with M-H algorithms. Specifically, for λ_k, the full conditional $[\lambda_k \mid \lambda_{-k}, \tilde{\boldsymbol{\beta}}, \sigma^2, y]$ is

$$h(\lambda_k) = \left[\prod_{i=1}^{n}\prod_{j=1}^{m} \phi\left(y_{j,-k}(\mathbf{s}), x_{jk}\tilde{\beta}_k(\mathbf{s}_i) I_{\lambda_k}[\tilde{\beta}_k(\mathbf{s}_i)], \sigma^2(\mathbf{s}_i)\right)\right] \Pi(\lambda_k),$$

where $\Pi(\lambda_k)$ is the uniform empirical Bayes prior for λ_k defined in the previous section. The proposal for λ_k is generate from zero mean Gaussian fluctuations as $\lambda_k + \Delta\lambda_k$, which will be accepted with probability: $\min\left\{1, \frac{h(\lambda_k + \Delta\lambda_k)}{h(\lambda_k)}\right\}$.

- Updating $\{\mathbf{u}_k\}_{k=1}^{p}$: we sequentially update $\mathbf{u}_1, ..., \mathbf{u}_p$ by drawing from their full conditionals $[\mathbf{u}_k \mid \tilde{\boldsymbol{\beta}}, \tau_k^2]$. Specifically, we update $\mathbf{u}_k$ by drawing from $N(\boldsymbol{\mu}_{\mathbf{u}_k}, \boldsymbol{\Sigma}_{\mathbf{u}_k})$ where $\boldsymbol{\mu}_{\mathbf{u}_k} = \boldsymbol{\Sigma}_{\mathbf{u}_k}\left(\theta^{-2}\sum_{g=1}^{G} \boldsymbol{\varphi}_g^{T} K_g^{-1} \tilde{\boldsymbol{\beta}}_k^{g}\right)$ and $\boldsymbol{\Sigma}_{\mathbf{u}_k} = \left(\sum_{g=1}^{G} \theta^{-2}\boldsymbol{\varphi}_g^{T} K_g^{-1} \boldsymbol{\varphi}_g + \tau_k^{-2} Z^{-1}\right)^{-1}$, with $Z = \mathrm{diag}(\zeta_1, ..., \zeta_L)$.

- Updating $\{\tau_k^2\}_{k=1}^{p}$: we sequentially update $\tau_1^2, ..., \tau_p^2$ by drawing from

their full conditionals $[\tau_k^2 \mid \mathbf{u}_k]$. Specifically, we update τ_k^2 by drawing from Inv-Ga$(a_{\tau_k^2}, b_{\tau_k^2})$ where $a_{\tau_k^2} = 0.001 + \frac{L}{2}$ and $b_{\tau_k^2} = 0.001 + \frac{1}{2}\mathbf{u}_k^T \mathbf{Z}^{-1}\mathbf{u}_k$.

Chapter 4

Bayesian Independent Component Analysis Involving Spatially Dependent Sources With Application to fMRI Data

This chapter is joint work with Dr. Ying Guo and Dr. Jian Kang.

4.1 Method

We discuss our Bayesian spatially dependent ICA model and its prior specifications in this section. To begin with, we briefly review the preprocessing steps prior to ICA analysis of fMRI data.

4.1.1 Preprocessing of fMRI data

We denote by $\mathcal{V}$, a compact subset of $\mathbb{R}^{\mathrm{d}}$ ($\mathrm{d} = 2$ or 3), the brain space in the fMRI study of interest. In a certain study, time series of the BOLD signals at a limited number of brain locations, $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{\mathrm{n}} \in \mathcal{V}$, are recorded and observed. Denoted by $\mathbf{Y} = [\mathbf{y}(\boldsymbol{\nu}_1), \ldots, \mathbf{y}(\boldsymbol{\nu}_{\mathrm{n}})]$ the $\mathrm{t} \times \mathrm{n}$ fMRI data matrix,

in which $\mathbf{y}(\boldsymbol{v}_i) = [y_1(\boldsymbol{v}_i), \ldots, y_t(\boldsymbol{v}_i)]^\top$ represents the time series recorded at t different time points for brain location $\boldsymbol{v}_i$, $i = 1, \ldots, n$.

Before applying an ICA algorithm to the fMRI data, some preprocessing steps such as centering, dimension reduction and whitening are needed (Hyvärinen et al., 2001). In our analysis, we apply the following linear transformation to the original data:

$$\begin{aligned} \mathbf{Y} &\leftarrow \mathbf{Y}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right), \\ \mathbf{Y} &\leftarrow \left[\left(\mathbf{\Lambda}_{q,Y} - \tilde{\lambda}_{q,Y}\mathbf{I}_q\right)\mathbf{U}_{q,Y}^\top\right]\mathbf{Y}. \end{aligned} \tag{4.1}$$

In the first transformation, we center the fMRI data matrix by column ($\mathbf{1}_n$ is a vector that has n repeated ones). In the second transformation, $\mathbf{\Lambda}_{q,Y}$ is a diagonal matrix that contains the q largest singular values of $\mathbf{Y}$ (the centered data matrix); the columns of $\mathbf{U}_{q,Y}$ are the corresponding singular vectors; $\tilde{\lambda}_{q,Y}$ is the average of the $t - q$ eigenvalues that are not included in $\mathbf{\Lambda}_{q,Y}$. We select the number of component, q, using the MDL criteria (Li et al., 2006). Transformations in (4.1) reduce the temporal domain dimension of the original data from t to q, i.e, we now have $\mathbf{y}(\boldsymbol{v}_i) \in \mathbb{R}^q$ for all $i = 1, \ldots, n$. It also centers and whitens the data such that the sample mean and covariance for columns of $\mathbf{Y}$ are $\mathbf{0}$ and $\mathbf{I}_q$. Throughout the rest part of this section, we consider $\mathbf{Y}$ being the preprocessed fMRI data.

4.1.2 The spatially dependent ICA model for fMRI data

Let $\mathbf{y}(\boldsymbol{v}) = [y_1(\boldsymbol{v}), \ldots, y_q(\boldsymbol{v})]^\top \in \mathbb{R}^q$ be the preprocessed data at an arbitrary brain location $\boldsymbol{v} \in \mathcal{V}$. Our spatial ICA model assumes that

$$\mathbf{y}(\boldsymbol{v}) = \mathbf{A}\mathbf{s}(\boldsymbol{v}), \quad \forall \boldsymbol{v} \in \mathcal{V}, \tag{4.2}$$

in which $\mathbf{s}(\boldsymbol{v}) = [s_1(\boldsymbol{v}), \ldots, s_q(\boldsymbol{v})]^\top$ represents the q latent spatial source signals (ICs); $\mathbf{A}$ is the $q \times q$ mixing matrix. The q elements in $\mathbf{s}(\boldsymbol{v})$ are mutually independent, i.e., $s_\ell(\boldsymbol{v}) \perp s_{\ell'}(\boldsymbol{v}')$ for all $1 \leqslant \ell < \ell' \leqslant q$ and $\boldsymbol{v}, \boldsymbol{v}' \in$

$\mathcal{V}$. Unlike traditional ICA model that assumes $\mathbf{y}(\boldsymbol{\nu})$ to be i.i.d. across all spatial locations, our spatial source signals are spatially dependent, i.e., $\text{Cov}[s_\ell(\boldsymbol{\nu}), s_\ell(\boldsymbol{\nu}')]$ is not constantly zero for all $\boldsymbol{\nu}, \boldsymbol{\nu}' \in \mathcal{V}$, $\boldsymbol{\nu} \neq \boldsymbol{\nu}'$. This assumption is specifically useful for fMRI data because, within each IC, different brain locations can be functionally connected and they can generate similar fMRI BOLD signals instead of being totally independent. At the same time, brain locations close to each other are more likely to demonstrate similar temporal activation patterns. In addition, when analyzing fMRI data, people commonly pre-smooth the raw data using Gaussian kernels, which also introduce artificial spatial dependence structures into the data. By incorporating spatial dependence within each IC, our model is more flexible and can capture these important aspects of fMRI data.

We restrict $\mathbf{A}$ to be an orthogonal matrix. This assumption is commonly adopted for pre-whitened fMRI data (Hyvärinen and Oja, 2000). To make Bayesian inference about the model, we assign priors to $\mathbf{A}$ and $\mathbf{s}(\cdot)$. We begin with the prior of $\mathbf{A}$, the orthogonal mixing matrix.

Prior of the mixing matrix: uniform distribution on the orthogonal group

The set of all orthogonal matrices defined on $\mathbb{R}^{q\times q}$, denoted by $\mathbb{O}(q)$, is a algebraic group. With no prior knowledge available about $\mathbf{A}$, we assign a uniform prior density for $\mathbf{A}$ on $\mathbb{O}(q)$ following Gupta and Nagar (2000) as follows

$$\mathbf{A} \sim \pi_a(\mathbf{A}), \quad \pi_a(\mathbf{A}) = \frac{1}{\text{vol}\,[\mathbb{O}(q)]} \text{I}\,[\mathbf{A} \in \mathbb{O}(q)]\,. \tag{4.3}$$

This prior is proper because the orthogonal group is a Stiefel manifold (Gupta and Nagar, 2000) with a finite volume, $\text{vol}\,[\mathbb{O}(q)]$. To be specific, the volume of $\mathbb{O}(q)$ with regard to the Haar measure is

$$\text{vol}\,[\mathbb{O}(q)] = \pi^{\frac{q+q^2}{4}} \frac{2^q}{\prod_{\ell=1}^{q} \Gamma\left(\frac{q+1-\ell}{2}\right)},$$

according to Corollary 2.1.16 in Muirhead (2005), where $\Gamma(\cdot)$ is the standard Gamma function.

Prior of the ICs: nonparametric Bayesian kernel models

For the latent spatial source signals $\mathbf{s}(\boldsymbol{v}) = [s_1(\boldsymbol{v}), \ldots, s_q(\boldsymbol{v})]^\top$, instead of assuming that each element is drawn identically and independently from a probability distribution regardless of the spatial locations, we model them using nonparametric regressions as follows: for a given $\ell, 1 \leqslant \ell \leqslant q$, and for all $\boldsymbol{v} \in \mathcal{V}$,

$$s_\ell(\boldsymbol{v}) = \mu_\ell(\boldsymbol{v}) + e_\ell(\boldsymbol{v}), \quad e_\ell(\boldsymbol{v}) \overset{\text{iid}}{\sim} \Pi_\ell. \tag{4.4}$$

In (4.4), the mean functions, $\mu_\ell(\boldsymbol{v}), \ell = 1, \ldots, q$, capture the spatially varying trend of the ICs. We model them using the Bayesian kernel models introduced by Pillai et al. (2007); Wolpert et al. (2011) as follows:

$$\mu_\ell(\boldsymbol{v}) = \int_{\boldsymbol{w} \in \mathcal{V}} \kappa(\boldsymbol{v}, \boldsymbol{w}) u_\ell(\mathrm{d}\boldsymbol{w}), \quad \forall \boldsymbol{v} \in \mathcal{V} \tag{4.5}$$

in which $\kappa(\cdot, \cdot) : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{R}$ is a kernel function and $u_\ell(\cdot)$ is a random signed Borel measure defined on $\mathcal{V}$. To be specific, we choose the square exponential (SE) kernel $\kappa(\boldsymbol{v}, \boldsymbol{w}) = \exp\left\{-\|\boldsymbol{v} - \boldsymbol{w}\|^2/\rho\right\}$ in our model. The main result in Pillai et al. (2007) states that the image of this Bayesian kernel model transformation for all $u_\ell(\cdot) \in \mathfrak{B}(\mathcal{V})$, where $\mathfrak{B}(\mathcal{V})$ is the collection of all signed Borel measures on $\mathcal{V}$, equals exactly to the reproducing kernel Hilbert space (RKHS) induced by $\kappa(\cdot, \cdot)$. To construct such a random signed Borel measure, we consider the decomposition of $u_\ell(\cdot)$ as:

$$u_\ell(\mathrm{d}\boldsymbol{w}) = z_\ell(\boldsymbol{w})\, \mathrm{d}F_\ell(\boldsymbol{w}), \quad \forall \boldsymbol{w} \in \mathcal{V}, \tag{4.6}$$

and assume that

$$z_\ell(\boldsymbol{w}) \sim \mathcal{GP}(0, c_\ell(w, w')), \quad F_\ell(\boldsymbol{w}) \sim \mathcal{DP}(\alpha, F_0), \tag{4.7}$$

which states that $z_\ell(\cdot), \ell = 1, \ldots, q$, are independently generated from Gaussian process priors with zero means and covariance kernel functions $c_\ell(\cdot, \cdot)$; $F_\ell(\cdot), \ell = 1, \ldots, q$, are independently drawn from a Dirichlet process prior with a concentration parameter α and a base measure F_0. Independent $z_\ell(\cdot)$ and $F_\ell(\cdot)$ lead to independent signed Borel measures, $u_\ell(\cdot)$, in (4.5). As suggested by Pillai et al. (2007), this generating procedure can cover all the elements in $\mathfrak{B}(\mathcal{V})$. As a result, $\mu_\ell(\boldsymbol{v})$ in (4.4) belongs to the RKHS induced by $\kappa(\cdot, \cdot)$. In fact, for a large family of kernel functions such as the square exponential kernels we use, their RKHS equals to the set of all continuous functions on $\mathcal{V}$ (Tokdar and Ghosh, 2007).

