Spatial Bayesian Modeling and Computation with Application to Neuroimaging Data

by

Cui Guo

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Professor Timothy D. Johnson, Co-Chair
Associate Professor Jian Kang, Co-Chair
Associate Professor Veronica Berrocal
Professor Stephan Taylor

Cui Guo

cuiguo@umich.edu

ORCID iD: 0000-0002-3297-119X

ACKNOWLEDGEMENTS

I would first like to express my sincere gratitude to my advisers Dr. Timothy D. Johnson and Dr. Jian Kang for their outstanding guidance and continuous support of my Ph.D. study and research. They have introduced me into the neuroimaging research and Bayesian analysis. I am fully motivated by their broad knowledge, deep insights and great enthusiasm in all the time of research. I also have learned a lot of useful computational techniques as well as communication, academic writing and presentation skills from them.

I wish to thank Dr. Veronica Berroca and Dr. Stephan Taylor for serving as members of my dissertation committee and providing me very insightful comments on my dissertation. Especially, I am very grateful to Dr. Veronica Berroca for teaching me spatial statistics and thoughtful discussions on the Bayesian spatial modeling, to Dr. Taylor for his useful suggestions on neuroimaging parcellations and emotion analysis.

My sincere thanks also goes to the Department of Biostatistics and the University of Michigan, where I have had very meaningful and enjoyable experience. I would like to thank all of the knowledgeable faculty members for their expertise in Biostatistics. In particular, I have very interesting teaching experience in Survival Analysis with Dr. Yi Li, and invaluable collaboration experience as a research assistance with Dr. Mousumi Banerjee and Dr. Thomas Braun. I am also grateful to all the staff members, fellow students and friends in the department for their great help and support.

Also, I wish to thank my parents for their love, patience and effort for providing me the best environment and education. Last but not least, I am greatly indebted to my boyfriend Peng Liao for his continuous support of my life and research in the past five years.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

As both clinical and cognitive neuroscience matures, the need for sophisticated neuroimaging analyses becomes more important. The use of imaging markers to predict clinical outcomes, or even imaging outcomes, can have great impact on public health. However, such analyses are still under development since it is challenging for several reasons: 1) the images are of high dimension, and 2) the images may exhibit complex spatial correlation structure. Bayesian methods play an important role in solving these problems by dealing with spatial data flexibly and applying efficient sampling algorithms. This dissertation aims to develop spatial Bayesian models to predict either scalar or imaging outcomes by using imaging predictors and seeks computationally efficient approaches.

In Chapter II, we propose a Bayesian scalar-on-image regression model with application to Multiple Sclerosis (MS) Magnetic Resonance Imaging (MRI) data. Specifically, we build up a multinomial logistic regression model to predict the clinical subtypes of MS patients by using their 3D MRI lesion data. Parameters corresponding to MRI predictors are spatially varying in the image space and are assumed to have a Gaussian Process (GP) prior distribution. Since the covariates are highly correlated, we use the Hamiltonian Monte Carlo algorithm, which is more statistically efficient than other Markov Chain Monte Carlo methods when the parameters are highly correlated. Finally, to reduce computational burden, we code the problem to run in parallel on a graphical processing unit. Results from simulation studies and a real MS data set show that our method has high prediction accuracy as evaluated by leave-one-out cross validation using an importance-sampling scheme.

In Chapter III, we propose a novel image-on-image regression model, by extending a spatial Bayesian latent factor model to neuroimaging data, where low dimensional latent factors are adopted to make connections between high-dimensional image outcomes and image predictors. We assign GP priors to the spatially varying regression coefficients in the model, which capture the complex spatial dependence among image outcomes as well as that among the image predictors. We perform simulation studies to evaluate the out-of-sample prediction performance of our method compared with linear regression and voxel-wise regression methods for different scenarios. We apply the proposed method to analysis of multimodal image data in the Human Connectome Project (HCP) where we predict task-related contrast maps using sub-cortical volumetric seed maps. The proposed method achieves a better prediction accuracy than simpler models by effectively accounting for the spatial dependence and efficient reduction of image dimension with latent factors.

In Chapter IV, we extend the image-on-image regression model proposed in Chapter III to the case where outcome is a cortical surface image and predictors images are volumetric seed maps. We expand the surface image on a set of spherical harmonics basis functions, where coefficients are linked to image predictors through a latent factor model. We assign GP priors to the spatially varying regression coefficients of the volumetric predictor images. Compared to ridge regression, the proposed method performs better in prediction according to simulation studies, and it can identify active brain regions in spherical z-score images from the HCP.

CHAPTER I

Introduction

The desire to understand the human brain has been one of scientific interest throughout the ages and our knowledge of the how the brain has exploded in recent years. Neuroimaging is an umbrella term encompassing a variety of medical imaging technologies used to non-invasively study the brain. These include various rapidly evolving techniques to image the brain properties related to structure, function, disease pathophysiology, or pharmacology of the nervous system. These techniques have found applications in a wide variety of fields, including neuroscience, cognitive science, psychology, neurology, and medicine. And the research of neuroimaging draws together a multi-disciplinary community of neuroscientists, psychologists, physicists, statisticians, mathematician, and computer scientists.

Statisticians have played a crucial role in this research and have contributed significantly to the field. However, the statistical analysis of neuroimaging data is still very challenging for a number of reasons. Currently, most classical statistical methods do not account for the spatial structure in the data resulting in underpowered statistical inferences (*Li et al.*, 2011; *Polzehl et al.*, 2010; *Wolfers et al.*, 2015; *Mwangi et al.*, 2014). There have been some attempts to model this complex and high-dimensional spatial correlation (*Bowman et al.*, 2008), however, there are still limitations due to high dimension, computational capability, and model flexibility (*Hyun et al.*, 2014;

Zhu et al., 2007). To help fill this gap, this dissertation proposes statistical approaches for the analysis of neuroimaging data by properly modelling the spatial dependence in images. The proposed methods are implemented using the Bayesian framework with fast computational algorithms and techniques. We apply our statistical models (1) to the associations of disease outcomes with neuroimaging data (Scalar-on-Image regression) and (2)to model task-related brain activity with resting state and structural imaging (so called Image-on-Image).

The remainder of this chapter is organized as follows: we start in Section 1.1 with a brief overview of neuroimaging including a variety of modalities. The statistical methods of neuroimaging data analysis are summarized in Section 1.2. The challenges of neuroimaging data analysis, in particular accounting for spatial dependence, are presented in Section 1.3. Finally, we outline the dissertation in Section 1.4.

1.1 Neuroimaging Data

The discovery of X-rays in 1895 by Professor W. C. Röntgen created a revolutionary step in the history of health sciences, which enabled physicians for the first time to noninvasively study the inside structure of a living subject. (*Röntgen*, 1896; *Glasser*, 1993; *Donya et al.*, 2015). X-rays were, and are still, widely used in hospitals to study damage to bones, teeth, and lungs but inherently limited to the study of anatomical structures because it projects a three-dimensional (3D) object onto a two-dimensional (2D) image (*Donya et al.*, 2015; *Ardran*, 1979). The limitations of X-rays encouraged the advent of modern advanced imaging modalities. And it is usually important to separate neuroimaging into structural imaging and functional imaging, each of which has several modalities.

1.1.1 Structural Imaging

Structural imaging, used for the study of brain structure and the diagnosis of disease and injury, typically including computerized tomography(CT), magnetic resonance imaging (MRI) and positron emission tomography (PET).

A CT scanner use computerized technology to combine many X-rays measurements taken from different angles to produce cross-sectional (tomographic) images (virtual "slices") of an object (*Epstein*, 2007; *Natterer*, 2001). In contrast to CT, PET performs emission tomography to observe metabolic processes in the body. A positron-emitting radio-ligand is introduced into the subject and distributed across the organ of interest. The PET scanner has a number of detector surrounding the subject. When the radioisotope decays, two photons are then detected by the photomultipliers, determining the line along which the decay happened (*Daghighian et al.*, 1990; *Shukla and Kumar*, 2006; *Nasrallah and Dubroff*, 2013). A 3D image of tracer concentration within the body is then reconstructed by a computer algorithm.

MRI provides more detailed anatomical information than CT without exposing the body to radiation. A MR scanner produces powerful magnetic field and radio frequency pulses to excite nuclei in the brain, which absorb external energy. Once the pulse is turned off, the nuclei emit this extra energy and return to their original aligned positions (a procedure called relaxation) (*Wills and Hector*, 1924; *Hashemi et al.*). The emitted energy is captured in turn in the scanner as the basic MR signal. A system of gradient coils vary the strength of the magnetic field, so that each location in the brain has its own resonance frequency as its raw data from the MR scanner. An inverse Fourier transform is applied to the raw data to reconstruct the image (*Heggie*, 2001; *Bracewell and Bracewell*, 1986).

MRI has high spatial resolution and is very adept at morphological imaging and functional imaging. PET permits the estimation of the density of a variety of neurochemical receptors across the brain (*Ombao et al.*, 2016) so that it is complementary

to MRI. CT is mostly used for bone injuries, chest or lung imaging and detecting cancer, while MRI is better suited for soft tissue damage. In addition, MRI is free of radiation, which is harmful in repeated exposures.

1.1.2 Functional Imaging

To study cognitive and affective processes, commonly used functional imaging modalities include PET, functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and magnetoencephalography (MEG).

EEG maps brain electrical activity with millisecond temporal resolution while MEG maps magnetic changes. EEG is non-invasive with a number of electrodes (usually 32, 64, or 256) placed along the scalp. It measures voltage fluctuations resulting from ionic current within the neurons of the brain. Similarly, MEG is also non-invasive and records magnetic fields produced by electrical currents occurring naturally in the brain, using very sensitive magnetometers.

fMRI extends the use of MRI to indirectly measure neuronal activity via the blood oxygenation level dependent (BOLD) contrast (*Ogawa et al.*, 1990, 1993). When the brain is active in response to a particular task, the neurons responsible for that task require more energy (oxygen) resulting in an increase in oxygenated blood flow. MR signals reflect the changes in the local magnetic fields caused by the changes of blood oxygenation in that region (*Kwong et al.*, 1992). During a standard fMRI experiment, the subject is asked to perform some tasks and the system records BOLD changes. Usually the number of voxels (n) is greater than 100,000 and the number of time points (T) varies from 100 to 2,000 (*Dale*, 1999; *Marshall et al.*, 2008; *Lindquist et al.*, 2008). Thus, the recorded signal consists of n time series, each with T observations.

All of these functional imaging modalities produce four-dimensional data but with different temporal and spatial resolutions. The temporal resolution of EEG and MEG, on the order of milliseconds, is much better than that of fMRI and PET (between

seconds and minutes). However, EEG and MEG have obvious limitation in spatial resolution, on the order of 6 cm^3. In contrast, the spatial resolution of PET is on the order of 200 mm^3. The spatial resolution of fMRI is typically on the order of 8-36 mm^3 and can be as small as 1 mm^3. In addition to the differences in temporal and spatial resolution, both PET and fMRI measure neuronal activity indirectly, while signals recorded by EEG and MEG directly measure the current generated by neurons in the brain.

1.2 Statistical Methods of Neuroimaging Data Analysis

Prior to statistical analysis, a series of preprocessing steps are performed on the data to minimize the influence of artifacts and to remove extraneous sources of variation. The steps in the fMRI preprocessing pipeline involve interpolation, slice timing correction, motion correction, registration, normalization and spatial smoothing (*Ombao et al.*, 2016), (*Lindquist et al.*, 2008). These preprocessing steps are crucial for further statistical analysis and validation of model assumptions.