The independence between these random signed Borel measures guarantees the independence between ICs. On the other hand, this construction leads to spatial dependence within each specific IC. This dependence structure is given in Theorem 5, the proof of which is relegated to the Appendix.

Theorem 5. *For the ℓth IC, i.e., $s_\ell(\boldsymbol{v})$, in our spatial ICA model* (4.2), (4.5)–(4.7) *implies that*

$$Cov[s_\ell(\boldsymbol{v}), s_\ell(\boldsymbol{v}')] = \frac{1}{|\mathcal{V}|^2} \iint_{(\boldsymbol{w}, \boldsymbol{w}') \in \mathcal{V} \times \mathcal{V}} \kappa(\boldsymbol{v}, \boldsymbol{w}) c_\ell(\boldsymbol{w}, \boldsymbol{w}') \kappa(\boldsymbol{w}', \boldsymbol{v}') \, d\boldsymbol{w} \, d\boldsymbol{w}',$$

where $|\mathcal{V}|$ is the volume of $\mathcal{V}$ with regard to the Lebesgue measure.

Theorem 5 implies that, within each of the ICs, even the kernel functions in model (4.5) and in the Gaussian processes for generating the signed Borel measures in (4.7) are stationary, the resulting spatial dependence structure for the ICs is still non-stationary, which admit extra flexibility from the modeling perspective.

The noise terms in (4.4) is modeled by the scale transformed Student-t distribution as proposed in many Bayesian literature for robust inference (West, 1984; Lange et al., 1989; Fonseca et al., 2008). Specifically, $e_\ell(\boldsymbol{v}) \sim \Pi_\ell$,

in which

$$\Pi_\ell = \frac{\Gamma\left(\frac{\nu_\ell+1}{2}\right)}{\Gamma\left(\frac{\nu_\ell}{2}\right)\sqrt{\nu_\ell \pi \sigma_\ell^2}}\left(1+\frac{x^2}{\sigma_\ell^2 \nu_\ell}\right)^{-\frac{\nu_\ell+1}{2}}. \tag{4.8}$$

This specification implies that $\frac{e_\ell(\nu)}{\sqrt{\sigma_\ell^2}} \sim t(\nu_\ell)$, a t distribution with ν_ℓ degree of freedom. A canonical heteroscedastic error term augmentation for this distribution is:

$$e_\ell(\nu) \mid \phi_\ell(\nu) \sim N\left(0,\ \frac{\sigma_\ell^2}{\phi_\ell(\nu)}\right), \quad \phi_\ell(\nu) \sim \text{Gamma}\left(\frac{\nu_\ell}{2},\ \frac{2}{\nu_\ell}\right). \tag{4.9}$$

We will adopt this generating process for $e_\ell(\nu)$ in our expositions throughout the rest of this paper.

4.1.3 Model representation, hyperprior specification and posterior inference

Model representation for finite number of observations

When analyzing real fMRI data, we only have a finite number of preprocessed observations, as described in section 4.1.1, at the n spatial locations: $y(\nu_1), \ldots, y(\nu_n)$, which all satisfy that $y(\nu_i) = As(\nu_i)$. Let A_ℓ be the ℓth column of the orthogonal mixing matrix A, then

$$s_\ell(\nu_i) = A_\ell^\top y(\nu_i). \tag{4.10}$$

The results from Liang et al. (2006); Pillai et al. (2007) imply that the DP prior specification in the kernel model (4.5) and (4.6) leads to an approximation at the finite number of brain locations, $\nu_1, \ldots, \nu_n$ as

$$s_\ell(\nu_i) \approx \sum_{j=1}^{n} z_{\ell,j} K_j(\nu_i) + e_\ell(\nu_i), \tag{4.11}$$

if the fraction, α/n, is small, which holds true for most imaging applications since n, the number of brain locations (or voxels), are very large while α is relatively small. In (4.11), $K_j(\nu_i) = \kappa(\nu_i, \nu_j)$ represents the value of kernel function $\kappa(\cdot,\cdot)$ evaluated at ν_i and ν_j; $z_{\ell,j} = z_\ell(\nu_j)$ is the value of

the Gaussian process $z_\ell(\cdot)$ defined in (4.7) at brain location $\boldsymbol{\nu}_j$.

Since in model (4.11), we still have a total number of n variables, $z_{\ell,j}$, to be estimated from n spatial observations, we assign the generalized Zellner's g-prior (Zellner, 1986; West, 2003; Maruyama and George, 2011) to the vector $\boldsymbol{z}_\ell = [z_{\ell,1}, \ldots, z_{\ell,n}]^\top$ to facilitate fully Bayesian inference and shrink small elements in $\boldsymbol{z}_\ell$ towards zeros. Specifically, we adopt the strategy of West (2003), which aims to achieve shrinkage estimation of $z_{\ell,1}, \ldots, z_{\ell,n}$ in different principal component directions on the design space induced by $\kappa(\cdot\,,\cdot)$. The prior on $\boldsymbol{z}_\ell$ is the given as follows

$$\boldsymbol{z}_\ell \sim \mathrm{N}_n(\boldsymbol{0}, \tilde{\boldsymbol{\Psi}}\tilde{\boldsymbol{\Lambda}}^{-1}\mathbf{G}_\ell\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{\Psi}}^\top), \tag{4.12}$$

where $\mathbf{K} = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^\top$ is the eigen-decomposition of the $n \times n$ empirical kernel matrix $\mathbf{K} = \{\kappa(\boldsymbol{\nu}_i, \boldsymbol{\nu}_j)\}_{1 \leqslant i,j \leqslant n}$; $\tilde{\boldsymbol{\Lambda}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K), \lambda_k > 0$ represents the diagonal matrix containing the K largest eigenvalues in $\boldsymbol{\Lambda}$, while $\tilde{\boldsymbol{\Psi}} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K]$ is an $n \times K$ matrix containing the corresponding eigenvectors in $\boldsymbol{\Psi}$ as its columns; $\mathbf{G}_\ell = g_\ell \mathbf{I}_K$ where $g_\ell > 0$ is g-parameter for the ℓth IC. Suppose that the Karhunen-Loève (KL) expansion (Rasmussen and Williams, 2006b) of $\kappa(\boldsymbol{\nu}, \boldsymbol{\nu}')$ is $\kappa(\boldsymbol{\nu}, \boldsymbol{\nu}') = \sum_{k=1}^{\infty} \lambda_k \psi_k(\boldsymbol{\nu})\psi_k(\boldsymbol{\nu}')$, then the prior specification for $\boldsymbol{z}_\ell$ in (4.12) implies that $c_\ell(\boldsymbol{\nu}, \boldsymbol{\nu}') = \sum_{k=1}^{K} \frac{g_\ell}{\lambda_k^2}\psi_k(\boldsymbol{\nu})\psi_k(\boldsymbol{\nu}')$ in (4.7).

Based on the notations above, if we define vectors $\boldsymbol{\beta}_\ell = [\beta_{\ell,1}, \ldots, \beta_{\ell,K}]^\top$, $\ell = 1, \ldots, q$, such that $\mathbf{K}\boldsymbol{z}_\ell = \tilde{\boldsymbol{\Psi}}\boldsymbol{\beta}_\ell$, then $\boldsymbol{\beta}_\ell \sim \mathrm{N}_K(0, \mathbf{G}_\ell)$ approximately since $\mathbf{K} \approx \tilde{\boldsymbol{\Psi}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Psi}}^\top$. Based on the re-parameterization above, we approximate (4.11) by

$$s_\ell(\boldsymbol{\nu}_i) = \sum_{k=1}^{K} \beta_{\ell,k}\psi_{k,i} + e_\ell(\boldsymbol{\nu}_i), \quad \beta_{\ell,k} \sim \mathrm{N}(0,\ g_{\ell,k}) \tag{4.13}$$

in which $\psi_{k,i}$ are elements in $\boldsymbol{\psi}_k = [\psi_{k,1}, \ldots, \psi_{k,n}]^\top$, the columns of $\tilde{\boldsymbol{\Psi}}$. The noise term $e_\ell(\boldsymbol{\nu}_i) \overset{\text{iid}}{\sim} \Pi_\ell$, follows the scale transformed t-distribution as given in (4.8), for all i. According to (4.9), we have the following het-

eroscedastic representation of $e_\ell(v_i)$ as follows

$$e_\ell(\boldsymbol{v}_i) \mid \phi_{\ell,i} \sim \mathrm{N}\left(0,\ \frac{\sigma_\ell^2}{\phi_{\ell,i}}\right), \quad \phi_{\ell,i} \sim \mathrm{Gamma}\left(\frac{\nu_\ell}{2},\ \frac{2}{\nu_\ell}\right). \tag{4.14}$$

Approximate computation using a modified square exponential exponential kernel

Functional eigen-decomposition of the kernel function $\kappa(\boldsymbol{v}, \boldsymbol{w}) = \exp\{-\|\boldsymbol{v}-\boldsymbol{w}\|^2/\rho\}$ does not have an analytical solution, which makes it computationally infeasible to fit our model to large imaging datasets where n might be greater than 10^5. To solve this issue, we introduce the modified square exponential kernel

$$\kappa^{\mathrm{mod}}(\boldsymbol{v}, \boldsymbol{w}) = \exp\left\{-a\|\boldsymbol{v}\|_2^2 - a\|\boldsymbol{w}\|_2^2 - \frac{\|\boldsymbol{v}-\boldsymbol{w}\|_2^2}{\rho}\right\},\ a, \rho > 0 \tag{4.15}$$

with a relatively small value for a as a numerical approximation to the square exponential kernel when dealing with massive neuroimaging data. The major benefit of this kernel function is that it has analytically tractable functional eigen-decomposition fomula. The detailed properties of this kernel is summarized in the Appendix. In this paper, we use this kernel function for approximate computation by fixing $a = 0.25$, which gives us competitive empirical performance. The choice of the scale parameter ρ in this kernel function is discussed below.

Choosing the scale parameter in the kernel function

Typically, this parameter can be chosen using a fully Bayesian approach by updating it with full conditional draws while sampling from the posterior. This fully Bayesian updating scheme will lead to changes of the dimensionality of our parameter space when working with representation (4.13), which required special treatment such as using the reversible jump MCMC (Green, 1995). Another way is to estimate ρ under the empirical Bayes paradigm, which requires repetitive posterior inferences about the

other parameters given different choices of ρ.

Since both approaches are know to be complex and inefficient, we propose an *ad-hoc* procedure to choose the scale parameter in the kernel function, which works well in practice. Specifically, for a chosen value of ρ and the resulting decomposition of $\mathbf{K}$, we use an existing ICA method to decompose $\mathbf{Y}_{q\times n} = [\mathbf{y}(\boldsymbol{\nu}_1), \ldots, \mathbf{y}(\boldsymbol{\nu}_n)]$, with the resulting unmixing matrix $\widetilde{\boldsymbol{W}}$. Off note, we use the Infomax ICA (Bell and Sejnowski, 1995) to estimate $\widetilde{\boldsymbol{W}}$ in our empirical study. We then calculate the generalized cross-validation (GCV) loss function defined as follows

$$\mathcal{L}(\rho) = \frac{\frac{1}{n}\left\| \left[\mathbf{I}_n - \tilde{\boldsymbol{\Psi}} \left(\tilde{\boldsymbol{\Psi}}^\top \tilde{\boldsymbol{\Psi}} \right)^{-1} \tilde{\boldsymbol{\Psi}}^\top \right] \mathbf{Y}^\top \widetilde{\boldsymbol{W}}^\top \right\|_{\mathcal{F}}}{\left\{ \frac{1}{n} \mathrm{Tr} \left[\mathbf{I}_n - \tilde{\boldsymbol{\Psi}} \left(\tilde{\boldsymbol{\Psi}}^\top \tilde{\boldsymbol{\Psi}} \right)^{-1} \tilde{\boldsymbol{\Psi}}^\top \right] \right\}^2}, \tag{4.16}$$

and choose ρ as the minimizer of $\mathcal{L}(\rho)$.

Choice of hyperpriors and posterior inference

We choose conjugate priors for σ_ℓ^2 by assuming $\sigma_\ell^2 \overset{\text{iid}}{\sim} \text{Inv-Ga}(a_\sigma, b_\sigma)$ for all $\ell = 1, \ldots, q$ in (4.14), where $\text{Inv-Ga}(\cdot, \cdot)$ is the inverse Gamma distribution. In (4.14), the priors on ν_ℓ are given as $\text{Inv-Ga}(a_\nu, b_\nu)$. We let $a_\sigma, b_\sigma, a_\nu, b_\nu$ be small numbers, say 0.001 (the values we use throughout this paper), to generate relatively uninformative priors on σ_ℓ^2 and ν_ℓ. The priors on g_ℓ is $g_\ell \overset{\text{iid}}{\sim} \text{Inv-Ga}(\eta_\ell,\ \xi_\ell)$ for all $1 \leqslant \ell \leqslant q$. We specify inverse Gamma priors on η_ℓ and ξ_ℓ as $\eta_\ell \overset{\text{iid}}{\sim} \text{Inv-Ga}(a_\eta,\ b_\eta)$, $\xi_\ell \overset{\text{iid}}{\sim} \text{Inv-Ga}(a_\xi,\ b_\xi)$ and also let $a_\eta = b_\eta = a_\xi = b_\xi = 0.001$ to reduce prior information.

Combining (4.3), (4.10), (4.13), (4.14) and the hyperprior specifications

above, we have:

$$\begin{aligned}
\boldsymbol{A}_\ell^\top \boldsymbol{y}(\boldsymbol{v}_i) \mid \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma_\ell^2 &\sim \mathrm{N}\left(\sum_{k=1}^{K} \beta_{\ell,k}\psi_{k,i}\ \frac{\sigma_\ell^2}{\phi_{\ell,i}}\right), \\
\boldsymbol{A} \sim \pi_a(\boldsymbol{A}), \quad \sigma_\ell^2 &\sim \text{Inv-Ga}(a_\sigma, b_\sigma), \\
\phi_{\ell,i} \sim \text{Gamma}\left(\frac{\nu_\ell}{2},\ \frac{2}{\nu_\ell}\right), \quad \nu_\ell &\sim \text{Inv-Ga}(a_\nu,\ b_\nu), \\
\beta_{\ell,k} \sim \mathrm{N}(0,\ g_\ell), \quad g_\ell &\sim \text{Inv-Ga}\left(\eta_\ell,\ \xi_\ell\right), \\
\eta_\ell \overset{\text{iid}}{\sim} \text{Inv-Ga}(a_\eta,\ b_\eta), \quad \xi_\ell &\sim \text{Inv-Ga}(a_\xi,\ b_\xi).
\end{aligned} \tag{4.17}$$

Our goal then is to examine the joint posterior distribution given the data matrix $\boldsymbol{Y}$, which contains all the preprocessed fMRI data. We approximate the posterior by drawing samples from

$$\left[\boldsymbol{A},\ \boldsymbol{\beta},\ \boldsymbol{\phi},\ \{\sigma_\ell^2,\ g_\ell,\ \nu_\ell, \eta_\ell,\ \xi_\ell\}_{\ell=1}^{q} \mid \boldsymbol{Y}\right]$$

using a Markov chain Monte Carlo (MCMC) algorithm to achieve this goal. More details about this algorithm is given in the Appendix.

Given the MCMC samples, we aim to achieve two major goals. The first is to estimate the orthogonal mixing matrix $\boldsymbol{A}$. Denote by $\boldsymbol{A}^{(1)}, \ldots, \boldsymbol{A}^{(T)}$ the valid the posterior samples for $[\boldsymbol{A} \mid \boldsymbol{Y}]$, we estimate this matrix as follows:

$$\widehat{\boldsymbol{A}} = \text{orth}(\boldsymbol{A}^*), \quad \boldsymbol{A}^*{}_{i,j} = \text{median}\{\boldsymbol{A}_{i,j}^{(1)}, \ldots, \boldsymbol{A}_{i,j}^{(T)}\},$$

where $\text{orth}(\cdot)$ is the orthogonalization operator. The use of posterior median is due to the heavy tails of $[\boldsymbol{A}_{i,j} \mid \boldsymbol{Y}]$, i.e., the marginal posterior of each element of $\boldsymbol{A}$, which we frequently observe in practice. The second goal is to make inference about the spatial source signals. Consider for the ℓth channel at brain location $\boldsymbol{v}_i$, posterior samples, $\boldsymbol{\beta}_\ell^{(1)}, \ldots, \boldsymbol{\beta}_\ell^{(T)}$ enable us to estimate the spatial source signals using the posterior of mean process $\mu_\ell(\boldsymbol{v}_i)$ as

$$\widehat{\mathbb{E}}[\mu_\ell(\boldsymbol{v}_i) \mid \boldsymbol{Y}] \approx \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \beta_{\ell,k}^{(t)}\psi_{k,i}$$

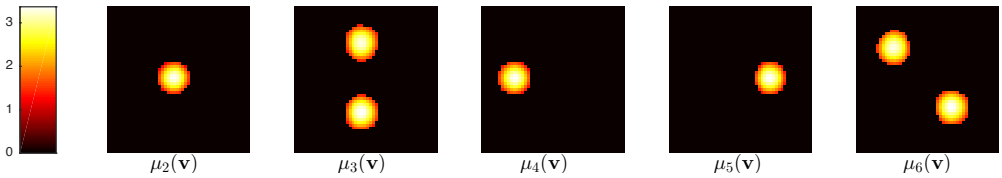


Figure 4.1: The true mean functions of the spatial sources signals used in our simulation studies. The mean function $\mu_1(s)$ is constantly zero.

Table 4.1: Summary of the data generating processes for $s(v)$ in the simulation study: The ean functions referred to here are all displayed in Figure 4.1; the nontaion DE(λ) stands for a double exponential distribution with rate parameter λ; the notation Gamma(k, θ) represents a gamma distribution with k being the shape parameter and θ being the scale parameter. σ parameter here is used to control the noise levels.

	IC channel	mean function	error distribution: $\epsilon(s)/\sigma \overset{iid}{\sim}$
$q = 2$	1	0	$N(0,1)$
	2	$\mu_2(s)$	$\{\chi^2(3)-3\}/\sqrt{6}$
$q = 4$	1	0	$N(0,1)$
	2	$\mu_2(s)$	$\{\chi^2(3)-3\}/\sqrt{6}$
	3	$\mu_3(s)$	$DE(1)/\sqrt{2}$
	4	$\mu_4(s)$	$\frac{1}{2}N\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right) + \frac{1}{2}N\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$
$q = 6$	1	0	$N(0,1)$
	2	$\mu_2(s)$	$\{\chi^2(3)-3\}/\sqrt{6}$
	3	$\mu_3(s)$	$DE(1)/\sqrt{2}$
	4	$\mu_4(s)$	$\frac{1}{2}N\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right) + \frac{1}{2}N\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$
	5	$\mu_5(s)$	$\{Gamma(4,2)-8\}/4$
	6	$\mu_6(s)$	$\{\chi^2(2)-2\}/2$

A specific 95% credible interval for the spatial source signals can be constructed from the 2.5%th and 97.5%th percentiles of

$$\left\{ \sum_{k=1}^{K} \beta_{\ell,k}^{(1)} \psi_{k,i}, \ldots, \sum_{k=1}^{K} \beta_{\ell,k}^{(T)} \psi_{k,i} \right\}$$

at brain location v_i. Voxels that have credible intervals excluding zero will be selected as activated areas.

4.2 Data Examples

4.2.1 Simulated data

In this section, we demonstrate the performance of our proposed method through simulation studies. We consider the number of ICs, q, to be 2, 4 and 6. For each specification of q, the first IC channel is a Gaussian white noise channel, i.e., $\mu_1(\mathbf{s}) \equiv 0$, $e_1(\mathbf{s})/\sigma \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. For the rest of the channels, the mean functions, i.e., $\mu_\ell(\boldsymbol{\nu}), \ell = 2, \ldots, q$, in model (4.4), is chosen from the five images shown in Figure 4.1, which are all 50×50 images. Different types of white noises are added to these mean functions to generate the ICs. Table 4.1 provides a detailed summary about the ICs in the data generating processes of our simulation studies. As we can see from Table 4.1, parameter σ governs the noise level of simulated datasets. By varying σ, we consider three different noise levels: low ($\sigma = 0.5$), medium ($\sigma = 1$) and high ($\sigma = 1.5$). Combining different values of q and σ, we have nine different scenarios. Under each scenario, a $q \times q$ mixing matrix, $\mathbf{A}$, is randomly generated via the orthogonalization of a square matrix with elements independently sampled from $\mathrm{N}(0,1)$. For each pair of (q, σ) combinations, 50 datasets are simulated.

We specify the modified SE kernel for our model and choose the scale parameter using the GCV loss criteria. We fit our model to the simulated datasets using MCMC. In practice, the chain mixes fast and we draw $2,000$ MCMC samples and discard the first 500 samples as burn-in observations. We also implement six other ICA algorithms including "FastICA" (Hyvärinen, 1999), "Infomax" ICA (Bell and Sejnowski, 1995), "JADE" ICA (Hyvärinen, 1999), kernel ICA (Bach and Jordan, 2003), consistent ICA based on empirical characteristic function (Chen and Bickel, 2005) and fast kernel density ICA (Chen, 2006).

To evaluate the accuracies of the mixing matrix estimates, we use the AMARI distance criteria (Amari et al., 1996), which has also been used in Bach and Jordan (2003); Chen and Bickel (2005). Specifically, let $\mathbf{A}_0$ be

the $q \times q$ true orthogonal mixing matrix and let $\widehat{\mathbf{A}}$ be an estimate, then the Amari error of this estimated mixing matrix is

$$d_{Amari}(\boldsymbol{A}_0, \widehat{\mathbf{A}}) = \frac{1}{2q}\sum_{i=1}^{q}\left(\frac{\sum_{j=1}^{q}|r_{i,j}|}{\max_j|r_{i,j}|} - 1\right) + \frac{1}{2q}\sum_{j=1}^{q}\left(\frac{\sum_{i=1}^{q}|r_{i,j}|}{\max_i|r_{i,j}|} - 1\right), \quad (4.18)$$

where $r_{i,j} = (\boldsymbol{A}_0^{\top}\widehat{\mathbf{A}})_{i,j}$. Figure 4.3 shows the boxplots of Amari errors across the 50 replications from our method as well as from the other competing methods. Off note, when an algorithm fails to converge, which happens frequently for the fastICA algorithm under our simulation settings, we will just treat the resulting Amari error as a missing value. We can see from Figure 4.3 that our method consistently generates more accurate mixing matrix estimates. The advantage of our model in terms of estimation accuracy is more pronounced under the high noise level settings ($\sigma = 1.5$).

In addition, an important goal when applying ICA to many fMRI studies is to identify the activation area within each IC. To evaluate the performance of different methods in terms of recovering the spatial activation patterns, we draw receiver operating characteristic (ROC) curves in Figure 4.4 for determining activated pixels for different ICA methods. Off note, fastICA is excluded here because it can fail to converge frequently under many scenarios. The large area under the curve (AUC) of our method in Figure 4.4 across all simulation settings implies that our method can separate spatial source signals from noise or background better very well.

Average computational time in second is reported in Figure 4.5. We can see that our method, given its larger number of iterations required, can complete model estimation and inference within a reasonable amount of time.

Simulated data: robustness checks

In addition to the simulation results reported above, we conduct further robustness check to evaluate the performance of our method against the competing ones under three additional settings: residuals estimated from real data; mean functions with sharper edges; mean functions with smaller areas of activation. We consider $\mathrm{q} = 4$ for under these three settings.

Under the first setting, we begin by using the Infomax ICA to analyze real fMRI data in our study and extracting residuals for four ICs of interest. Then we generate error terms $e_\ell(\boldsymbol{\nu})$ by randomly choosing $2,500$ voxels. This simulation reflects the true noise level and spatial dynamics of the error terms indicated by the real data. In order to cover more voxels, we replicate the simulation for 100 times.