After preprocessing, which statistical method that is applied to neuroimaging data much depends on the goal of the study (*Ombao et al.*, 2016; *Lazar*, 2008). High-resolution structural imaging has been extensively used in the clinical setting, since it provides detailed anatomical information, is sensitive to many pathologies and assists in diagnosis of disease (*Atlas*, 2009). In particular, with the help of statistical methods, structural MRI (sMRI) plays a fundamental role in the diagnosis, management, and study of multiple sclerosis (MS), stroke, cancer, traumatic brain injury, and Alzheimer's disease (AD). For example, lesion segmentation is a classification problem and solved by training statistical classifiers (e.g. logistic regression, support vector machine, or random forest) given the voxel-level intensity information from sMRI (*Pham et al.*, 2000; *Balafar et al.*, 2010; *Lladó et al.*, 2012; *Sweeney et al.*, 2013b). Another example is the study of the association between lesion localization

and health outcomes (e.g. treatment groups, disease stages or subtype groups), simple summary statistics and voxel-wise regression methods are commonly used (*Charil et al.*, 2007, 2003; *Sepulcre et al.*, 2009; *Rossi et al.*, 2012; *Ge et al.*, 2014a). In summary, almost all of the statistical methods used for sMRI data analyses are simple and disregard spatial dependence amongst voxels.

Many statistical methods have been proposed for various analysis of functional neuroimaging data, in particular fMRI data (*Lindquist et al.*, 2008; *Lazar*, 2008), as both temporal and spatial information provide researchers unprecedented access to brain activity. Here, we primarily focus on the massive univariate approach, the most common and the standard statistical method to measure brain activity in designed experiments, that constructs brain activation maps (*Lazar*, 2008). This approach analyzes fMRI time series, one voxel at a time. In this context, the general linear model (GLM) (*Martin and Maes*, 1979) is the statistical model of choice (*Worsley and Friston*, 1995). Statistical inference is performed using the statistical parametric map (SPM), summarizing test statistics (e.g. Student t-tests) for all the voxels. The classical massive univariate approach ignores spatial dependence in both model fitting and inference.

1.3 Accounting For Spatial Dependence

The statistical models and methods described in section 1.2 are for voxel-wise analysis of fMRI data. However, an important aspect of both structural and functional neuroimaging data is spatial dependence amongst voxels.

For some analyses of structural neuroimaging data the particular classification algorithm or regression method is less important than characteristics of features (*Sweeney et al.*, 2014; *Hand*, 2006) and that spatial dependence is one of the crucial and widely ignored characteristics in the statistical analysis of neuroimaging data. Meanwhile, some voxel-wise approaches proposed for fMRI data analysis introduce

substantial estimation bias and are not optimal for prediction because they do not account for spatial dependence (*Li et al.*, 2011; *Polzehl et al.*, 2010). In addition, whole-brain voxel-wise approaches are not optimal for multivariate analyses such as pattern recognition since voxels lack biological meaning (*Wolfers et al.*, 2015; *Mwangi et al.*, 2014).

There is some literature that incorporates spatial dependence in neuroimaging data analyses. See the review in (*Bowman et al.*, 2008). The key in modelling spatial dependence in neuorimages is the construction and estimation of the covariance matrix amongst voxels. Because neuroimaging data typically has a very large number of voxels it is computational prohibitive to estimate large unstructured covariance and precision matrices (*Bowman et al.*, 2008; *Hyun et al.*, 2014). An alternative approach is to reduce the dimension of the covariance matrix by using a region of interest approach or by partitioning the brain into a small number of parcels (*Bowman*, 2007). Under the Bayesian framework, spatial priors are commonly used such as conditional autoregressive (CAR) priors, Gaussian process priors and Markov random field priors (*Gössl et al.*, 2001; *Katanoda et al.*, 2002; *Bowman*, 2005; *Brezger et al.*, 2007; *Groves et al.*, 2009). However, these Bayesian approaches come at a huge computational cost. usually brings heavy computation to calculate the tuning parameters defined for those priors.

1.4 Dissertation Outline

This dissertation is organized as follows. In Chapter II, we propose a scalar-on-image regression model to predict clinical subtypes of Multiple Scelrosis using structural MRI data. In Chapter III, we develop a spatial Bayesian latent factor model for image-on-image regression to predict task-related contrast maps given a set of task-free images. In Chapter IV, the model developed in Chapter II is extended to predict surface images from volumetric images. We wrap up the dissertation with a

summary and ideas for future extensions.

CHAPTER II

Scalar-on-Image Regression with Application to Multiple Sclerosis MRI Data

Multiple sclerosis (MS) is an autoimmune disease that attacks the central nervous system. Furthermore, MS lesions are visible on magnetic resonance (MR) images. Hence, magnetic resonance imaging (MRI) plays a central in the diagnosis and management of MS patients. A research question of interest is whether MS lesion data, via MR images, can predict the subtypes of MS. To answer this question we propose a Bayesian scalar-on-image regression model. Specifically, a polytomous logistic regression model to predict a patient's clinical MS subtype using her 3D MRI lesion data as predictors. We assume that the parameters corresponding to these predictors are spatially smooth (correlated) and use a 3D Gaussian Process (GP) prior to model these parameters and their correlation structure. These parameters are of high dimension, one parameter for each of the thousands of voxels, and tend to be highly correlated, at least locally. This high-dimensional problem results in a computationally intense Bayesian algorithm. Thus, to speed up the algorithm we adopt three ideas: 1) we sample the discretized GP via the Hamiltonian Monte Carlo (HMC) algorithm which increases mixing compared to standard Markov chain Monte Carlo algorithms in high-dimensional problems with highly correlated parameters; 2) since the data reside on a regular lattice, we embed the covariance matrix (a nested block Toeplitz matrix)

of the approximate Gaussian Process in a nested block circulant matrix and leverage the relationship between these matrices and the 3D Fourier transform; and 3) we leverage the parallelism of the fast Fourier transform and of the HMC algorithm and port these portions of the algorithm onto a graphical processing unit. We show via simulation studies and via an MS data set that our model has high prediction accuracy as evaluated by leave-one-out cross validation.

2.1 Introduction

MS is an unpredictable and often disabling autoimmune disease of the central nervous system. Chronic inflammation and neuronal demyelination in the brain and the spinal cord disrupt action potentials within the brain and between the brain and body, that then cause a multitude of clinical symptoms including, but not limited to, blindness, muscle weakness and gait impairment (*Compston and Coles*, 2002). Symptoms occur in isolation (relapsing form), gradually develop over time (progressive form), or in a combination of both (*Lublin et al.*, 1996). MS is subtyped based on progression pattern. Until recently, MS was subtyped into 4 distinct clinical courses: relapsing-remitting (RLRM), secondary progressive (SCP), primary progressive (PRP) and progressive relapsing (*Lublin et al.*, 2014). However, now, progressive relapsing is no longer considered to be a subtype as its definition was vague and its progression pattern overlapped with other MS subtypes (*Lublin et al.*, 2014). While currently there is no cure for MS, several treatment therapies are beneficial in reducing attacks in RLRM disease. These drugs are only indicated for RLRM disease and thus it is crucial to precisely determine subtype as early as possible. (*Lövblad et al.*, 2010).

MRI is an important tool used in the diagnosis of MS and in monitoring its progression (*Lövblad et al.*, 2010). MS lesions are visible on MRI images. On T1-weighted MR images, lesions from long-term demyelination appear as "black holes". on T2-weighted MR images, inflammatory lesions appear as hyper-intense regions due

to edema. Hence, MRI is a useful tool to monitor MS disease course in both time and space (*Lövblad et al.*, 2010; *Bakshi et al.*, 2005).

Although MRI is an important tool for management of MS, conventional MRI findings are poorly correlated with clinically observed disease progression (*Ge et al.*, 2014b). Some methodological and statistical analyses have been developed to explore the association between MRI and clinical progression of MS. These analyses focused on two main issues: 1) the distribution of lesions counts across patients from different MS subtypes and 2) the role of MRI variables as surrogate markers in MS clinical trials (*Sormani and Filippi*, 2007). Among these studies, "lesion load" is the most common and simple to use. Lesion load is simply the total lesion volume obtained from MRI and attempts to predict clinical outcome as well as changes over time as predictors (*Calabrese et al.*, 2012; *Moodie et al.*, 2012). Other authors have focused on "mass univariate" techniques. That is independent voxel-by-voxel analyses of lesion probability maps (LPM) to compare the distribution of lesions from different MS subtypes (*Filli et al.*, 2012; *Holland et al.*, 2012) or correlate the lesion maps with clinical subtypes (*Bates et al.*, 2003). Both of these methods ignore spatial correlation inherent in images. Furthermore, several studies have shown poor predictive performance of clinical MS subtypes by using these methods. (*Lövblad et al.*, 2010; *MacKay Altman et al.*, 2012; *Morgan et al.*, 2010).

In contrast to these somewhat naive methods, scalar-on-image regression may be a more appropriate statistical model in this context. In its simplest form, scalar-on-image regression is a linear model with scalar response y_i, and image covariate $\mathbf{x}_i$ (written as a vector):

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \cdots, N$$

for each of N subjects, where $\boldsymbol{\beta}$ is an image of parameters (also written as a vector).

In this chapter, $\mathbf{x}_i$ is a 3-dimensional discretized image of say L voxels. Typically $L \gg N$ so that this model is unidentifiable unless we impose some constraints on $\boldsymbol{\beta}$. Within the Bayesian context one can impose a spatial prior on $\boldsymbol{\beta}$ such as a Gaussian Process (GP) (*Rasmussen and Williams*, 2006) prior or a Gaussian Markov random field (GMRF) prior (*Besag*, 1974a; *Rue and Held*, 2005). In some contexts, authors have imposed sparsity on $\boldsymbol{\beta}$ as well (*Goldsmith et al.*, 2014; *Kang et al.*, 2016; *Li et al.*, 2015). In contrast, some authors have chosen to approximate $\boldsymbol{\beta}$ via a basis expansion. The idea behind this basis function approach is that the unknown image coefficients $\boldsymbol{\beta}$ can be approximated by a span of K known basis functions $B_1, B_2, \cdots, B_K$. The total number of image coefficients is hence reduced from L to K, typically with $K \ll L$, since $\beta_v \approx \sum_{k=1}^{K} b_k B_k(v)$. The choice of the basis functions can be 1) fixed basis functions, e.g. B-Splines (*Marx and Eilers*, 2005) or Wavelets (*Reiss et al.*, 2015), 2) data-driven basis functions, e.g. principle component regression (*Müller and Stadtmüller*, 2005; *Allen*, 2013), or 3) a combination of the two (*Reiss and Ogden*, 2010a).

All above methods overcome the issue of non-identifiability and make estimation feasible by giving structural model assumptions. However, these assumptions come at the price of inducing bias. Both basis function approaches and random field methods are based on assumptions of smoothness. In contrast, the image coefficients $\boldsymbol{\beta}$ are modeled directly by using random field methods without any approximations other than discretization. What's more, sparsity is not an appropriate assumption for prediction as an MS patient whose subtype is to be predicted may have lesions in areas of the brain outside the assumed sparse model solution.

Therefore, in this paper, we only assume smoothness of the image coefficients $\boldsymbol{\beta}$ by proposing a Gaussian Process (GP) prior for $\boldsymbol{\beta}$ without sparsity and projection assumptions. To our knowledge, a GP prior has never been used in this context with such a large 3-dimensional image parameter $\boldsymbol{\beta}$. A GP is a probabilistic measure in a

continuous domain, e.g. time or space, where any finite number of random variables have a joint Gaussian distribution (*Rasmussen and Williams*, 2006). It is only determined by a mean function, commonly assumed to be zero, and a covariance function (or kernel) with just a few hyperparameters. Because it provides fully probabilistic predictive distributions and it is a non-parametric model, the GP prior can be used for a wide variety of tasks and is a popular prior model in Bayesian non-parametric data analysis. However, in this 3D imaging context with several thousand voxels, Gaussian processes are limited by their computational complexity—estimating more parameters than voxels. Many approximation techniques have been proposed to overcome this issue within the machine learning community. Some sparse approximation methods (*Smola and Bartlett*, 2001; *Seeger et al.*, 2003; *Snelson and Ghahramani*, 2006) introduce latent variables which are then treated exactly, while all other parameters are approximated to overcome the computation limitation (*Quiñonero-Candela and Rasmussen*, 2005). The choice of the number of latent variables determines the performance in these sparse models, but no criteria are set forth for determining the optimal number of latent variables that balance computational cost and estimation accuracy. Some other straightforward method use iterative methods, e.g. conjugate gradients (*Golub and Van Loan*, 2012), to speed up GP regression and then reduce the number of iterations to get approximate solutions (*Gibbs and MacKay*, 1996; *Yang et al.*, 2005). This algorithmic strategy, however, is most effective when the input space is low dimensional. No matter the type of GP approximation method used, computational speed is improved but at the price of estimation bias. Accurate estimation is crucial for prediction of MS subtypes due to the overlap in the spatial distribution of MS lesions amongst the three subtypes. Estimation bias reduces prediction accuracy.

Therefore, in this chapter we do not attempt to approximate the Gaussian Process, rather, we overcome the computational limitations by imbedding the covariance ma-

trix into a larger matrix. In three dimensions our proposed covariance matrix of the discretized GP (a multivariate normal distribution) has a nested block Toeplitz structure. This structure can be embedded into a nested block circulant matrix (*Wood and Chan*, 1994). Leveraging the relationship between this circulant matrix and the 3D Fourier transform, computation cost is significantly reduced *Wood and Chan* (1994). Furthermore, both the fast Fourier transform (FFT) and the HMC algorithm are ameniable to parallelization. Hence, leveraging parallel computing techniques and porting much of the computation to a graphical processing unit (GPU), we significantly reduce the computational burden of our model.

The goals of our work include 1) to build a predictive scalar-on-image regression model of MS subtypes based on both clinical covariates and MRI lesion; 2) to model the spatial dependence in the images using GP without specifying assumptions and approximations other than discretization; and 3) to estimate model parameters, in particular the GPs, efficiently under the Bayesian framework. In Section 2.2, we first formulate our scalar-on-image regression model and demonstrate how we consider the spatial dependence existing in voxel-wise parameters. A Bayesian framework and the HMC algorithm used for estimation are then introduced. At the end of this section, we explain how to conduct efficient calculations based on Fast Fourier Transform (FFT) algorithm. In Section 2.4, we present a simple 2-dimensional simulation study. Section 2.5 includes a real data analysis of MS MRI lesion data. In section 2.6, we conclude the paper with a brief discussion.

2.2 The Model

2.2.1 Bayesian Scalar-on-Image Regression Model

We develop a Bayesian scalar-on-image regression model to estimate the probabilities of the different possible outcomes of a scalar variable, given a 3D binary image

as well as some non-spatially varying variables. The cornerstone of this scalar-on-image regression model is the multinomial logistic regression model with a logit link function.

Let $\mathbb{R}^d$ be a d-dimensional space of real values for any integer $d \geq 1$. Suppose there are N subjects in the dataset. For each subject i, we have their scalar response variable $Y_i \in \{0, 1, \cdots, K-1\}$, a p-vector of non-spatially distributed variables $\boldsymbol{Z}_i^{\mathrm{T}} = (\boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \cdots, \boldsymbol{Z}_{ip}) \in \mathbb{R}^p$. Let $\mathcal{B} \subset \mathbb{R}^3$ denote the common brain space of all subjects. Note here that each subject's brain is warped to a representative brain atlas, $\mathcal{B}$. Let $s \in \mathcal{B}$ and let $X_i : \mathcal{B} \to \{0, 1\}$ denote the function that maps the brain space to the set $\{0, 1\}$. In other words, X_i is a binary image for subject i where if subject i has a lesion at location s, $X_i(s) = 1$, otherwise $X_i(s) = 0$. Let $\pi_{ik} = \Pr(Y_i = k)$. The baseline-categories logistic link function, with baseline category $k = 0$, is then used to relates the expectation of the random outcome to the systematic component as below

$$\begin{aligned} \log\left(\frac{\pi_{ik}}{\pi_{i0}}\right) &= \eta_{ik}, \quad k = 1, \ldots, K-1 \\ \eta_{ik} &= \alpha_k + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_k + \int_{\mathcal{B}} X_i(s)\beta_k(s)\mathrm{d}s \quad \text{so that} \\ \pi_{ik} &= \frac{\exp\left(\eta_{ik}\right)}{1 + \sum_{l=1}^{K-1}\exp\left(\eta_{il}\right)} \end{aligned} \tag{2.1}$$

The systematic component η_{ik} is given by (2.1), containing two main terms. The non-spatially varying term $\alpha_k + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_k$ consists of the intercept α_k and the subject-specific covariate effects $\boldsymbol{\gamma}_k^{\mathrm{T}} = (\gamma_{k1}, \gamma_{k2}, \cdots, \gamma_{kp})$ corresponding to $\boldsymbol{Z}_i$ and the spatially varying term, $\int_{\mathcal{B}} X_i(s)\beta_k(s)\mathrm{d}s$, that integrates lesion location information over $\mathcal{B}$ with spatially varying coefficient function $\beta_k : \mathcal{B} \to \mathbb{R}$.

2.2.2 Modelling β_k with a Gaussian Process Prior

To model spatial dependence, we assume that the spatially varying coefficient functions, β_k, $k = 1, \ldots, K-1$, independently follow a GP priors with a zero mean function and covariance (kernel) function $\sigma_k^2 \boldsymbol{C}(s-s'; \rho_k)$: $\beta_k \sim \mathcal{GP}(\mathbf{0}, \sigma_k^2 \boldsymbol{C}(s-s'; \rho_k))$. The marginal variance is σ_k^2 and the correlation function $\boldsymbol{C}(s-s'; \rho_k)$ are determined by a strictly positive decay parameter ρ_k, which controls the smoothness of the process and the distance between two points s and s', respectively. In this chapter we adopt the Gaussian kernel: $\boldsymbol{C}(s-s'; \rho_k) = \exp(-\rho_k \|s-s'\|^2)$, where $\|\cdot\|$ signifies Euclidean distance. As ρ_k decreases, the process becomes increasingly smooth and as the distance between s and s' increases, the correlation decreases.

2.2.3 Discretization

Obviously, the above model is infinite dimensional and not amenable to computation. By definition a GP is a stochastic process (in our case indexed by space) such that every finite dimensional collection of these random variables follows a multivariate normal distribution (MVN). Thus by discretizing space, the brain, into a set of disjoint rectangular prisms (voxels) and evaluating the GP at the centriod of each voxel, we can approximate each GP with a multivariate normal distribution. The voxels are selected to correspond to the voxels that define the discreted brain atlas. Within the space occupied by each voxel, the value of the approximated GP is taken to be constant and is equal to the expectation at the centroid of this voxel defined by the associated MVN.

As noted above, we descritize the brain into M voxels denoted by s_j, $j = 1, \ldots, M$. Let $\mathbf{S}^{\mathrm{T}} = (s_1, \ldots, S_M)$ where the three dimensional image has been vectorized. We next discretize each function β_k into an M-dimensional column vector $\boldsymbol{\beta}_k^{\mathrm{T}} = (\beta_{ks_1}, \ldots, \beta_{ks_M})$ with a slight abuse of notation where s_j represents both the space occupied by a voxel when referring to the GP and represents the centroid of the same

voxel when referring to the MVN distribution. So $\boldsymbol{\beta}_k \sim N_M(\mathbf{0}, \sigma_k^2 \mathbf{C}(\rho_k))$ where the (i,j)th element of the correlation matrix is given by $c_{ij}(\rho_k) = \exp(-\rho_k \|s_i - s_j\|^2)$. Correspondingly we descritize the binary-valued fuctions X_i on the same grid of voxels: $\mathbf{X}_i^{\mathrm{T}} = (X_{is_1}, \ldots, X_{is_M})$. For computational reasons (see Section 2.3) we reparametrize $\boldsymbol{\beta}_k = \sigma_k \mathbf{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{\zeta}_k$ where $\boldsymbol{\zeta}_k \sim N_M(\mathbf{0}, \mathbf{I})$.

Thus, the scalar-on-image regression model on the discretized space is given by

$$\log\left(\frac{\pi_{ik}}{\pi_{i0}}\right) \approx \alpha_k + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_k + \sigma_k \mathbf{X}_i^{\mathrm{T}} \mathbf{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{\zeta}_k. \tag{2.2}$$

2.2.4 Prior and Posterior Distributions

A Bayesian framework is formulated for inference and prediction. Briefly, we assume that the intercept term α_k following the standard normal distribution that $\alpha_k \sim N(0,1), k = 1, \ldots, K-1$. The non-spatially varying parameter vectors $\boldsymbol{\gamma}_k$ are given a prior MVN distributions:

$$\boldsymbol{\gamma}_k | \boldsymbol{U}_k, \boldsymbol{\Sigma}_k \sim \mathrm{N}_p(\boldsymbol{U}_k, \boldsymbol{\Sigma}_k)$$

with the hyperpriors assigned MVN and an inverse-Wishart distributions

$$\boldsymbol{U}_k | \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0 \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$
$$\boldsymbol{\Sigma}_k | \nu_0, \boldsymbol{\phi}_0 \sim \text{Inverse-Wishart}(\nu_0, \boldsymbol{\phi}_0).$$

As defined in Section 2.2.1, for the discretized spatially varying parameter vectors $\boldsymbol{\zeta}_k$, we place a prior distribution on it as below

$$\boldsymbol{\zeta}_k \overset{i.i.d.}{\sim} \mathrm{N}_M(\mathbf{0}, \boldsymbol{I}).$$

The hyper-parameters in GP priors follow gamma distributions

$$\begin{aligned}\sigma_k|a_\sigma, b_\sigma &\sim \text{Gamma}(a_\sigma, b_\sigma)\\ \rho_k|a_\rho, b_\rho &\sim \text{Gamma}(a_\rho, b_\rho)\end{aligned}$$

where the gamma distribution is parametrized such that the mean is a/b and variance is a/b^2.