The first panel in Figure 4.6 demonstrates the results under this setting, which confirms that our method still dominates the others in terms of both estimating mixing matrix and identifying spatially activated regions.

Under the second setting, instead of imposing gradually varying structures to the mean functions as shown in Figure 4.1, we force the signals within activated areas to be 2, in order to discourage spatial dependence, especially on the boundaries. Under the third setting, we reduce the area of activation across source signals (three quarters of activated pixels shown in Figure 4.1 are forced to be zero, the resulting ratio of activation is 1.24% and the original one is 5.00%). The spatial pattern of these new sources are presented in Figure 4.2. This specification aims to decrease the amount of information that helps our method to borrow due to neighborhood dependence. For both of these two settings, we let the noise parameter σ be 0.5, corresponding to the lowest noise level in our previous simulation studies, which favors competing methods.

The second panel in Figure 4.6 shows the results under the sharp edge

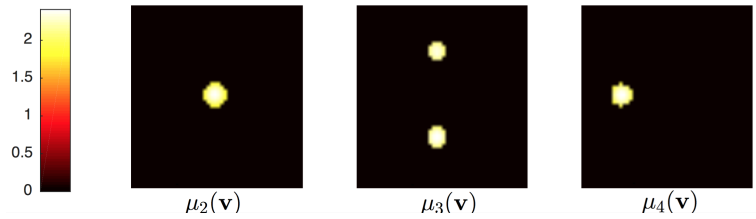


Figure 4.2: The new true mean functions of the spatial sources signals used in additional simulation studies (setting three: smaller area of activation). The mean function $\mu_1(\mathbf{s})$ is constantly zero.

setting. Under this setting, our method still performs very well in terms of both estimating $\mathbf{A}$ and identifying activations within $\mathbf{s}(\cdot)$. At the same time, with constant strength of activation featuring larger values around the boundary, traditional ICA methods tend to have a better performance in terms of finding the activated pixels. This is because the amplified signals around the boundary are more distinguishable from zero. The third panel in Figure 4.6 shows the results when the activation area becomes smaller. Under this setting, our method is still competitive while the kernel ICA and fast kernel density ICA now give us the best mixing matrix estimates. Kernel ICA also gives us the best ROC curve and our method is comparable to the fast kernel density ICA in terms of ROC analysis. This phenomenon is intuitive by thinking about the limiting scenario: When the true activation area keeps becoming smaller ($\sup_{\boldsymbol{\nu}\in\mathcal{V}}|\mu_\ell(\boldsymbol{\nu})|\to 0$, $\ell = 1,\ldots,q$), the source signals will be closer to iid white noises, which agrees better with the assumption of existing ICA methods.

4.2.2 Real resting-state fMRI data

We demonstrate our method using a resting-state fMRI (rs-fMRI) data example from the Philadelphia Neuro-developmental Cohort (PNC) study. This study spans over $9,500$ young individuals aged between eight and twenty-one. Among all the participants, $1,445$ individuals received neuroimaging scans including resting-state fMRI. One appealing feature of the PNC study is that all their imaging data were acquired on the same scanner (Siemens Tim Trio 3 Tesla, Erlangen, Germany; 32 channel head

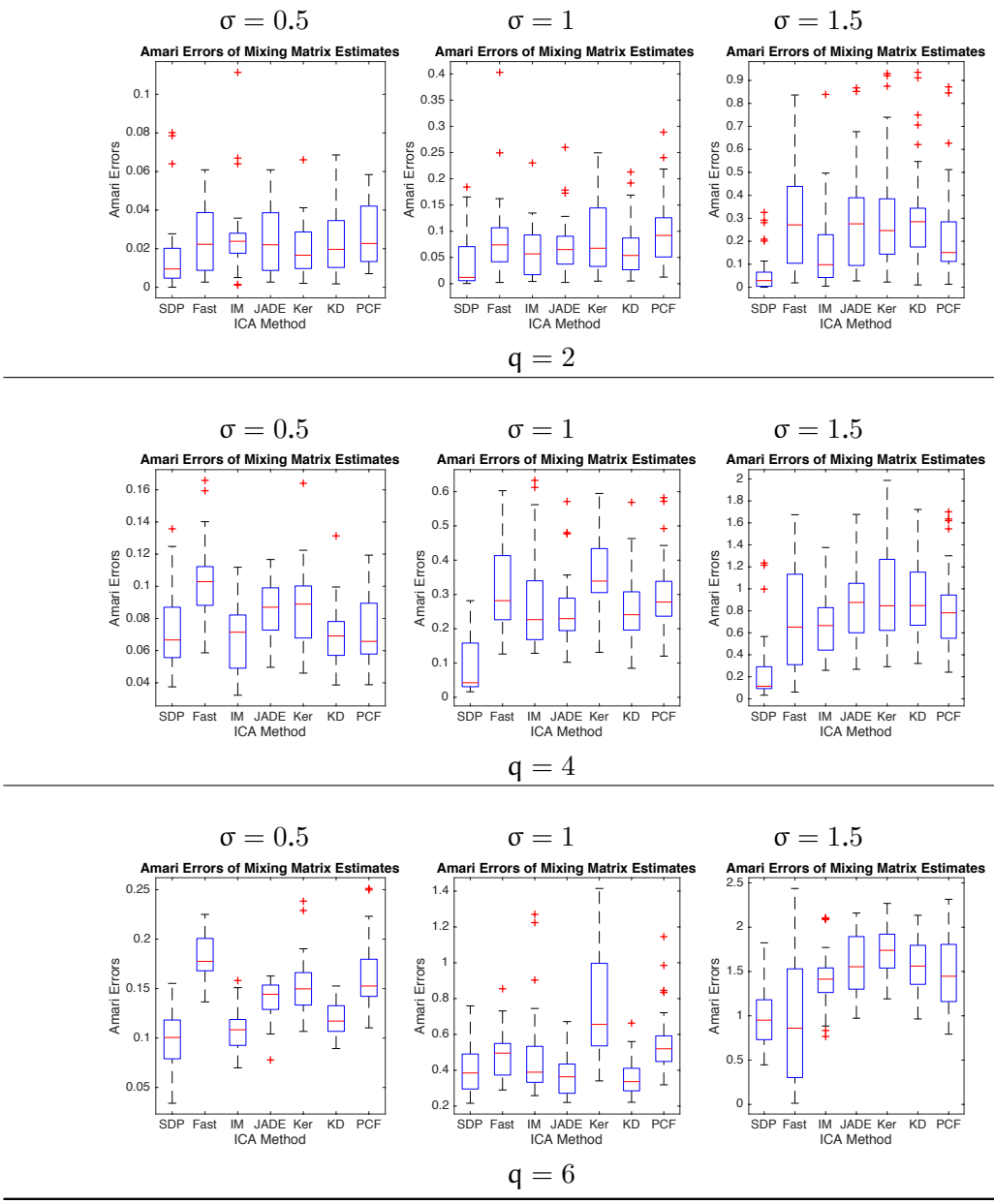
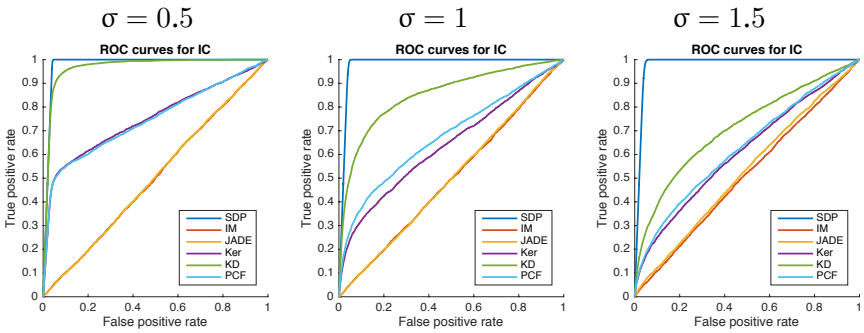
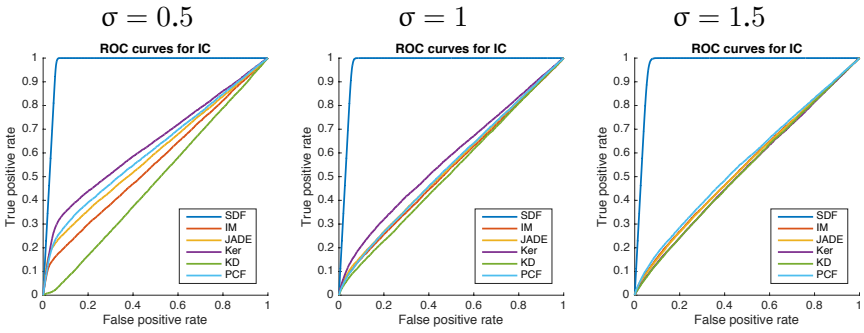


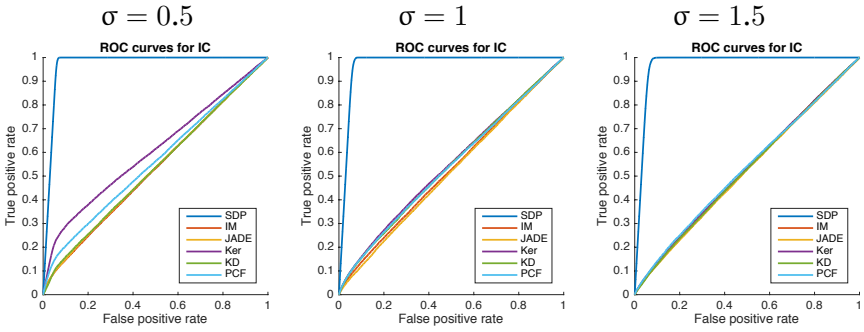
Figure 4.3: The Amari errors from different ICA methods. SDP: our ICA method with spatially dependent sources; Fast: FastICA; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function



q = 2



q = 4



q = 6

Figure 4.4: The ROC curves for different ICA methods. SDP: our ICA method with spatially dependent sources; Fast: FastICA; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function

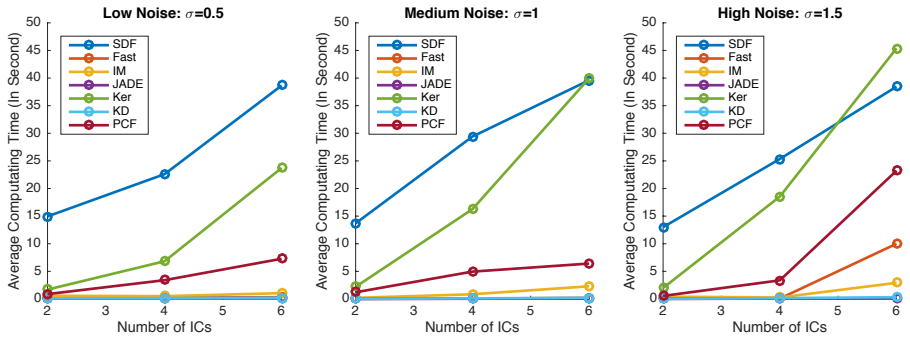


Figure 4.5: The average running time for different ICA methods (FastICA is excluded for its convergence issues). SDP: our ICA method with spatially dependent sources; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function.

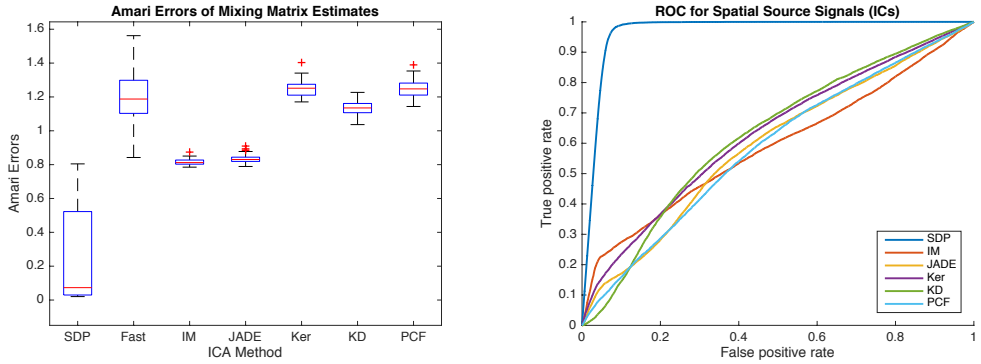
coil) using the same imaging sequences. More detailed description about this dataset is available in (Satterthwaite et al., 2014).