We denote all parameters in our model by the set $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_{K-1}\}$ where $\boldsymbol{\theta}_k = \{\alpha_k, \boldsymbol{\gamma}_k, \sigma_k, \rho_k, \boldsymbol{\zeta}_k, \boldsymbol{U}_k, \boldsymbol{\Sigma}_k\}$. And $\boldsymbol{\theta}_{-k}$ denotes all other parameters in $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_k$. Since response category $Y_i = 0$ serves as the baseline group, α_0, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\zeta}_0$ are all set to $= 0$. Therefore, the joint posterior density of $\boldsymbol{\theta}_k$, $k \neq 0$, is

$$\begin{aligned}\pi(\boldsymbol{\theta_k}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta_{-k}}, \boldsymbol{\Omega}) = &\left[\prod_{i=1}^{N}\prod_{l=0}^{K-1} \pi_{il}^{I(y_i=l)}\right] \times \pi(\alpha_k)\pi(\boldsymbol{\gamma}_k; \boldsymbol{U}_k, \boldsymbol{\Sigma}_k)\pi(\boldsymbol{\zeta}_k)\\ &\times \pi(\sigma_k; a_\sigma, b_\sigma)\pi(\rho_k; a_\rho, b_\rho)\pi(\boldsymbol{U}_k; \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)\pi(\boldsymbol{\Sigma}_k; \boldsymbol{\nu}_0, \Phi_0)\end{aligned} \tag{2.3}$$

where $\boldsymbol{\Omega}$ is the set of all fixed, known hyperparameters: $a_\sigma, b_\sigma, a_\rho, b_\rho, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, \nu_0$, and $\boldsymbol{\Phi}_0$. Thus the log posterior density is

$$\begin{aligned}\log \pi(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) = &\left[C + \sum_{i=1}^{N}\sum_{l=0}^{K-1} I(y_i = l)\log(\pi_{il})\right] - \frac{1}{2}\alpha_k^2 - \frac{1}{2}\boldsymbol{\zeta}_k^{\mathrm{T}}\boldsymbol{\zeta}_k\\ &- \frac{1}{2}(\boldsymbol{\gamma}_k - \boldsymbol{U}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\gamma}_k - \boldsymbol{U}_k)\\ &+ (a_\sigma - 1)\log(\sigma_k) - b_\sigma\sigma_k + (a_\rho - 1)\log(\rho_k) - b_\rho\rho_k\end{aligned} \tag{2.4}$$

where $I(\cdot)$ is the indicator function and C is a constant with respect to $\boldsymbol{\theta}_k$.

2.3 Algorithms

By conjugacy, $\boldsymbol{U}_k$ and $\boldsymbol{\Sigma}_k$ are sampled via a Gibbs updates (*Gelman et al.*, 2014). Based on the joint posterior distribution of $\boldsymbol{\theta}_k$ in (2.3), however, all other parameters α_k, $\boldsymbol{\gamma}_k$, $\boldsymbol{\zeta}_k$, σ_k and ρ_k don't have closed-form full conditional posterior distributions. The Metropolis-Hastings algorithm (*Gelman et al.*, 2014) is a possible way to sample these parameters but inefficient due to the high-dimensional and highly correlated parameter space of the spatially varying parameters. Thus, we resort to the HMC algorithm (*Neal et al.*, 2011) to sample α_k, $\boldsymbol{\gamma}_k$, $\boldsymbol{\zeta}_k$, σ_k as well as ρ_k. HMC is well suited to sample from high-dimensional, highly correlated target distributions. The HMC algorithm is introduced in the following section.

2.3.1 Hamiltonian Monte Carlo Algorithm

The HMC algorithm is based on Hamiltonian dynamics and consists of a d-dimensional position vector, $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\gamma}_k, \boldsymbol{\zeta}_k, \sigma_k, \rho_k)$, the variables that we wish to sample using HMC, and an artificial d-dimensional momentum vector $\boldsymbol{\xi}_k$, one for each position variable. The seperable Hamiltonian function, the total system energy, $H(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)$ is partitioned into potential energy, $U(\boldsymbol{\theta}_k)$, and kinetic energy, $K(\boldsymbol{\xi}_k)$:

$$
\begin{aligned}
H(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k) &= U(\boldsymbol{\theta}_k) + K(\boldsymbol{\xi}_k) \\
U(\boldsymbol{\theta}_k) &= -\log\left[P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega})\right] \\
K(\boldsymbol{\xi}_k) &= \boldsymbol{\xi}_k^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{\xi}_k / 2
\end{aligned}
$$

The potential energy is defined as the negative log of the joint posterior distribution for $\boldsymbol{\theta}_k$. The momentum term is a quadratic form (not necessary, but is in our implementation) where $\boldsymbol{M}$ denotes a "mass matrix", which is typically diagonal and often a scalar multiple of the identity matrix. $K(\boldsymbol{\xi}_k)$ is thus the negative log density of the mean zero MVN vector $\boldsymbol{\xi}_k$ with covariance M.

The dynamics are defined as the partial derivatives of the Hamiltonian $H(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)$ and determine how $\boldsymbol{\theta}_k$ and $\boldsymbol{\xi}_k$ change over time t. Mathematically,

$$\frac{d\theta_{k,j}}{dt} = \frac{\partial H(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)}{\partial \xi_{k,j}} = \frac{\partial K(\boldsymbol{\xi}_k)}{\partial \xi_{k,j}} \tag{2.5}$$

$$\frac{d\xi_{k,j}}{dt} = -\frac{\partial H(\boldsymbol{\theta}_k, \boldsymbol{\xi}_k)}{\partial \theta_{k,j}} = -\frac{\partial U(\boldsymbol{\theta}_k)}{\partial \theta_{k,j}} \tag{2.6}$$

for $j = 1, 2, ..., d$, where j represents the j^{th} element in the corresponding random vector.

In practice, the Hamiltonian equations (2.5) and (2.6) are approximated by discretizing time t. By using a small step size, δ, and starting from time $t = 0$, we iteratively and approximately compute the state at time δ, 2δ, 3δ and so on. A commonly used approximation algorithm is the leapfrog algorithm (*Neal et al.*, 2011):

$$\boldsymbol{\xi}_k(t + \delta/2) = \boldsymbol{\xi}_k(t) - \delta/2 \frac{\partial U}{\partial \boldsymbol{\theta}_k}\big(\boldsymbol{\theta}_k(t)\big) \tag{2.7}$$

$$\boldsymbol{\theta}_k(t + \delta) = \boldsymbol{\theta}_k(t) + \delta \boldsymbol{M}^{-1} \boldsymbol{\xi}_k(t + \delta/2)$$

$$\boldsymbol{\xi}_k(t + \delta) = \boldsymbol{\xi}_k(t + \delta/2) - \delta/2 \frac{\partial U}{\partial \boldsymbol{\theta}_k}\big(\boldsymbol{\theta}_k(t + \delta)\big) \tag{2.8}$$

Starting from the current state $\boldsymbol{\theta}_k(t)$ and $\boldsymbol{\xi}_k(t)$ at time t, Hamiltonian dynamics simulates L steps using the leapfrog algorithm and the resulting approximate solution is used as a proposed value for the next state of the Markov chain in a Metropolis-Hastings algorithm. We tune the total number of leapfrog steps L and the step size δ by using the No U-turns sampling algorithm (*Hoffman and Gelman*, 2014).

To implement the HMC algorithm, we need to calculate the log joint posterior density (2.4) and the gradients (2.7) and (2.8). Let $P(\boldsymbol{\theta}_k|\cdot) = P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega})$.

The gradients of the log joint posterior are

$$\begin{aligned}
\nabla_{\alpha_k} \log P(\boldsymbol{\theta}_k|\cdot) &= C_1 + \left[\sum_{i=1}^{N} \big(I(y_i = k) - \pi_{ik}\big)\right] - \alpha_k \\
\nabla_{\boldsymbol{\zeta}_k} \log P(\boldsymbol{\theta}_k|\cdot) &= C_2 + \left[\sum_{i=1}^{N} \big(I(y_i = k) - \pi_{ik}\big)\sigma_k \boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{X}_i\right] - \boldsymbol{\zeta}_k \\
\nabla_{\boldsymbol{\gamma}_k} \log P(\boldsymbol{\theta}_k|\cdot) &= C_3 + \left[\sum_{i=1}^{N} \big(I(y_i = k) - \pi_{ik}\big)\boldsymbol{Z}_i\right] - \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\gamma_k} - \boldsymbol{U}_k) \\
\nabla_{\sigma_k} \log P(\boldsymbol{\theta}_k|\cdot) &= C_4 + \left[\sum_{i=1}^{N} \big(I(y_i = k) - \pi_{ik}\big)\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{\zeta}_k\right] + \frac{a_\sigma - 1}{\sigma_k} \\
\nabla_{\rho_k} \log P(\boldsymbol{\theta}_k|\cdot) &= C_5 + \left[\sum_{i=1}^{N} \big(I(y_i = k) - \pi_{ik}\big)\sigma_k \boldsymbol{X}_i^{\mathrm{T}}\left(\frac{\partial \boldsymbol{C}^{\frac{1}{2}}(\rho_k)}{\partial \rho_k}\right)\boldsymbol{\zeta}_k\right] + \frac{a_\rho - 1}{\rho_k},
\end{aligned} \tag{2.9}$$

where C_1, C_2, C_3, C_4, C_5 are constants.

Since σ_k and ρ_k are constrained to be positive, these parameters must satisfy these constraints during the leapfrog updating. For example, given upper and lower bounds of $\boldsymbol{\theta}_k$ denoted as $\boldsymbol{u}$ and $\boldsymbol{l}$, we repeat following steps until $\boldsymbol{u} \leq \boldsymbol{\theta}_k(t+\delta) \leq \boldsymbol{l}$

- if $\boldsymbol{\theta}_k(t+\delta) > \boldsymbol{u}$, then

$$\boldsymbol{\theta}_k(t+\delta) = \boldsymbol{u} - \big(\boldsymbol{\theta}_k(t+\delta) - \boldsymbol{u}\big) \quad \text{and} \quad \boldsymbol{\xi}_k(t+\delta/2) = -\boldsymbol{\xi}_k(t+\delta/2)$$

- if $\boldsymbol{\theta}_k(t+\delta) < \boldsymbol{l}$, then

$$\boldsymbol{\theta}_k(t+\delta) = \boldsymbol{l} + \big(\boldsymbol{l} - \boldsymbol{\theta}_k(t+\delta)\big) \quad \text{and} \quad \boldsymbol{\xi}_k(t+\delta/2) = -\boldsymbol{\xi}_k(t+\delta/2).$$

Of course in practice, we only apply lower bounds to σ_k and ρ_k,

Given the above algorithm, we obtain proposed momentum and parameter vectors: $\boldsymbol{\xi}_k^*$ and $\boldsymbol{\theta}_k^*$. This proposed state is then accepted as a draw from the posterior at time

$t+1$ with probability

$$\min\left[1, \exp\Big(-H(\boldsymbol{\xi}_k^*, \boldsymbol{\theta}_k^*) + H(\boldsymbol{\xi}_k, \boldsymbol{\theta}_k)\Big).\right]$$

2.3.2 Fast Fourier Transform Algorithm

As seen in the last section, two of the gradients depend on the Cholesky decomposition of $\mathbf{C}(\rho_k)$ and one on its partial derivative with respect to ρ_k. $\mathbf{C}(\rho_k)$ is a very large matrix. In our data example the image dimension is $64 \times 64 \times 64$. Therefore the length of the MVN vector $\boldsymbol{\beta}_k$ is $M = 64^3 = 262144$ and $\mathbf{C}(\rho_k)$ is of dimension $64^3 \times 64^3$ and the calculation of its Cholesky decompostion and its partial derivative is too computationally expensive to take directly. Therefore we follow *Wood and Chan* (1994) and embed $\mathbf{C}(\rho_k)$—a nested block toeplitz matrix into a nested block circulant matrix that is symmetric, say $\mathbf{G}(\rho_k)$. Although this appears counter intuitive, there is a direct relationship between the eigen decomposition of these matrices and the 3D discrete Fourier transform. Thus, computation of the Cholesky decomposition can be achieved efficiently in the Fourier domain.

In the remainder of this section, we will give details of this embedding for a 3D image. Suppose the image dimension is $n_x \times n_y \times n_z$ with voxel dimensions $v_x \times v_y \times v_z$. Recall that for the image the correlation between any two spatial locations s_i and s_j (centers of two voxels) for subtype k is $\exp(-\rho\|s_i - s_j\|^2)$. Since all voxels have the same dimension, this correlation matrix has a nested block Toeplitz structure. To embed this correlation matrix into a symmetric nested block circulant matrix we first embed the image into a larger image with dimensions $m_x \times m_y \times m_z$ where $m_d = 2^{q_d}, d \in \{x, y, z\}$, q_d an integer with the restriction that $2^{q_d} \geq 2(n_d - 1)$ with the additional requirement that each new voxel has dimensions $v_x \times v_y \times v_z$. We vectorize this image and call the resulting vector $\tilde{\mathbf{S}}$. The choice of restricting q_d to an integer implies that the new dimensions m_d is highly composite so that we can

efficiently use the the 3D FFT algorithm.

Next the vectors $\boldsymbol{\beta}_k$, $\boldsymbol{\zeta}_k$, both of length $n_x n_y n_z$, are embedded into larger vectors, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}}$ and $\tilde{\boldsymbol{\zeta}}_{\boldsymbol{k}}$ of lengths $m_x m_y m_z$. $\tilde{\boldsymbol{\beta}}_{\boldsymbol{k}}$ has a MVN distribution with mean zero and nested block circulant correlation matrix $\mathbf{G}(\rho_k)$, however care must be taken to ensure that this new matrix is symmetric. The restriction on the q_d is the minimum value that allows this matrix to be symmetric, but distances between voxels must be defined appropriately. In the 1D case, the extended line is wrapped onto a circle and the distance between two points is defined as the shortest distance on the circle (see Figure 2.1). In the 2D case, the extended grid is wrapped onto the surface of a torus and the distance between two pixels is defined as the shortest distance on the torus. Analogous results apply to dimensions 3 and higher (see *Wood and Chan* (1994)). $\mathbf{G}(\rho_k)$ is thus an $m_x m_y m_z \times m_x m_y m_z$ symmetric, nested block circulant matrix with $\mathbf{C}(\rho_k)$ embedded within. In the remainder of this section we will let $Q = m_x m_y m_z$.

A standard result of symmetric nested block circulant matrices is that there always exists an eigen decomposition of the matrix: $\boldsymbol{G}(\rho_k) = \boldsymbol{Q}\boldsymbol{\Lambda}(\rho_k)\boldsymbol{Q}^*$. Here, $\boldsymbol{\Lambda}(\rho_k)$ is the $m_x m_y m_z \times m_x m_y m_X$ diagonal matrix of eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_Q$ for $\boldsymbol{G}(\rho_k)$. $\mathbf{Q}$ is the eigen-matrix, a unitary matrix, and $\mathbf{Q}^*$ is its complex conjugate transpose. This eigen decomposition of $\boldsymbol{G}(\rho_k)$ is equivalent to the 3D discrete Fourier transform of the base of this matrix. The base can be taken as the first row of this matrix. Call it $G = (g_0, g_1, \ldots, g_{Q-1})$ with associate voxel locations $s_0, \ldots, s_{Q-1}$ where $g_l = \exp(-\rho_k d_l^2)$ and d_l is the minimum distance between voxels s_l and s_0 defined on the 3D torus about which the extended grid is wrapped. Also, let d_l denote the shortest distance between voxel v_0 and voxel v_l on the 3D torus on the extended grid wrapped onto the surface of a 3D torus The jth eigenvalue of G is

$$\lambda_j = \sum_{l=0}^{Q-1} g_l \exp\left(-\frac{2\pi i l j}{Q}\right), \quad j = 0, 2, \cdots, Q-1 \tag{2.10}$$

where $i = \sqrt{-1}$. This eigen decomposition of the matrix $G(\rho_k)$ can be efficiently computed via the 3D FFT.

Once we have the eigenvalues of $\boldsymbol{G}(\rho_k)$, we can efficiently calculate $\boldsymbol{G}^{\frac{1}{2}}(\rho_k) = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}(\rho_k)\mathbf{Q}^*$ and $\mathbf{Q}\boldsymbol{\Lambda}^{1/2}(\rho_k)\mathbf{Q}^*\tilde{\boldsymbol{\zeta}}_k$ using the 3D FFT algorithm. Finally, by setting the corresponding augmented elements in $\tilde{\boldsymbol{\zeta}}_k$ to be 0, we recover $\boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{\zeta}_k = \boldsymbol{G}^{\frac{1}{2}}(\rho_k)\tilde{\boldsymbol{\zeta}}_k$. Analogously we can derive $\boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{X}_i$ necessary to calculate the gradient with respect to $\boldsymbol{\zeta}_k$.

We also need to calculate the partial derivative of $\rho_k\boldsymbol{C}^{\frac{1}{2}}(\rho_k)$ with respect to ρ_k based on the partial derivative of $\boldsymbol{G}^{\frac{1}{2}}(\rho_k)$ with respect to ρ_k. Let d_l denote the shortest distance between

$$
\begin{aligned}
\frac{\partial \boldsymbol{G}^{\frac{1}{2}}(\rho_k)}{\partial \rho_k} &= \boldsymbol{Q}\mathrm{diag}\left\{\frac{d\lambda_j^{1/2}}{d\rho_k}\right\}\boldsymbol{Q}^*\tilde{\boldsymbol{\zeta}}_k \\
&= \boldsymbol{Q}\mathrm{diag}\left\{\lambda_j^{-1/2}\sum_{l=1}^{Q}\left[-\frac{1}{2}d_l^2 g_l \exp(-\frac{2\pi i l j}{Q})\right]\right\}\boldsymbol{Q}^* \\
&= \boldsymbol{Q}\mathrm{diag}\left\{\lambda_j^{-1/2}\psi_j\right\}\boldsymbol{Q}^* \\
&= (\boldsymbol{Q}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^*)(\boldsymbol{Q}\boldsymbol{\Psi}\boldsymbol{Q}^*) \\
&= \boldsymbol{G}^{-\frac{1}{2}}(\rho_k)\boldsymbol{S}(\rho_k)
\end{aligned}
$$

From the above, we can see that ψ_j is the jth eigenvalue of the symmetric nested block circulant matrix $\mathbf{S}(\rho_k)$ and that $-0.5d_l^2 g_l$ is the lth element of the first row of $\mathbf{S}(\rho_k)$ and of the base of $\mathbf{S}(\rho_k)$. Hence, we only need to calculate the eigenvalues λ_j of $\mathbf{G}^{-1/2}(\rho_k)$ and the eigenvalues ψ_j of $\mathbf{S}(\rho_k)$ using the FFT perform the above partial derivative.

Furthermore, the FFT algorithm is highly parallelizable and therefore can be efficiently computed on a GPU. Also the gradients of log-joint-posterior are amenable to paralleization and can be efficiently computed on a GPU. GPUs have a massively parallel architecture consisting of thousands of smaller, more efficient cores designed

for handling identical simultaneous tasks as compared to Central Processing Units (CPUs). Our algorithm has been programmed on an NIVIDA Tesla K20 GPU that has 5GB global memory, 1,024 threads and 49,152 bytes shared memory per block.

2.4 Simulation Study

We conduct a simulation study by generating two groups of 2D images of dimension 64×64 with outcome labels $y = 0$ and $y = 1$. We follow two different probability generating matrices $\mathbf{P}_0$ and $\mathbf{P}_1$ as below

$$\mathbf{P}_0 = \begin{pmatrix} 0.8 & 0.8 - \triangle_p & 0.8 - 2\triangle_p & \cdots & 0 \\ 0.8 - \triangle_p & 0.8, & 0.8 - \triangle_p & \cdots & \triangle_p \\ 0.8 - 2\triangle_p & 0.8 - \triangle_p & 0.8 & \cdots & 2\triangle_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \triangle_p & 2\triangle_p & \cdots & 0.8 \end{pmatrix}$$

and

$$\mathbf{P}_1 = \begin{pmatrix} 0 & \triangle_p & 2\triangle_p & \cdots & 0.8 \\ 0 & \triangle_p & 2\triangle_p & \cdots & 0.8 \\ 0 & \triangle_p & 2\triangle_p & \cdots & 0.8 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \triangle_p & 2\triangle_p & \cdots & 0.8 \end{pmatrix}$$

where $\triangle_p = 0.8/(64 - 1)$. Each element in $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ is the probability that a lesion is located at the site in the image. These matrices were chosen such that the two groups have non-overlapping as well as overlapping lesion distributions.

For each group, we generate 55 images. 50 of these in each group are randomly selected for training and 5 for testing. Images of the empirical lesion probabilities

averaged over the 50 simulated images in each of the training sets are displayed in the first row of Figure 2.3. Figure 2.3 displays the difference of the two empirical lesion images.

We estimate the posterior distribution via MCMC with a total of 10,000 simulated draws and a burn-in of 5,000. The priors distributions of σ_k and ρ_k are set to be Gamma(1, 1) and Gamma(1, 1), respectively. The diagonal elements in the mass matrix $\boldsymbol{M}$ are set to be 1.0, 1.5, 10.0 and 0.5 for corresponding intercept α_k, spatially varying coefficient $\boldsymbol{\zeta}_k$, marginal variance parameter σ_k and decay parameter ρ_k, respectively. The starting number of steps while using Leapfrog method is $L = 100$ and the step size is set at $\delta = 0.001$. Both L and δ are updated using the No U-turns algorithm (*Hoffman and Gelman*, 2014) and the targeted acceptance rate is restricted to be in the range of 0.55 to 0.75. In addition, while drawing σ_k and ρ_k in the HMC portion of the algorithm, σ_k is required to be positive and ρ_k is bounded between $[0.5, 25]$.

The mean posterior estimates of the spatially varying parameters $\boldsymbol{\beta}$ are displayed in Figure 2.3 (second row, left). The pattern is roughly consistent with the difference image in the same figure.

To determine prediction accuracy, we simulated data, as noted above, 10 different times. Each time we predicted accuracy, defined as whether we could correctly predict in which group each of the 5 test data sets belong. Overall, we correctly identified 49 out of the 50 (98%).

2.5 MS Lesion Data Analysis

We apply our model to a a data set that consists 228 MS patients classified into one of three clinical subtypes of MS. In increasing order of severity, the three clinical subtypes are Relapsing Remitting (RLRM, 172 patients), Primary Progressive (PRP, 13 patients) and Secondary Chronic Progressive (SCP, 43 patients). RLRM is used

as the baseline group in our model. We treat the subtypes as a multinomial variable since patients do not progress through the three subtypes, although some RLRM MS patients will convert to SCP.

We consider three patient specific covariates, including sex, age and disease duration (DD) (Table 2.1). The variables age and disease duration have been mean-zero centered. More female develop RLRM disease compared to other subtypes. The average mean centered ages of patients are -2.89, 2.13 and 10.87 for RLRM, PRP and SCP, respectively (true ages and disease duration are not available). This may indicate that MS subtype severity increases with age of patients. However, the range of age for RLRM patients is much wider than the other two groups and also covers the other two groups' ages ranges. Meanwhile, SCP patients have the longest disease duration on average, while PRP patients have the shortest disease duration on average. However, the distributions of disease duration do overlap one another.