Prior to the analysis, we removed subjects who had more than 20 volumes with relative displacement larger than 0.25mm to avoid subject-specific excessive motion (Satterthwaite et al., 2015). Out of all the subjects who received brain scans, 515 participants met this inclusion criterion and their rs-fMRI data were used in our analysis.

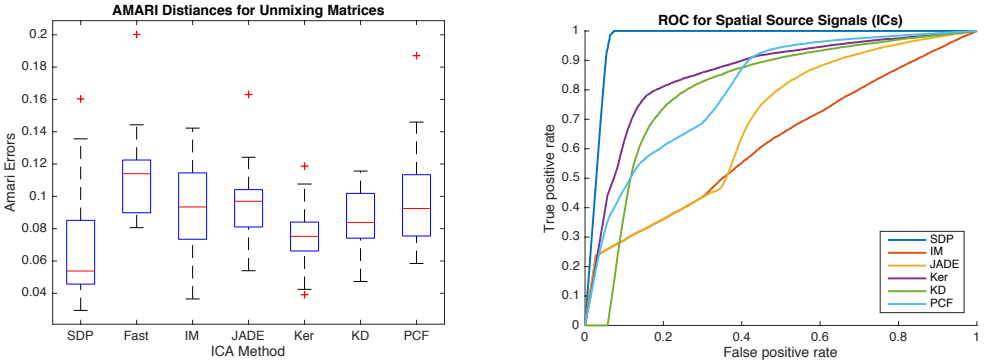
Our research question is to identify fundamental patterns of functional connectivities or brain functional networks in the resting states among the PNC subjects. To achieve our goal, we adopt a group ICA procedure which starts with a two-stage dimension reduction step (Beckmann and Smith, 2005; Calhoun et al., 2001; Guo and Pagnoni, 2008). Specifically, we denote by $\widetilde{Y}_j$ the original $t \times n$ fMRI data matrices for subject j, $j = 1, \ldots, 515$, in the study, where $t = 120$ is the number of scans over time and $n = 18,5405$ is the number of voxels recorded in the gray matter. The in the first stage dimension reduction, we transform the original data as follows

$$Y_j^* = U_R^\top \widetilde{Y}_j,\ j = 1, \ldots, 515$$

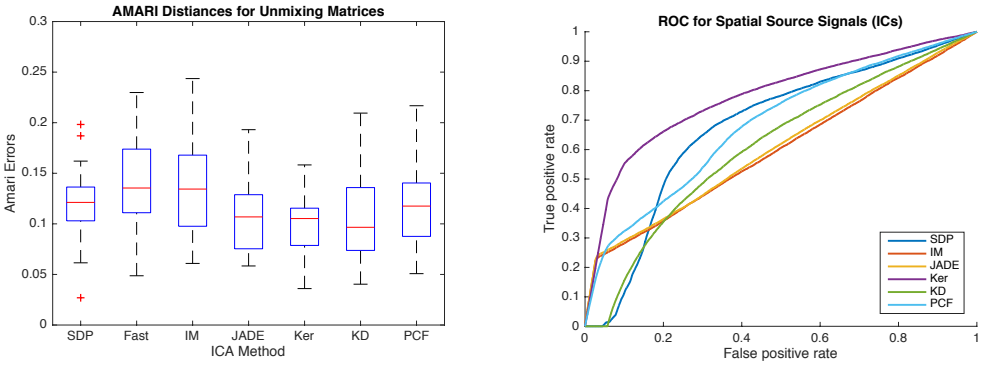
in which columns of U_R $(R = 20)$ are eigen-vectors corresponding to the top twenty dominating eigen-values of $\sum_{j=1}^{515} Y_j Y_j^\top$. We the concatenate



Simulation with residuals from real data



Simulation with constant values for activated regions



Simulation with smaller areas of activation

Figure 4.6: Additioanl simulation studies for robustness checking. SDP: our ICA method with spatially dependent sources; Fast: FastICA; IM: Informax ICA; JADE: JADE ICA; Ker: kernel ICA; KD: fast kernel density ICA; PCF: ICA based on empirical characterist function

data from the previous step as $\boldsymbol{Y}^* = [\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_{515}^\top]^\top$ and apply the joint centering, dimension reduction as well as whitening pre-process routines described in Section 4.1.1 to $\boldsymbol{Y}^*$. The number of ICs in the second stage processing is chosen as $q = 8$ using the GIFT Matlab toolbox (`http://mialab.mrn.org/software/gift/index.html`) after artifact removal. The resulting data matrix is the feed into our proposed ICA method as well as other competing methods.

We use the modified square exponential kernel in our model. We fit our model to the fMRI data using MCMC with $12,000$ iterations (first $4,000$ samples discarded; samples recorded every 10 steps). We choose the scale parameter by plotting the logarithms GCV loss against different (logarithm) values of the parameter. Figure 4.7 shows this plot, which implies an optimal choice of ρ as $e^{-4.4} = 0.0123$. Our method can successfully recover four biologically meaningful brain functional networks including the right parietal frontal network, the default mode network, the primary visual cortex and the auditory network. All these networks are prominent findings among various rs-fMRI studies (Smith et al., 2009). We show in Panel (A) of Figure 4.8 brain slices featuring the structures of these networks. In Figure 4.8, All the spatial source signals quantifying these networks are thresholded by excluding voxels that have Bayesian credible intervals containing zero. For comparison, we also report results from two competing ICA algorithms: the Infomax ICA (Bell and Sejnowski, 1995), as a representative of the parametric approaches, and the empirical characteristic function based consistent ICA (Chen and Bickel, 2005), as a representative of the semi-parametric/nonparametric methods. Since these alternative methods do not allow statistically valid inference for the ICs, we threshold the resulting spatial source signals based on the heuristic z-scores defined in (Mckeown et al., 1998), which has no valid statistical interpretations, to determine activation patterns of the functional networks. Panel (B) and Panel (C) of Figure 4.8 shows the network structures for these two competing methods. By comparing three panels in

Figure 4.8, we can see that our method produce more pronounced estimates of brain activities at those characteristic regions. At the same time, our method can recover the activation regions within the brain functional networks with high completeness.

4.3 Discussion

In this paper, we propose a new ICA model featuring spatially dependent source signals for fMRI data analysis. We model the ICs by nonparametric regression using Bayesian kernel models. We adopt a fully Bayesian inference procedures using MCMC to estimate the mixing matrix and make inference about the spatial source signals using posterior credible intervals. When the noise level is high and the data have distinct spatial dependence structures, our method can estimate the mixing matrix and identify activated regions within IC sources more accurately compared with existing methods.

This study can be improved in two aspects. First, we can evaluate the theoretical properties of our posterior inference from a frequentist perspective by studying its posterior consistency and rate of contraction. A major difficulty to overcome along this direction is to come up with good testing functions with exponentially decaying tails with regard to the number of spatial observations. Second, we can develop computationally more efficient algorithms, using variational approximation to the true posterior, to make inference about our model. This direction is extremely useful for handling big neuroimaging data. To achieve this goal, one major problem is how to propose a good specification to the approximate marginal posterior of the mixing matrix $\mathbf{A}$. In practice, a good specification needs to mimic the true posterior while facilitates tractable integration with regard to the mixing matrix.

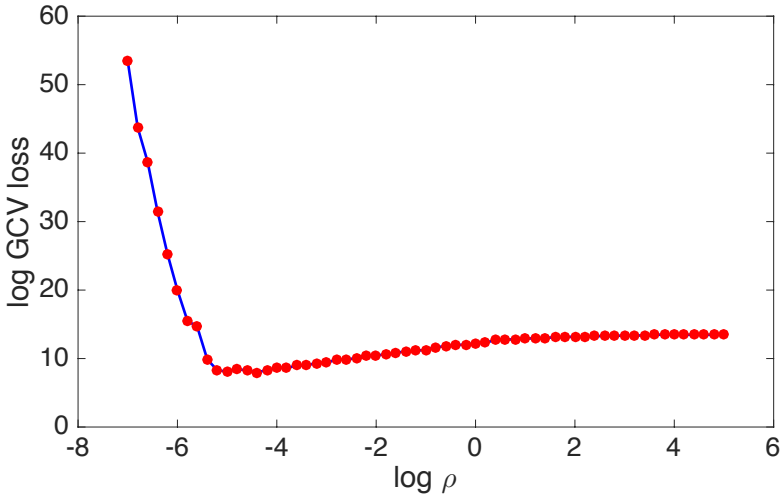


Figure 4.7: Plot the log GCV losses against different values of ρ, the scale parameter in the kernel function, for the PNC dataset.

4.4 Appendices

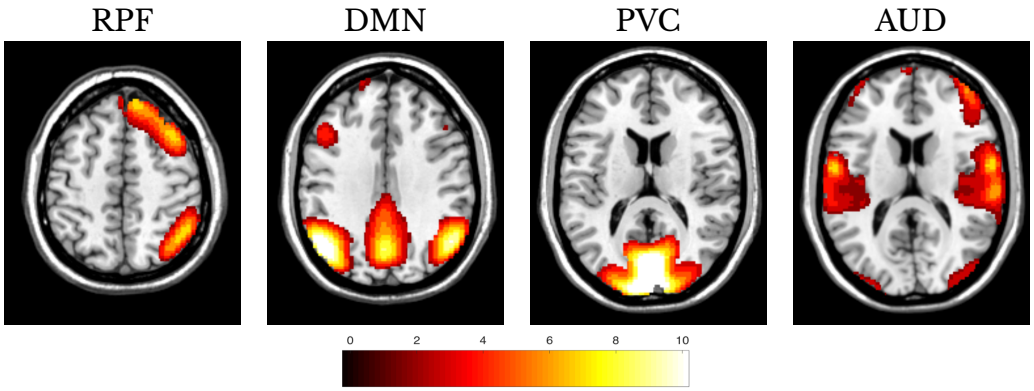
4.4.1 Proof of Theorem 5

For the ease of exposition, we drop the subscript ℓ here. Since $s(v) = \int_{w \in \mathcal{V}} \kappa(v, w) u(\mathrm{d}w) + e(v)$ with $\mathbb{E}[e(v)] = 0$, then

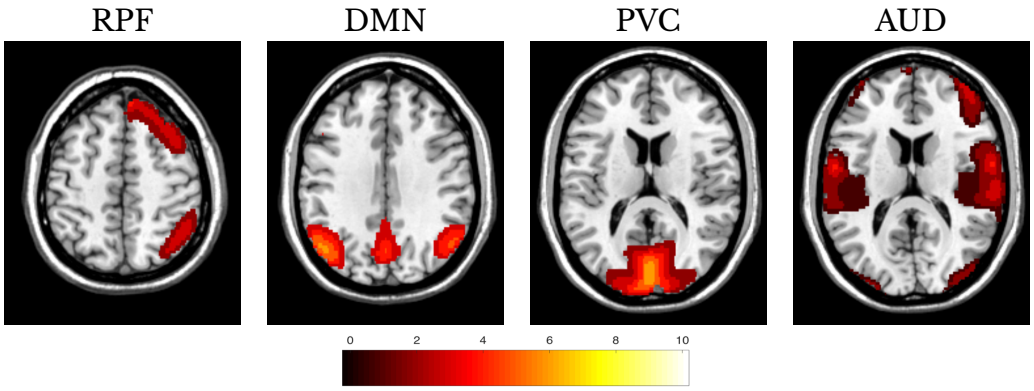
$$\begin{aligned} \mathbb{E}[s(v)] &= \mathbb{E}\left[\int_{w \in \mathcal{V}} \kappa(v, w) u(\mathrm{d}w) + e(v)\right] \\ &= \mathbb{E}_F\left[\int_{w \in \mathcal{V}} \mathbb{E}_z\left[\kappa(v, w) z(w)\right] \mathrm{d}F(w)\right] \\ &= 0 \end{aligned} \tag{4.19}$$

Now consider $\forall s \in \mathcal{V}$ and an arbitrary $s' \in \mathcal{V}$, such that $s \neq s'$,