All patients were scanned on 1.5T Siemens Avant scanner. We only consider the T_2-weighted images with a native voxel resolution of $0.997 \times 0.997 \times 3.000\text{mm}^3$. Lesions are identified on T_2-weighted images by a semi-automatic procedure so that each patient has a binary lesion map with 1 denoting the presence of a lesion and 0 the absence of a lesion at each voxel. The resulting maps are then affine registered to the Montreal Neurological Institute (MNI) brain template at $2 \times 2 \times 2\text{mm}^3$ resolution using trilinear interpolation, and thresholded at 0.5 to retain binary values (*Ge et al.*, 2014b). The 3D binary lesion maps ($91 \times 109 \times 91$) with voxel size $2 \times 2 \times 2\text{mm}^3$ contain a total of 274,596 in-mask voxels.

To reduce computational burden, we down sample the images to dimensions $64 \times 64 \times 64$. Then there are a total of 129,088 in-mask voxels. This image dimension reduction process is implemented by dividing the original images into larger equal-sized voxels and interpolating. The lesion status in the new voxels are set to be 1 if any lesion that appears in the combined neighbor voxels in the original images. Further

more, since MS is typically a white-matter disease, our analysis is constrained to a set of voxels with high-probability white matter ($> 50\%$). Finally, we have a total of 70,823 in-mask and in-white matter voxels out of 262,144 voxels in a single $64 \times 64 \times 64$ brain image. Our results show that the prediction accuracy and estimations are not affected by limiting to high-probability white matter voxels.

There are 38,620 (54.53%) voxels in which at least one patient (no matter MS subtype) has a lesion, out of the 70,823 voxels (Figure 2.2). With respect to MS subtypes, patients with RLRM disease have the greatest average lesion load (50.15%), while the average lesion load is 13.64% and 29.01% for subject with PRP and SCP disease.

Figure 2.4 shows an axial view (slice 33) of the empirical lesion probability maps for the three MS subtypes. Since only 13 patients in our data set are classified as PRP, their empirical lesion probability map has a less spatially extensive distribution than the other two groups. Overall, the shapes of areas with lesions in this axial slice are similar for the three groups. The highest empirical lesion probabilities are 3.26×10^{-1}, 3.85×10^{-1} and 4.88×10^{-1} for RLRM, PRP and SCP MS subtypes, respectively.

Given that patients with RLRM disease serve as the baseline group , Figure 2.5 (first row) shows the difference of empirical lesion probabilities for patients with PRP and SCP disease. The area with positive differences means more lesions exist in those voxels for PRP or SCP than RLRM disease.

Similar to the simulation study, we estimate the posterior distribution via MCMC using a total of 10,000 simulated draws and a burn-in of 5,000. The diagonal elements in the mass matrix $\boldsymbol{M}$ are set to be 1.0, 1.0, 1.5, 10.0 and 0.5 for corresponding intercept α_k, non-spatially varying coefficient $\boldsymbol{\gamma}_k$, spatially varying coefficient $\boldsymbol{\zeta}_k$, marginal variance parameter σ_k and decay parameter ρ_k, respectively. The augmented image size is $128 \times 128 \times 128$.

We include three non-spatially varying covariates so that $\boldsymbol{U}_k$ is the mean vector

with three elements and $\boldsymbol{\Sigma}_k$ is a 3×3 covariance matrix. Their prior distributions are set to be MVN$(\mathbf{0}, \boldsymbol{I}_3)$ and Inverse-Wishart$(3, \boldsymbol{I}_3)$, where $\boldsymbol{I}_3$ is an identity matrix. The prior for ρ_k is Gamma$(1, 1)$. In contrast, we propose a tighter prior distribution Gamma$(18, 6)$ to σ_k with prior mean 3 and smaller prior variance 0.5. Different prior means are placed on σ_k and it turns out that the large the mean of the prior for σ_k, the better the prediction performance. This is reasonable because there are many voxels that are not covered by a lesion in the full image space and their estimated spatially varying coefficients are around zero. Higher values of σ_k increase the scale of $\boldsymbol{\beta}_k$. This increases significant non-zero estimates of $\boldsymbol{\beta}_k$ in areas with many lesions for but not in regions without lesions. Areas with no lesions only introduce "noise" in the modelling.

We first include both binary lesion maps and patient specific covariates in our model. For non-spatially varying parameters $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, we show their posterior estimates in Table 2.3. The estimates indicate that none of the patient specific covariates have significant effects on the prediction of MS subtypes except the covariate age for the SCP subtype (posterior mean: 2.03, 95% credible interval: [0.22, 3.91]). Next, we exclude the three patient specific covariates and only use the binary lesion maps in our scalar-on-image regression model.

Maps of the mean posterior estimates of the spatially varying parameters $\boldsymbol{\beta}$ show some similar patterns to the difference of empirical lesion probabilities, see Figure 2.5. First, those mean posterior estimates of β's in voxels outside the high-probability white matter areas are near zero. Second, the estimated $\boldsymbol{\beta}_k$s are positive/negative (red/blue) in areas where the difference of empirical lesion probabilities are positive/negative (red/blue), particularly in areas with high absolute difference of empirical lesion probabilities.

Prediction results, computed by using Importance-Sampling Leave-One-Out cross validation, are shown in Table 2.4. For Table 2.4(a), the results are from a single

Markov chain of the analysis using binary lesion maps as the only predictors and there are only 6 prediction errors out of the total of 228 patients (prediction accuracy is 97.3%). In contrast, if weinclude patient specific covariates as predictors, see Table 2.4(c), the prediction accuracy is reduced to 82.9% with 39 errors. In particular, 28 RLRM patients are predicted as SCP patients. If we remove the binary lesion maps and only use the three patient-specific covariates, the predictions from a multinomial regression model are displayed in Table 2.4(b) with 47 prediction errors and 79.39% prediction accuracy and none of the PRP patiens are correctly prediction correctly. These results indicate that the inclusion of sex, age, disease duration are not strong predictors and introduce noise in the prediction of MS subtypes.

2.6 Discussion

In this chapter we propose a Bayesian scalar-on-image regression model with application to MS MRI data. Both non-spatially and spatially varying variables can be used as predictors in the model. Spatial correlation is modelled using a GP prior distribution and estimated under the Bayesian framework by using the HMC algorithm. In addition, using the connection between symmetric nested block circulant matrices and the 3D discrete Fourier transform we can take advantage of the efficiency of the FFT algorithm and parallelism make to invert the extremely large correlation matrix. Lastly, we use importance-sampling leave-one-out cross validation to evaluate prediction performance for the MS data set.

Compared with other methods used for modelling scalar outcomes based on brain images, our scalar-on-image regression model fully considers and estimates the spatially correlations by assuming a flexible GP prior distribution for spatially varying parameters. Then, information within the entire brain is used for prediction without resorting to region-of-interest analysis or sparsity assumptions. In addition, our Bayesian estimation framework, adopting HMC and FFT algorithms, makes it fea-

sible and efficient to deal with the high-dimensional imaging data with complicated correlation structure. The simulation and application demonstrate that our method can result high prediction accuracy, especially for the MS data set we analyzed. Finally, the use of a GPU significantly reduces the computing time from at least 45 days to less than 24 hours for a single run of 10,000 iterations (at least 45 times faster).

One of the drawbacks of our method is the limitation of the spatial structure of the covariance matrix for the spatially varying parameters. The covariance matrix must have a nested block Toeplitz struction so that it can be embedded in a symmetric nested block circulant matrix so that we may take advantage of the relationship between the eigen decomposition of these matrices and the discrete Fourier transform. We greatly increase the computational efficiency of the algorithm at the expense of a restricted correlation structure, which may not be correct.

Another area of concern is the use of importance-sampling leave-one-out cross validation. Exact cross-validation requires re-fitting the model each time a subject is removed for prediction. This would be prohibitively time consuming. Approximate leave-one-out cross-validation easily calculated using an importance sampling algorithm. However, the approximation results may be biased and noisy if the variance of importance weights is very large (*Vehtari et al.*, 2016). We conducted exact leave-one-out cross validation to evaluate our model's accuracy in our simulation study and then compared with the accuracy measured by importance-sampling leave-one-out cross validation, both of which resulted in perfect prediction accuracy.

Table 2.1: Descriptive statistics of patient specific covariates

Covariates		RLRM MS	PRP MS	SCP MS
Sex[a,c]	Female	134(78%)	6(46%)	23(53%)
	Male	38(22%)	7(54%)	20(47%)
Age[b,c]		-2.89[-25.48, 21.02]	2.13[-10.48, 13.62]	10.87[-4.98, 22.82]
Disease Duration[b,c]		-1.12[-12.92, 27.08]	-4.61[-10.92, 14.08]	6.54[-9.92, 34.08]

[a] Sex: number of patients (percentage).
[b] Age/Disease duration: mean(range).
[c] All patient-specific covariates are mean-zero centered

Table 2.2: Descriptive statistics of MS MRI data restricted to the high-probability white matter region.

Subtype	Sample Size	Empirical Lesion Probability				Voxels (at least one patient has lesion)
		Mean	Range	Median	Mode	NO.(%)
Overall	227	1.85×10^{-2}	$[0.00, 4.04\times10^{-1}]$	4.39×10^{-3}	0.00	38,620 (54.53%)
RLRM	172	1.80×10^{-2}	$[0.00, 3.90\times10^{-1}]$	5.8×10^{-3}	0.00	35,520 (50.15%)
PRP	13	1.57×10^{-2}	$[0.00, 5.39\times10^{-1}]$	0.00	0.00	9,661 (13.64%)
SCP	43	2.13×10^{-2}	$[0.00, 6.28\times10^{-1}]$	0.00	0.00	20,549 (29.01%)

Table 2.3: Posterior estimates of non-spatially varying parameters $\boldsymbol{\beta}_k$ from MS data set: mean, standard deviation (SD), 95% confidence interval (CI)

	PRP MS			SCP MS		
Covariates	Posterior Mean	Posterior SD	95% CI	Posterior Mean	Posterior SD	95% CI
Sex	0.07	1.27	[-2.43, 2.56]	0.31	1.50	[-2.63, 3.24]
Age	0.91	0.97	[-0.99, 2.81]	2.03	0.94	[0.22, 3.91]
Disease Duration	0.54	1.03	[-1.49, 2.57]	1.35	1.00	[-0.62, 3.31]

Table 2.4: Predictions of MS subtypes (a) using lesion maps only (b)using patient-specific covariates only (c) using both lesion maps and patient specific covariates

(a)

		True subtypes			
		RLRM	PRP	SCP	Total
Predicted Subtypes	RLRM	167(97.1%)	0	1	168
	PRP	1	13(100%)	0	14
	SCP	4	0	42(97.7%)	46
	Total	172	13	43	228

(b)

		True subtypes			
		RLRM	PRP	SCP	Total
Predicted Subtypes	RLRM	161(93.6%)	13	23	197
	PRP	0	0(0.0%)	0	0
	SCP	11	0	20(46.5%)	31
	Total	172	13	43	228

(c)

		True subtypes			
		RLRM	PRP	SCP	Total
Predicted Subtypes	RLRM	141(90.0%)	1	3	145
	PRP	3	11(84.6%)	3	17
	SCP	28	1	37(86.0%)	66
	Total	172	13	43	228

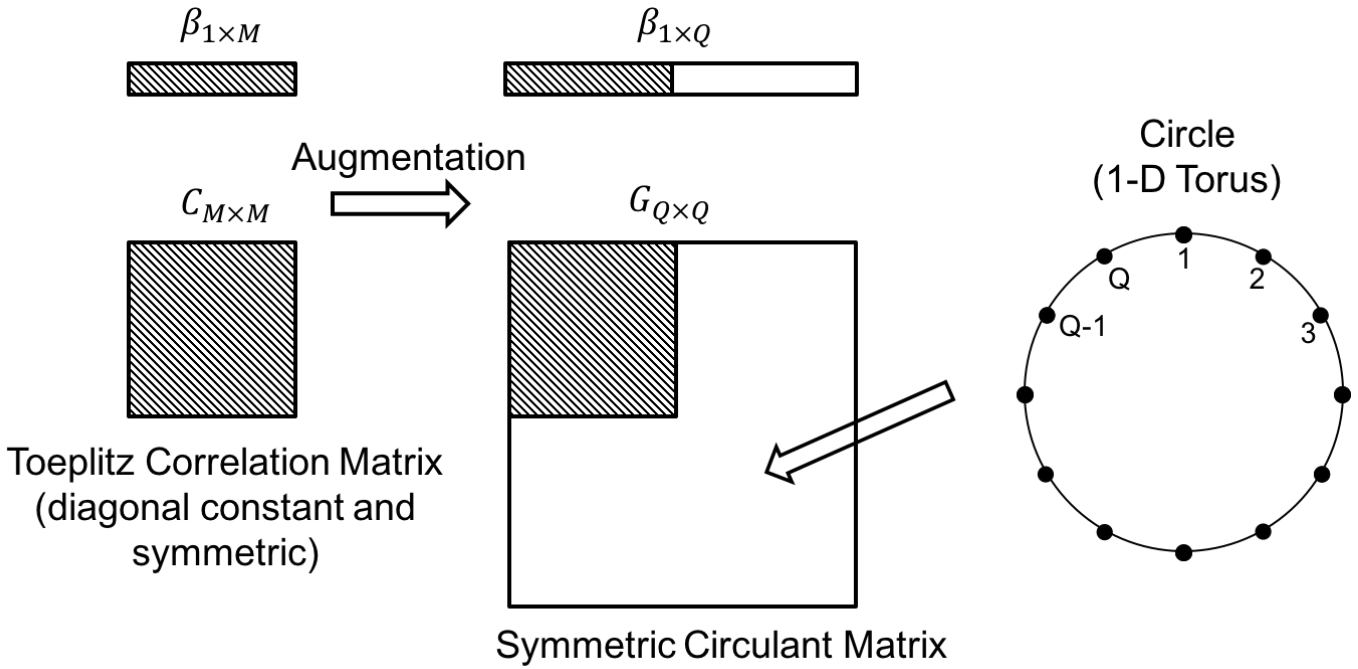


Figure 2.1: Example of the embedding process for a one-dimensional Gaussian Process $\boldsymbol{\beta}$ and its correlation matrix $\boldsymbol{C}$. The M-length vector $\boldsymbol{\beta}$ has a $M \times M$ Toeplitz correlation matrix $\boldsymbol{C}$. The augmented Q-length vector $\tilde{\boldsymbol{\beta}}$ has a $Q \times Q$ correlation matrix $\boldsymbol{G}$, which is symmetric and circulant. The smaller correlation matrix $\boldsymbol{C}$ is placed on the left upper corner of $\boldsymbol{G}$. By wrapping $\boldsymbol{G}$ on a circle (1D torus), the correlation values are determined by the minimum distance between any two points s and s' on the circle.

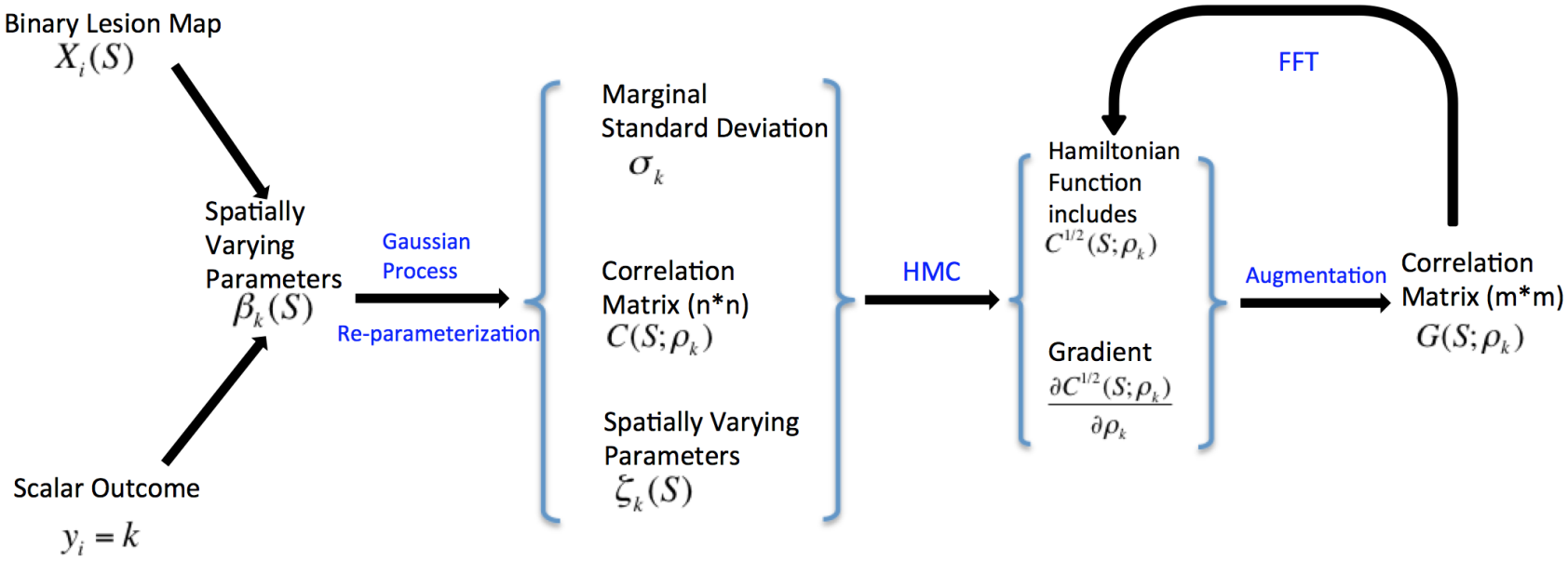


Figure 2.2: Schematic view of the scalar-on-image regression model.

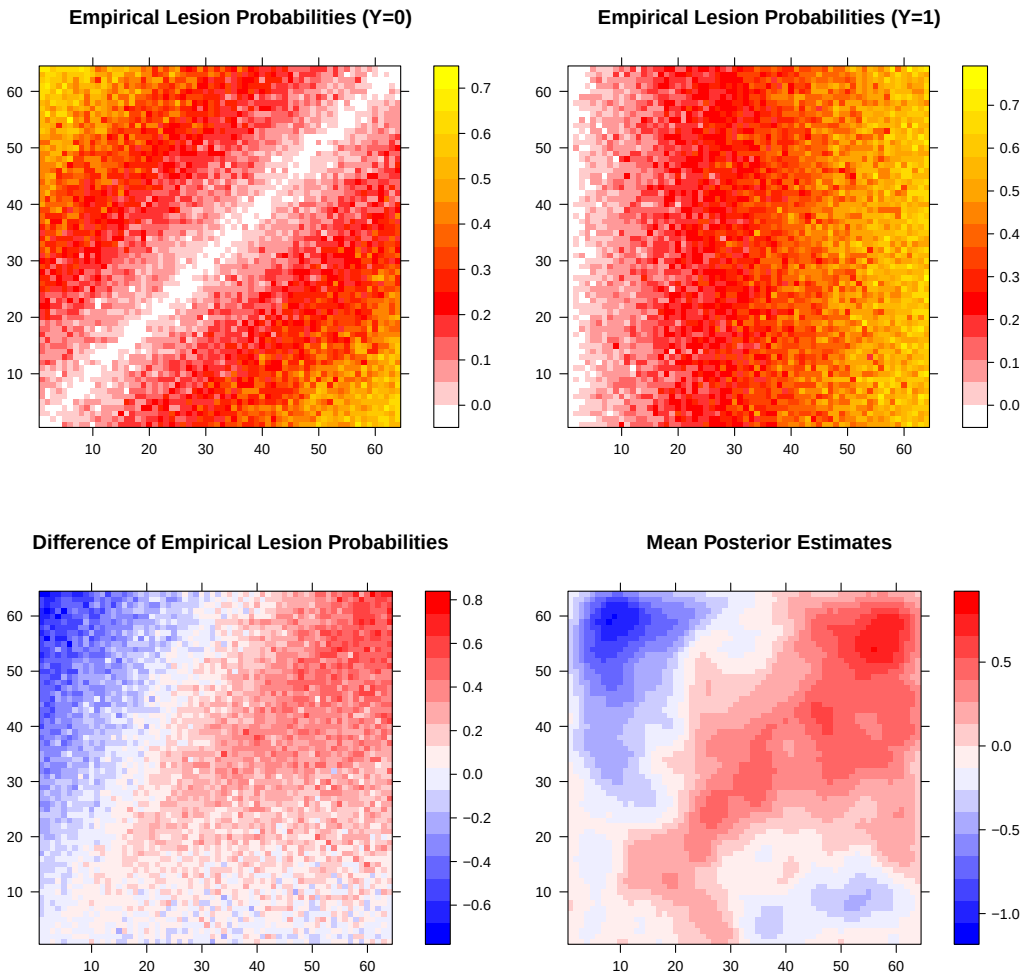


Figure 2.3: Maps of simulated data and estimates in the simulation study. (1) Maps of empirical lesion probabilities of training images in simulated group with outcome label $Y = 0$ (first row, left) and group with $Y = 1$ (first row, right). (2) Difference of empirical lesion probabilities for training images with label $Y = 1$ to those with label $Y = 0$ (second row, left). (3) Mean posterior estimates of spatially varying parameters $\boldsymbol{\beta}$ (second row, right).