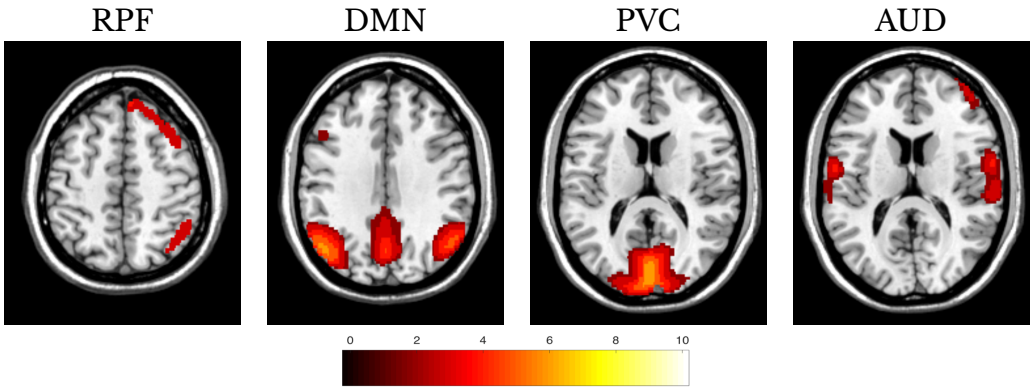
$$\begin{aligned} & \mathrm{Cov}[s(v), s(v')] \\ = \; & \mathbb{E}\left[s(v) s(v')\right] \\ = \; & \mathbb{E}\left[\int_{w \in \mathcal{V}} \kappa(v, w) u(\mathrm{d}w) \int_{w \in \mathcal{V}} \kappa(v', w) u(\mathrm{d}w)\right] \\ = \; & \mathbb{E}_F\left[\iint_{(w, w') \in \mathcal{V} \times \mathcal{V}} \mathbb{E}_z\left[\kappa(v, w) z(w) \kappa(v', w') z(w')\right] \mathrm{d}F(w)\, \mathrm{d}F(w')\right] \\ = \; & \mathbb{E}_F\left[\iint_{(w, w') \in \mathcal{V} \times \mathcal{V}} \kappa(v, w) \kappa(v', w') c(w, w') \, \mathrm{d}F(w)\, \mathrm{d}F(w')\right]. \end{aligned} \tag{4.20, 4.21}$$



(A) Results from our proposed method



(B) Results from the Infomax ICA



(C) Results from the empirical characteristic function based ICA

Figure 4.8: ICA results for the PNC data example from three representative ICA methods (RPF: right parietal frontal network; DMN: default mode network; PVC: primary visual cortex; AUD; auditory netork)

The stick-breaking construction of the DP gives $F = \sum_{h=1}^{\infty} [V_h \prod_{l<h}(1-V_l)]\, \delta_{w_h}$ with $w_1, \ldots, w_h, \ldots \overset{\text{iid}}{\sim} F_0$ and $V_1, \ldots, V_h, \ldots \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, which implies that:

$$
\begin{aligned}
& \text{R.H.S. of (4.20)} && (4.22)\\
= \;& \mathbb{E}_{(V,w)} \left[\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left(\kappa(v, w_i) V_i \prod_{l<i} (1-V_l) \right) \left(\kappa(v', w_j) V_j \prod_{l<j} (1-V_l) \right) c(w_i, w_j) \right] \\
= \;& \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{\alpha^{i-1}}{(1+\alpha)^i} \frac{\alpha^{j-1}}{(1+\alpha)^j} \iint_{(w,w') \in \mathcal{V} \times \mathcal{V}} \frac{1}{|\mathcal{V}|^2} \kappa(v, w) \kappa(v', w') c(w, w')\, dw\, dw' \\
= \;& \frac{1}{|\mathcal{V}|^2} \iint_{(w,w') \in \mathcal{V} \times \mathcal{V}} \kappa(v, w) c(w, w') \kappa(w', v')\, dw\, dw'. && (4.23)
\end{aligned}
$$

□

4.4.2 Details about the algorithm to draw from the posterior

We use a Markov chain Monte Carlo (MCMC) algorithm to approximate the posterior distribution drawing samples from the posterior distributions of the parameters in the ICA model.

Let $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$, $\Phi_i = \text{diag}(\phi_{1,i}, \ldots, \phi_{q,i})$, $B = [\beta_1, \ldots, \beta_q]^\top$ (a $q \times K$ matrix) and $\psi^{(i)} = [\psi_{1,i}, \ldots, \psi_{K,i}]^\top$ (the ith row of $\tilde{\Psi}$). The joint posterior is proportional to

$$
\begin{aligned}
& \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(A^\top y(v_i) - B\psi^{(i)} \right)^\top \Sigma^{-1} \Phi_i \left(A^\top y(v_i) - B\psi^{(i)} \right) \right\} \exp\left\{ -\frac{1}{2} \sum_{\ell=1}^{q} g_\ell^{-1} \beta_\ell^\top \beta_\ell \right\} \\
& \times \prod_{i=1}^{n} \left\{ \prod_{\ell=1}^{q} \phi_{\ell,i}^{\frac{1}{2}} \left(\sigma_\ell^2\right)^{-\frac{1}{2}} \right\} \left[\prod_{\ell=1}^{q} g_\ell^{-\frac{K}{2}} \right] \left[\prod_{\ell=1}^{q} \left(\sigma_\ell^2\right)^{-a_\sigma - 1} \exp\left(-\frac{b_\sigma}{\sigma_\ell^2} \right) \right] \\
& \times \left[\prod_{i=1}^{n} \prod_{\ell=1}^{q} \frac{\phi_{\ell,i}^{\nu_\ell/2-1} \exp(-\phi_{i,\ell} \nu_\ell / 2)}{\Gamma(\frac{\nu_\ell}{2}) \left(\frac{2}{\nu_\ell} \right)^{\nu_\ell/2}} \right] \left[\prod_{\ell=1}^{q} \frac{\xi_\ell^{\eta_\ell}}{\Gamma(\eta_\ell)} g_\ell^{-\eta_\ell - 1} \exp\left(-\frac{\xi_\ell}{g_\ell} \right) \right] \\
& \times \pi_a(A) \left[\prod_{\ell=1}^{q} \pi(\nu_\ell) \pi(\eta_\ell) \pi(\xi_\ell) \right]
\end{aligned}
$$

- Update A: the full conditional of A given all other parameters in this model is proportional to

$$
\exp\left\{ \text{tr}\left[\left(\sum_{i=1}^{n} \Sigma^{-1} \Phi_i B \psi^{(i)} y(v_i)^\top \right) A \right] \right\}, \quad A \in \mathbb{O}(q)
$$

This is a matrix von-Mises Fisher distribution. We sample from this distribution following the rejection sampling algorithm by Hoff (2009):

1. Obtain the singular value decomposition $\mathbf{UDV}^\top$ of

$$\sum_{i=1}^{n} \mathbf{y}(\boldsymbol{\nu}_i)\boldsymbol{\psi}^{(i)\top}\mathbf{B}^\top\boldsymbol{\Phi}_i\boldsymbol{\Sigma}^{-1}$$

   and let $\mathbf{H} = \mathbf{UD}$ with columns $[\mathbf{h}_1, \ldots, \mathbf{h}_q]$.

2. Sample pairs $\{u, \tilde{\mathbf{A}}\}$ until

$$u < \prod_{\ell=2}^{q-1} \frac{I_{(q-\ell-1)/2}(\|\mathcal{N}_\ell^\top \mathbf{h}_\ell\|)}{I_{(q-\ell-1)/2}(\|\mathbf{h}_\ell\|)} \left(\frac{\|\mathbf{h}_\ell\|}{\|\mathcal{N}_\ell^\top \mathbf{h}_\ell\|}\right)^{(q-\ell-1)/2},$$

   where $\mathcal{N}_\ell \in \mathbb{R}^{q\times(q-\ell+1)}$ contains a complete set of orthogonal basis for the null space of $[\tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_{\ell-1}]$; $I_\alpha(\cdot)$ is the modified Bessel function of the first kind. We complete this step by:

   i. Sample $u \sim \text{Unif}(0, 1)$.

   ii. a. Sample $\tilde{\mathbf{A}}_1 \sim \text{vMF}(\mathbf{h}_1)$, the vector von-Mises Fisher distribution.

   b. For $\ell = 2, \ldots, (q-1)$, sample $\mathbf{b} \sim \text{vMF}(\mathcal{N}_\ell^\top \mathbf{h}_\ell)$ and set $\tilde{\mathbf{A}}_\ell = \mathcal{N}_\ell \mathbf{b}$.

   c. Set $\tilde{\mathbf{A}}_q = \mathcal{N}_{q-1}$.

3. Set $\mathbf{A} = \tilde{\mathbf{A}}\mathbf{V}^\top$.

- Update $\{\boldsymbol{\beta}_\ell\}_{\ell=1}^q$: we draw $\boldsymbol{\beta}_\ell$ from $N(\boldsymbol{\mu}_{\boldsymbol{\beta}_\ell}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}_\ell})$ where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}_\ell} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}_\ell}\left[\sigma_\ell^{-2}\sum_{i=1}^{n}\phi_{\ell,i}\mathbf{A}_\ell^\top \mathbf{y}(\boldsymbol{\nu}_i)\boldsymbol{\psi}^{(i)}\right]$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}_\ell} = \left(\sum_{i=1}^{n}\sigma_\ell^{-2}\phi_{\ell,i}\boldsymbol{\psi}^{(i)}\boldsymbol{\psi}^{(i)\top} + g_\ell^{-1}\mathbf{I}_K\right)^{-1}.$$

- Update $\{\sigma_\ell^2\}_{\ell=1}^q$ by drawing from

$$\text{Inv-Ga}\left(a_\sigma + \frac{n}{2},\ b_\sigma + \frac{1}{2}\sum_{i=1}^{n}\phi_{\ell,i}\left(\boldsymbol{A}_\ell^\top \boldsymbol{y}(\boldsymbol{\nu}_i) - \boldsymbol{\beta}_\ell^\top \boldsymbol{\psi}^{(i)}\right)^2\right).$$

- Update $\{\phi_{\ell,i}\}_{1\leqslant \ell\leqslant q, 1\leqslant i\leqslant n}$ by drawing from

$$\text{Gamma}\left(\frac{1+\nu_\ell}{2},\ \frac{2}{\sigma_\ell^{-2}\left(\boldsymbol{A}_\ell^\top \boldsymbol{y}(\boldsymbol{\nu}_i) - \boldsymbol{\beta}_\ell^\top \boldsymbol{\psi}^{(i)}\right)^2 + \nu_\ell}\right).$$

- Update $\{g_\ell\}_{\ell=1}^q$ by drawing g_ℓ from

$$\text{Inv-Ga}\left(\frac{2\eta_\ell + K}{2},\ \frac{2\xi_\ell + \boldsymbol{\beta}_\ell^\top \boldsymbol{\beta}_\ell}{2}\right)$$

  for all $\ell = 1, \ldots, q$.

- Update $\{\eta_\ell, \xi_\ell\}_{\ell=1}^q$: we update η_ℓ and ξ_ℓ together via a Metropolis updating scheme embedded in the Gibbs sampler. We first generate a symmetric proposal $[\eta_\ell^*, \xi_\ell^*]^\top \sim N([\eta_\ell, \xi_\ell]^\top, {}^2\mathbf{I}_2)$ and then accept it with probability

$$\min\left\{1,\quad \frac{i\gamma(g_\ell; \eta_\ell^*, \xi_\ell^*)}{i\gamma(g_\ell; \eta_\ell, \xi_\ell)} \exp\left(\frac{b_\eta}{\eta_\ell} - \frac{b_\eta}{\eta_\ell^*} + \frac{b_\xi}{\xi_\ell} - \frac{b_\xi}{\xi_\ell^*}\right)\left(\frac{\eta_\ell}{\eta_\ell^*}\right)^{a_\eta+1}\left(\frac{\xi_\ell}{\xi_\ell^*}\right)^{a_\xi+1}\right\},$$

  where $i\gamma(\cdot; \alpha, \beta)$ is the density function of an inverse Gamma distribution with shape parameter α and scale parameter β.

- Update $\{\nu_\ell\}_{\ell=1}^q$: this updating step is also achieved with Metropolis algorithms. We first sample a proposal $\nu_\ell^* \sim N(\nu_\ell, {}^2_\nu)$ and then accept it with probability

$$\min\left\{1,\quad \left[\prod_{i=1}^{n}\frac{\gamma(\phi_{\ell,i}; \nu_\ell^*/2, 2/\nu_\ell^*)}{\gamma(\phi_{\ell,i}; \nu_\ell/2, 2/\nu_\ell)}\right]\exp\left(\frac{b_\nu}{\nu_\ell} - \frac{b_\nu}{\nu_\ell^*}\right)\left(\frac{\nu_\ell}{\nu_\ell^*}\right)^{a_\nu+1}\right\},$$

  where $\gamma(\cdot; a, b)$ is the density function of a Gamma distribution with shape parameter a and scale parameter b.

4.4.3 Functional eigen-decomposition for the generalized SE kernel

For a modified SE kernel function: $\kappa^{mod}(\boldsymbol{v},\boldsymbol{w}) = \exp\left\{-a\|\boldsymbol{v}\|_2^2 - a\|\boldsymbol{w}\|_2^2 - \frac{\|\boldsymbol{v}-\boldsymbol{w}\|_2^2}{\rho}\right\}$, $a, \rho > 0$, let $b = \rho^{-1}$. Then if $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^d$, its functional eigen-decomposition has an analytical form given in the following proposition:

Proposition 2. *For a specific* $k \in \{1, ..., \infty\}$*, define series* $\{k_i\}_{i=0}^d$*,* $\{l_i\}_{i=0}^d$ *and* $\{m_i\}_{i=1}^d$ *as follows*

$$k_i = \left\{k_i \in \mathbb{N}^0 : \binom{k_i + d - i - 1}{d-i} \leqslant l_i \leqslant \binom{k_i + d - i}{d-i} - 1\right\},\ 0 \leqslant i \leqslant d-1,\ k_d = 0,$$

$$l_0 = l - 1,\ l_i = l_{i-1} - \binom{k_{i-1} + d - i}{d - i + 1},\ i \geqslant 1,$$

$$m_i = k_{i-1} - k_i,\ i \geqslant 1,$$

where $\mathbb{N}^0$ *is the set of nonnegative integers;* $\binom{n}{k} = 0$ *if* $k > n$*. Define* $K = \binom{m+d}{d} = \sum_{k=0}^{m}\binom{k+d-1}{d-1}$*. For* $\boldsymbol{v} = [v_1, ..., v_d]^\top \in \mathbb{R}^d$*, let* $\psi_k(\boldsymbol{v})$ *and* λ_k *be the* k*th (sort* λ_k*s from large to small) combination of eigenfunctions and eigenvalues for the modified square exponential kernel* $\kappa(\boldsymbol{v},\boldsymbol{w})$*, then*

$$\lambda_k = \left(\frac{\pi}{A}\right)^d B^{k_0}, \quad \frac{\sum_{k=1}^{K}\lambda_k}{\sum_{k=1}^{\infty}\lambda_l} = (1-B)^d \sum_{k=0}^{m}\binom{k+d-1}{d-1}B^k,$$

$$\psi_k(\boldsymbol{v}) = (2c)^{\frac{d}{4}} \exp\left(-c\|\boldsymbol{v}\|_2^2\right)\prod_{i=1}^{d} H_{m_i}(\sqrt{2c}v_i),$$

where $c = \sqrt{a^2 + 2ab}$*,* $A = a + b + c$ *and* $B = b/A$*;* $H_j(\cdot)$ *is the* j*th (*$j \in \mathbb{N}^0$*) order normalized Hermite polynomial.*

This result is a direct extension from the one-dimensional case ($d = 1$) derived in Zhu et al. (1998).

Chapter 5

Summary and Future Work

We provide concluding remarks on our contributions to the existing literature and briefly discuss some future directions in this chapter.

The first topic of this dissertation contributes to the existing literature mainly in four aspects. First, our proposed model is the first hierarchical ICA model for fMRI data analysis which incorporates covariate effects into the source signals. Our model provides the first statistical framework to estimate and test the differences between brain functional networks using fMRI data. Second, we propose a fast approximate EM algorithm for model estimation, which scales linearly to the number of source signals. This fast computation algorithm is extremely suitable for analyzing big imaging data. Third, our model can naturally control for potential confounding factors to the primary covariate effects of interest, which is suitable for the PTSD dataset given its observational nature. Fourth, empirical simulation studies confirm that our method can address the “cross-talk" issues in the existing method and provide high statistical power in terms of testing differences between brain functional networks. A future direction along this line is to conduct shinkage estimation of covarate effects on brain functional networks.

The contribution of the second topic can be summarized as three main points. First, we propose a new functional class with rigorous math-

ematical definitions for the covariate functions in spatially varying coefficient models, which can capture common features in neuroimaging data: region-wise spatial smoothness, jump discontinuity and sparsity. Second, we propose a new family of priors, namely the thresholded Gaussian process priors, for the covariate functions to facilitate nonparametric Bayesian inference. The proposed priors have large support property and induce posterior consistency, as well as posterior regional selection consistency, under the SVCM framework. Third, empirical performance of our method is superior to existing GLM based methods: It has lower Type I error rates compared with the false discovery rate control procedure; it is statistically more powerful compared with the family-wise error rate control method based on random field theory. One potential improvement to this topic is to drop the constant thresholding paradigm and conduct adaptive SVCF support estimation using dynmanic thresholding processes (from λ to $\lambda(\mathbf{s})$).

The third topic presented in this dissertation lead to three major contributions. First, it is the first ICA model that can account for spatial dependence within the spatial source signals, or the ICs. The nonparametric regression modeling paradigm to the ICs extends the existing literature which mainly focuses on using density estimation to handle the ICs by treating them as white noises. It separates the informative spatial mean dynamics from residual noises. The adoption of Bayesian kernel models enable us to capture various sophisticated spatial dependence structure flexibly. Second, we show through extensive simulation studies that our method can estimate the mixing matrix and identify spatial activation patterns more accurately, especially when the signal-to-noise ratio is low in the data. The advantages over existing ICA methods are mainly due to the incorporation of spatial dependence as well as the separation of spatial mean trends from residual noises. Third, our method is the first one that can make formal model-based inference about the spatial source signals when performing ICA decompositions of fMRI data. Our fully

Bayesian inference algorithm based on MCMC gives us a natural way of constructing posterior credible sets for the spatial source signals at each brain voxel. Testing the significance of activation is straightforward as we only need to determine whether the credible intervals include zeros. Since the Bayesian kernel models can be defined on general Euclidean spaces, one direct, yet useful, extension to this topic is to bring in covariate effects into the kernel function and use anisotropic scale parameters to select and quantify these effects.

Bibliography

Adler, R. J. and Taylor, J. E. (2009), *Random Fields and Geometry*, vol. 115, Springer.

Amari, S.-i., Cichocki, A., Yang, H. H., et al. (1996), "A new learning algorithm for blind signal separation," *Advances in neural information processing systems*, 757–763.

Anand, A., Li, Y., Wang, Y., Wu, J., Gao, S., Bukhari, L., Mathews, V. P., Kalnin, A., and Lowe, M. J. (2005), "Activity and connectivity of brain mood regulating circuit in depression: a functional magnetic resonance study," *Biological psychiatry*, 57, 1079–1088.

Attias, H. (1999), "Independent factor analysis," *Neural computation*, 11, 803–851.

— (2000), "A variational Bayesian framework for graphical models," *Advances in neural information processing systems*, 12, 209–215.

Bach, F. R. and Jordan, M. I. (2003), "Kernel independent component analysis," *The Journal of Machine Learning Research*, 3, 1–48.

Beck, A. T., Steer, R. A., Brown, G. K., et al. (1996), "Manual for the beck depression inventory-II," .

Beck, A. T., Steer, R. A., and Carbin, M. G. (1988), "Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation," *Clinical psychology review*, 8, 77–100.

Beckmann, C. F., DeLuca, M., Devlin, J. T., and Smith, S. M. (2005), "Investigations into resting-state connectivity using independent compo-

nent analysis," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360, 1001–1013.

Beckmann, C. F., Mackay, C. E., Filippini, N., and Smith, S. M. (2009), "Group comparison of resting-state FMRI data using multi-subject ICA and dual regression," *Neuroimage*, 47, S148.

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

— (2005), "Tensorial extensions of independent component analysis for multisubject FMRI analysis," *Neuroimage*, 25, 294–311.

Bell, A. J. and Sejnowski, T. J. (1995), "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, 7, 1129–1159.

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

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2014), "Dirichlet-Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, 0, 00–00.

Biswal, B. B. and Ulmer, J. L. (1999), "Blind source separation of multiple signal sources of fMRI data sets using independent component analysis," *Journal of computer assisted tomography*, 23, 265–271.

Bondell, H. D. and Reich, B. J. (2012), "Consistent high-dimensional Bayesian variable selection via penalized credible regions," *Journal of the American Statistical Association*, 107, 1610–1624.

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

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. (1996), "Statistical methods of estimation and inference for functional MR image analysis," *Magnetic Resonance in Medicine*, 35, 261–277.

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

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001), "A method for making group inferences from functional MRI data using independent component analysis," *Human brain mapping*, 14, 140–151.

Campbell, D. G., Felker, B. L., Liu, C.-F., Yano, E. M., Kirchner, J. E., Chan, D., Rubenstein, L. V., and Chaney, E. F. (2007), "Prevalence of depression–PTSD comorbidity: Implications for clinical practice guidelines and primary care-based interventions," *Journal of General Internal Medicine*, 22, 711–718.

Cardoso, J.-F. (1999), "High-order contrasts for independent component analysis," *Neural computation*, 11, 157–192.

Chen, A. (2006), "Fast kernel density independent component analysis," in *Independent Component Analysis and Blind Signal Separation*, Springer, pp. 24–31.

Chen, A. and Bickel, P. J. (2005), "Consistent independent component analysis and prewhitening," *Signal Processing, IEEE Transactions on*, 53, 3625–3632.

— (2006), "Efficient independent component analysis," *The Annals of Statistics*, 34, 2825–2855.

Chen, C.-H., Ridler, K., Suckling, J., Williams, S., Fu, C. H., Merlo-Pich, E., and Bullmore, E. (2007), "Brain imaging correlates of depressive symptom severity and predictors of symptom improvement after antidepressant treatment," *Biological psychiatry*, 62, 407–414.

Choudhuri, N., Ghosal, S., and Roy, A. (2004), "Bayesian estimation of the spectral density of a time series," *Journal of the American Statistical Association*, 99, 1050–1059.

Chumbley, J., Worsley, K. J., Flandin, G., and Friston, K. J. (2009), "False Discovery Rate Revisited: FDR and Topological Inference Using Gaussian Random Fields," *NeuroImage*, 44, 62–70.

Cole, L. J., Farrell, M. J., Gibson, S. J., and Egan, G. F. (2010), "Age-related differences in pain sensitivity and regional brain activity evoked by noxious pressure," *Neurobiology of aging*, 31, 494–503.

Cressie, N. A. and Cassie, N. A. (1993), *Statistics for Spatial Data*, vol. 900, Wiley New York.

Daniels, J. K., Frewen, P., McKinnon, M. C., and Lanius, R. A. (2011), "Default mode alterations in posttraumatic stress disorder related to early-life trauma: a developmental perspective," *Journal of psychiatry & neuroscience: JPN*, 36, 56.

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'ardenne, K., Richter, W., Cohen, J., and Haxby, J. (2009), "Independent component analysis for brain fMRI does not select for independence," *Proceedings of the National Academy of Sciences*, 106, 10415–10422.

Diggle, P. J., Tawn, J., and Moyeed, R. (1998), "Model-based geostatistics," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47, 299–350.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., Matthews, P. M., Beckmann, C. F., and Mackay, C. E. (2009), "Distinct patterns of brain activity in young carriers of the APOE-ε4 allele," *Proceedings of the National Academy of Sciences*, 106, 7209–7214.

First, M. B. (1995), "Structured Clinical Interview for the DSM (SCID)," *The Encyclopedia of Clinical Psychology*.

Flandin, G. and Penny, W. D. (2007), "Bayesian fMRI data analysis with sparse spatial basis function priors," *NeuroImage*, 34, 1108–1125.

Fonseca, T. C., Ferreira, M. A., and Migon, H. S. (2008), "Objective Bayesian analysis for the Student-t regression model," *Biometrika*.

Fox, E. and Dunson, D. B. (2012), "Multiresolution Gaussian Processes," in *Advances in Neural Information Processing Systems 25*, eds. Pereira, F., Burges, C., Bottou, L., and Weinberger, K., Curran Associates, Inc., pp. 737–745.

Friston, K. and Penny, W. (2003), "Posterior probability maps and SPMs," *Neuroimage*, 19, 1240–1249.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. (1995), "Statistical parametric maps in functional imaging: a general linear approach," *Human brain mapping*, 2, 189–210.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003), "Spatial modeling with spatially varying coefficient processes," *Journal of the American Statistical Association*, 98, 387–396.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002), "Thresholding of statistical maps in functional neuroimaging using the false discovery rate," *Neuroimage*, 15, 870–878.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Ghosal, S., Roy, A., et al. (2006), "Posterior consistency of Gaussian process prior for nonparametric binary regression," *The Annals of Statistics*, 34, 2413–2429.

Gramacy, R. B. and Lee, H. K. H. (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association*, 103, 1119–1130.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., Reiss, A. L., and Schatzberg, A. F. (2007), "Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus," *Biological psychiatry*, 62, 429–437.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014), "ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging," *Neuroimage*, 95, 232–247.

Guo, Y. (2011), "A general probabilistic model for group independent component analysis and its estimation methods," *Biometrics*, 67, 1532–1542.

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

Guo, Y. and Tang, L. (2013), "A hierarchical model for probabilistic independent component analysis of multi-subject fMRI studies," *Biometrics*, 69, 970–981.

Gupta, A. and Nagar, D. (2000), *Matrix Variate Distributions*, Monographs and Surveys in Pure and Applied Mathematics. Boca Raton, Fla, USA: Chapman and Hall/CRC.

Hadjikhani, N., Joseph, R. M., Snyder, J., Chabris, C. F., Clark, J., Steele, S., McGrath, L., Vangel, M., Aharon, I., Feczko, E., et al. (2004), "Activation of the fusiform gyrus when individuals with autism spectrum disorder view faces," *NeuroImage*, 22, 1141–1150.

Hallin, M. and Mehta, C. (2015), "R-Estimation for Asymmetric Indepen-

dent Component Analysis," *Journal of the American Statistical Association*, 110, 218–232.

Hendler, T., Rotshtein, P., Yeshurun, Y., Weizmann, T., Kahn, I., Ben-Bashat, D., Malach, R., and Bleich, A. (2003), "Sensing the invisible: differential sensitivity of visual cortex and amygdala to traumatic context," *Neuroimage*, 19, 587–600.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004), "Validating the independent components of neuroimaging time series via clustering and visualization," *Neuroimage*, 22, 1214–1222.

Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, 58, 13–30.

Hoff, P. D. (2009), "Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data," *Journal of Computational and Graphical Statistics*, 18.

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

Hyvärinen, A., Karhunen, J., and Oja, E. (2001), *Independent component analysis*, vol. 46, John Wiley & Sons.

Hyvärinen, A. and Oja, E. (2000), "Independent component analysis: algorithms and applications," *Neural networks*, 13, 411–430.

Ilmonen, P., Paindaveine, D., et al. (2011), "Semiparametrically efficient inference based on signed ranks in symmetric independent component models," *the Annals of Statistics*, 39, 2448–2476.

Johnson, V. E. and Rossell, D. (2012), "Bayesian model selection in high-dimensional settings," *Journal of the American Statistical Association*, 107, 649–660.

Katanoda, K., Matsuda, Y., and Sugishita, M. (2002), "A spatio-temporal

regression model for the analysis of functional MRI data," *NeuroImage*, 17, 1415–1428.

Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., and Nelson, C. B. (1995), "Posttraumatic stress disorder in the National Comorbidity Survey," *Archives of general psychiatry*, 52, 1048–1060.

Kim, H.-M., Mallick, B. K., and Holmes, C. (2005), "Analyzing nonstationary spatial data using piecewise Gaussian processes," *Journal of the American Statistical Association*, 100, 653–668.

Kostantinos, N. (2000), "Gaussian mixtures and their applications to signal processing," *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*.

Lange, K. L., Little, R. J., and Taylor, J. M. (1989), "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, 84, 881–896.

Lee, S., Shen, H., Truong, Y., Lewis, M., and Huang, X. (2011), "Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging," *Journal of the American Statistical Association*, 106, 1009–1024.

Li, F. and Zhang, N. R. (2010), "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics," *Journal of the American Statistical Association*, 105.

Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. (2011), "Multiscale adaptive regression models for neuroimaging data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 559–578.

Li, Y.-O., Adali, T., and Calhoun, V. D. (2006), "Sample dependence correction for order selection in fMRI analysis," in *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, IEEE, pp. 1072–1075.

Liang, F., Mao, K., Liao, M., Mukherjee, S., and West, M. (2006), "Non-parametric Bayesian kernel models," *Department of Statistical Science, Duke University, Discussion Paper*.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g priors for Bayesian variable selection," *Journal of the American Statistical Association*, 103.

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society, Series B*, 44, 226–233.

Maruyama, Y. and George, E. I. (2011), "Fully Bayes factors with a generalized g-prior," *The Annals of Statistics*, 2740–2765.

Mckeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Kindermann, R. S., Bell, A. J., and Sejnowski, T. J. (1998), "Analysis of fMRI Data by Blind Separation Into Independent Spatial Components," *Human Brain Mapping*, 6, 160–188.

McLachlan, G. and Peel, D. (2004), *Finite mixture models*, John Wiley & Sons.

Meilijson, I. (1989), "A fast improvement to the EM algorithm on its own terms," *Journal of the Royal Statistical Society. Series B. Methodological*, 51, 127–138.

Meng, X.-L. and Rubin, D. B. (1991), "Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Minka, T. P. (2000), "Automatic choice of dimensionality for PCA," in *NIPS*, vol. 13, pp. 598–604.

Mitchell, T. J. and Beauchamp, J. J. (1988), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Muirhead, R. J. (2005), *Aspects of multivariate statistical theory*, Hoboken, NJ: John Wiley & Sons.

Narisetty, N. N., He, X., et al. (2014), "Bayesian variable selection with shrinking and diffusing priors," *The Annals of Statistics*, 42, 789–817.

Nichols, T. and Hayasaka, S. (2003), "Controlling the familywise error rate in functional neuroimaging: a comparative review," *Statistical Methods in Medical Research*, 12, 419–446.

Park, T. and Casella, G. (2008), "The bayesian lasso," *Journal of the American Statistical Association*, 103, 681–686.

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005), "Bayesian fMRI time series analysis with spatial priors," *NeuroImage*, 24, 350–362.

Pillai, N. S., Wu, Q., Liang, F., Mukherjee, S., and Wolpert, R. L. (2007), "Characterizing the Function Space for Bayesian Kernel Models." *Journal of Machine Learning Research*, 8.

Polzehl, J. and Spokoiny, V. G. (2000), "Adaptive weights smoothing with applications to image restoration," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 335–354.

Qiu, P. (2007), "Jump surface estimation, edge detection, and image restoration," *Journal of the American Statistical Association*, 102, 745–756.

Quiton, R. L. and Greenspan, J. D. (2007), "Sex differences in endogenous pain modulation by distracting and painful conditioning stimulation," *Pain*, 132, S134–S149.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001), "A default mode of brain function," *Proceedings of the National Academy of Sciences*, 98, 676–682.

Rasmussen, C. E. and Williams, C. K. (2006a), "Gaussian processes for machine learning," *the MIT Press*, 2, 4.

Rasmussen, C. E. and Williams, C. K. I. (2006b), *Gaussian processes for machine learning*, MIT Press.

Reich, B. J., Fuentes, M., Herring, A. H., and Evenson, K. R. (2010), "Bayesian variable selection for multivariate spatially varying coefficient regression," *Biometrics*, 66, 772–782.

Reineberg, A. E., Andrews-Hanna, J. R., Depue, B. E., Friedman, N. P., and Banich, M. T. (2015), "Resting-state networks predict individual differences in common and specific aspects of executive function," *NeuroImage*, 104, 69–78.

Roberts, G. O. and Tweedie, R. L. (1996), "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, 341–363.

Samarov, A. and Tsybakov, A. (2004), "Nonparametric independent component analysis," *Bernoulli*, 10, 565–582.

Samworth, R. J. and Yuan, M. (2012), "Independent component analysis via nonparametric maximum likelihood estimation," *The Annals of Statistics*, 40, 2973–3002.

Satterthwaite, T., Wolf, D., Roalf, D., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E., Elliott, M., Smith, A., Hakonarson, H., et al. (2015), "Linked Sex Differences in Cognition and Functional Connectivity in Youth." *Cerebral cortex (New York, NY: 1991)*, 25, 2383.

Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., et al. (2014), "Neuroimaging of the Philadelphia neurodevelopmental cohort," *Neuroimage*, 86, 544–553.

Seber, G. A. and Lee, A. J. (2012), *Linear regression analysis*, vol. 936, John Wiley & Sons.

Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., Mintun, M. A., Wang, S., Coalson, R. S., and Raichle, M. E. (2009), "The default mode network and self-referential processes in depression," *Proceedings of the National Academy of Sciences*, 106, 1942–1947.

Smith, D. V., Utevsky, A. V., Bland, A. R., Clement, N., Clithero, J. A., Harsch, A. E., Carter, R. M., and Huettel, S. A. (2014), "Characterizing individual differences in functional connectivity using dual-regression and seed-based approaches," *NeuroImage*, 95, 1–12.

Smith, M. and Fahrmeir, L. (2007), "Spatial Bayesian variable selection with application to functional magnetic resonance imaging," *Journal of the American Statistical Association*, 102, 417–431.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009), "Correspondence of the brain's functional architecture during activation and rest," *Proceedings of the National Academy of Sciences*, 106, 13040–13045.

Tabelow, K., Polzehl, J., Ulug, A. M., Dyke, J. P., Watts, R., Heier, L. A., and Voss, H. U. (2008), "Accurate localization of brain activity in presurgical fmri by structure adaptive smoothing," *Medical Imaging, IEEE Transactions on*, 27, 531–537.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R. and Hastie, T. J. (2002), "Independent components analysis through product density estimation," in *Advances in neural information processing systems*, pp. 649–656.

Tohka, J., Foerde, K., Aron, A. R., Tom, S. M., Toga, A. W., and Poldrack, R. A. (2008), "Automatic independent component labeling for artifact removal in fMRI," *Neuroimage*, 39, 1227–1245.

Tokdar, S. T. and Ghosh, J. K. (2007), "Posterior consistency of logistic Gaussian process priors in density estimation," *Journal of Statistical Planning and Inference*, 137, 34–42.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002), "Automated anatomical labeling of activations in SPM using a macroscopic anatom-

ical parcellation of the MNI MRI single-subject brain," *NeuroImage*, 15, 273–289.

Van Der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Wang, J., Zhu, H., Fan, J., Giovanello, K., and Lin, W. (2013), "Multiscale adaptive smoothing models for the hemodynamic response function in fMRI," *The Annals of Applied Statistics*, 7, 904.

West, M. (1984), "Outlier models and prior distributions in Bayesian linear regression," *Journal of the Royal Statistical Society. Series B (Methodological)*, 431–439.

— (2003), "Bayesian factor regression models in the âĂIJlarge p, small nâĂİ paradigm," *Bayesian statistics*, 7, 733–742.

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., Shenton, M. E., Green, A. I., Nieto-Castanon, A., LaViolette, P., et al. (2009), "Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia," *Proceedings of the National Academy of Sciences*, 106, 1279–1284.

Wolpert, R. L., Clyde, M. A., and Tu, C. (2011), "Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels," *The Annals of Statistics*, 1916–1962.

Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith, S. M. (2004), "Fully Bayesian spatio-temporal modeling of fMRI data," *Medical Imaging, IEEE Transactions on*, 23, 213–231.

Xu, L., Cheung, C., Yang, H., and Amari, S. (1997), "Maximum equalization by entropy maximization and mixture of cumulative distribution functions," in *Proc. of ICNNŠ97*, pp. 1821–1826.

Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 49–67.

Yue, Y., Loh, J. M., and Lindquist, M. A. (2010), "Adaptive spatial smoothing of fMRI images," *Statistics and its Interface*, 3, 3–13.

Zellner, A. (1986), "Bayesian estimation and prediction using asymmetric loss functions," *Journal of the American Statistical Association*, 81, 446–451.

Zhu, H., Fan, J., and Kong, L. (2014), "Spatially Varying Coefficient Model for Neuroimaging Data with Jump Discontinuities," *Journal of the American Statistical Association*.

Zhu, H., Williams, C. K., Rohwer, R., and Morciniec, M. (1998), "Gaussian Regression and Optimal Finite Dimensional Linear Models," in *Neural Networks and Machine Learning*, ed. Bishop, C. M., Berlin: Springer-Verlag.

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

Zou, Q.-H., Zhu, C.-Z., Yang, Y., Zuo, X.-N., Long, X.-Y., Cao, Q.-J., Wang, Y.-F., and Zang, Y.-F. (2008), "An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF," *Journal of Neuroscience Methods*, 172, 137–141.