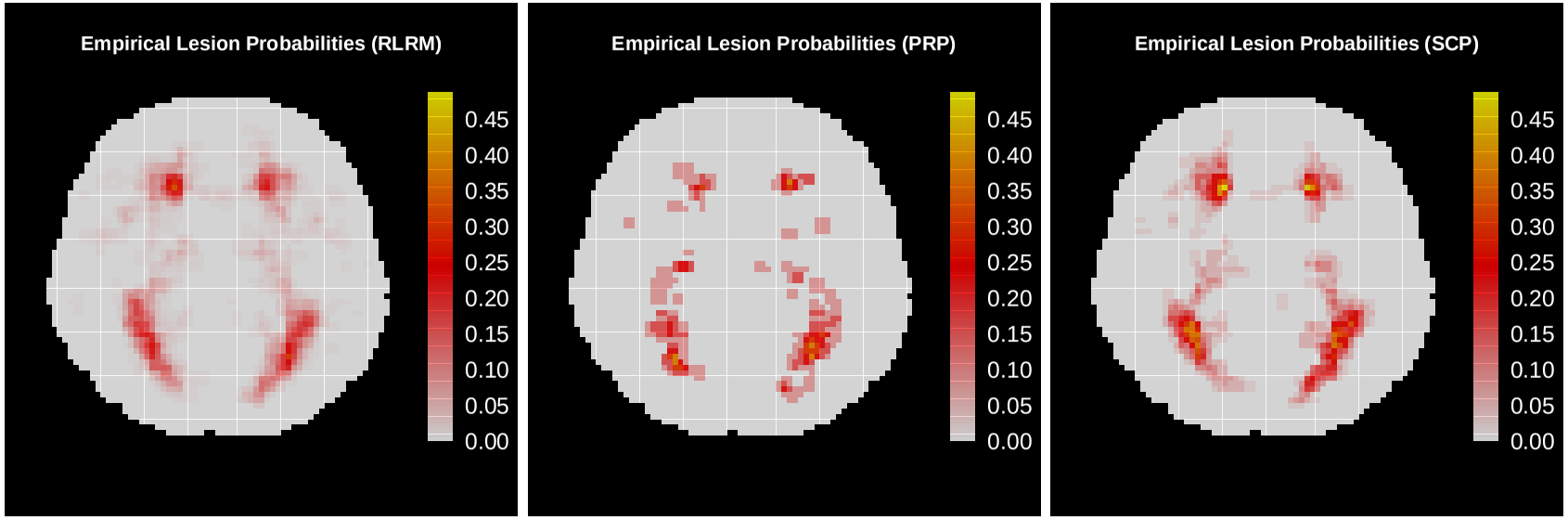


Figure 2.4: Maps of Empirical Lesion Probabilities for patients from three MS subtypes. (Left)RLRM (N=172); (Middle) PRP (N=13); (Right) SCP (N=43).

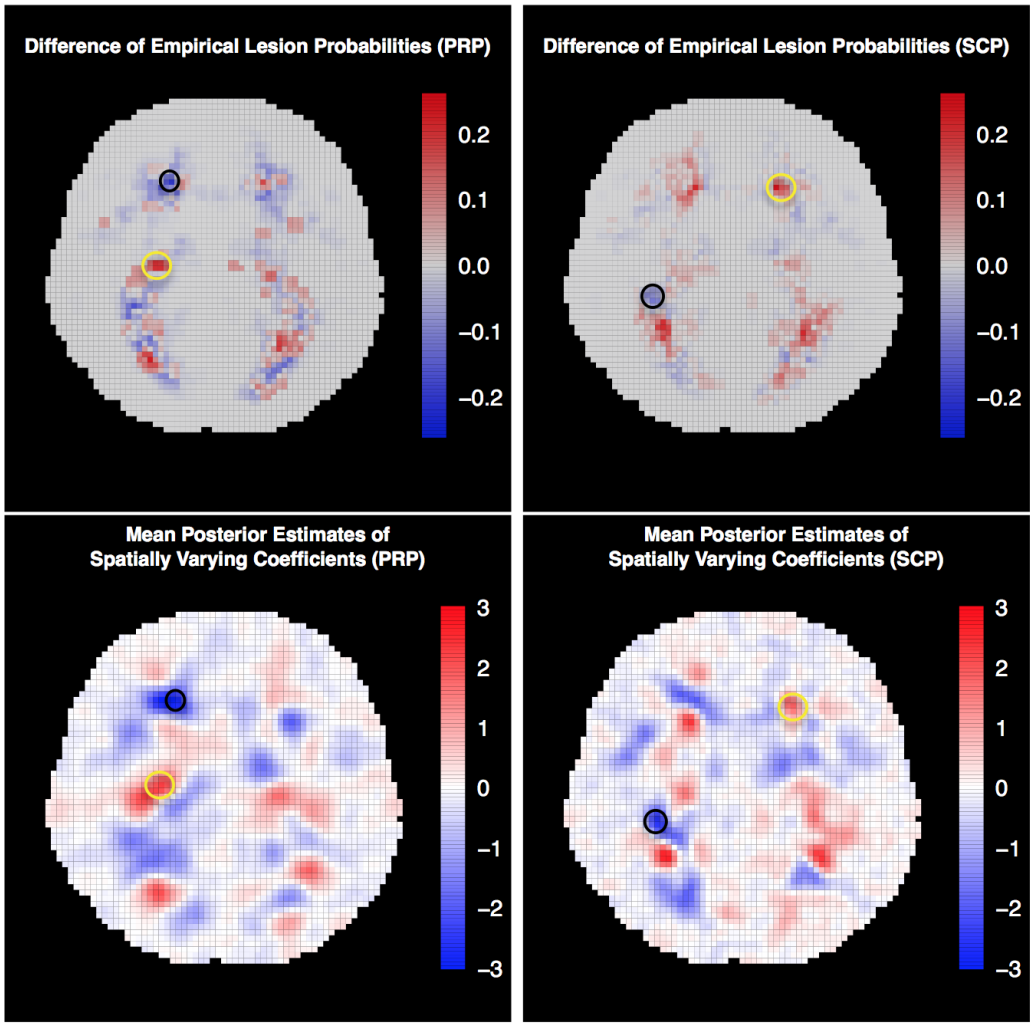


Figure 2.5: Axial view maps for PRP (left) and SCP (right) MS patients, respectively. (First row) The difference of empirical lesion probabilities by subtracting empirical lesion probabilities from RLRM patients. (Second row) mean posterior estimates of spatially varying parameters $\boldsymbol{\beta}_k$.

CHAPTER III

A Spatial Bayesian Latent Factor Model for Image-on-Image Regression

Image-on-image regression analysis, using images to predict images, is a challenging task, due to 1) the high dimensionality and 2) the complex spatial dependence structures in image predictors and image outcomes. In this Chapter, we propose a novel image-on-image regression model, by extending a spatial Bayesian latent factor model to image data, where low-dimensional latent factors are adopted to make connections between high-dimensional image outcomes and image predictors. We assign Gaussian process priors to the spatially varying regression coefficients in the model, which can well capture the complex spatial dependence among image outcomes as well as that among the image predictors. We perform simulation studies to evaluate the out-of-sample prediction performance of our method compared with linear regression and voxel-wise regression methods for different scenarios. We apply the proposed method to analysis of multimodal image data in the Human Connectome Project where we predict task-related contrast maps using sub-cortical volumetric seed maps. The proposed method achieves a better prediction accuracy by effectively accounting for the spatial dependence and efficiently reduce image dimensions with latent factors.

3.1 Introduction

Image-related regression analysis has attracted an increasing scientific interest in many areas, including medicine (*Duff et al.*, 2015), disease diagnosis (*Suk et al.*, 2017) and neuroscience (*Bowman*, 2014; *Zhou et al.*, 2013). In these applications, researchers are often interested in identifying the association between the quantitative images, e.g., functional magnetic resonance imaging (fMRI) (*Glover*, 2011), and other variables of interest, e.g. the patient clinical characteristics, where the quantitative images are either considered as the outcome variables or predictors. Two types of image-related regression models have been extensively studied: scalar-on-image regression which uses images to predict scalar outcome (*Reiss and Ogden*, 2010b; *Goldsmith et al.*, 2011; *Huang et al.*, 2013; *Wang et al.*, 2017; *Kang et al.*, 2018), and image-on-scalar regression which uses a set of scalar predictors to predict the image outcome (*Gelfand et al.*, 2003; *Reiss et al.*, 2010; *Goldsmith and Kitago*, 2016; *Chen et al.*, 2016; *Yan and Liu*, 2017). Recently there are increasing interests in developing regression models where both outcomes and predictors are multiple images, to which we refer as the image-on-image regression. It has many important applications in neuroimaging studies, for instance, *Tavor et al.* (2016) showed that structural and resting-state fMRI images could predict task-based fMRI images using a large set of task conditions and spanning several behavioral domains. They trained a simple linear regression model to predict imaging outcome using many feature images (predictor). After parcellation, they fitted their model one parcel and one subject at a time, assuming independence and identical linear relationship across voxels in each parcel. Then they predicted outcomes for unseen subjects by averaging fitted models across subjects. Although their linear method is simple to implement for image-on-image regression analysis, it ignores the spatial dependence structures in images by assuming voxels are mutually independent. Meanwhile, they fitted their model on the individual level but predicted for new subjects using averaged estimations, which

conflicts their assumption of individual variations in their imaging outcomes. In addition, this model fails to select relevant predictors while giving an amount of correlated task-independent images.

The spatial correlations and associations can be hard to model due to its heterogeneity, complexity and high-dimensionality with limited sample size. For example, the spatial correlations between neural activity at different voxels might extend beyond neighbouring voxels, and may not decrease with increasing distance (*Bowman*, 2014). Also, the spatial patterns may vary across different types of brain images and even subjects. Moreover, the effects of features on outcome images may be from the whole image space but not voxel by voxel. Hence, in order to precisely describe the spatial patterns and associations of outcome and predictor images, one needs to define voxel-wise parameters in image space. Those parameters are high-dimensional and have complex spatial correlations, leading to the potentially over-parameterized regression model and computational challenges.

In this work, we propose a novel spatial Bayesian latent factor model for the image-on-image regression. We introduce low-dimensional latent factors to link high-dimensional outcome images and predictor images. In particular, for each subject, we represent the expectation of the outcome image as a linear combination of multiple spatial basis functions, where each basis function is associated with one spatial knot. The knot-dependent coefficients are decomposed into several latent factors. Each latent factor summarizes one feature of the outcome image. We model each latent factor as a scalar outcome and assume its expectation is equal to the summation of linear predictors. Each linear predictor is a linear transformation of one predictor image, where the corresponding spatially-varying coefficients represent the spatial effects of the predictor image. We assign Gaussian process (GP) priors to the spatially-varying coefficients and adopt the basis expansion approach for GP representations, which are sufficient, flexible and convenient to preserve the complex spatial correlations. With

an appropriate choice of the basis function, the model can well capture the complex spatial patterns of the outcome image using a smaller number of spatial knots. Thus, our basis expansion approach may effectively reduce the model dimensionality. More importantly, by integrating the whole feature image space, our model can well predict the outcome images borrowing the strength not only from the same voxels in the predictor image but also neighbouring and even long-distance voxels. A small number of spatial latent factors in our model can effectively capture the association between the predictor images and the outcome image. Thus, our model can make more accurate prediction compared to the linear model (*Tavor et al.*, 2016) with a small sample size.

A few image-on-image regression models have been proposed recently motivated by various applications in medical imaging. For example, *Sweeney et al.* (2013a) applied the voxel-wise logistic regression models incorporating Multiple Sclerosis imaging sequences to predict lesion incidence with T1-weighted, T2-weighted, FLAIR, and PD volumes from a longitudinal study. The voxel-wise regression method evaluates the population effects and is simple to implement. However, it ignores the spatial correlations among voxels and thus may lose power to detect the association between the predictor images and the outcome images. Another voxel-wise regression model proposed by *Hazra et al.* (2017) includes prediction effects from the neighboring voxels within a given Euclidean distance. The prediction effects are assumed to be the same when voxels in the predictor image have the same distance to the voxel in the outcome image. In the outcome images, voxels are assumed to be independent over space. Their spatial association with the predictor images is restricted to small regions and only related to the spatial distance. In contrast to the above two voxel-wise regression methods, our spatial latent factor model may capture more complex spatial dependence between outcome image and the predictor image. Deep learning has been applied to medical imaging for image-on-image regression analysis, such as the image recovering and disease diagnosis (*Zhu et al.*, 2017; *Isola et al.*, 2017; *Huang et al.*,

2018). However, the performance of deep learning methods relies on very sample size, which we typically do not have in medical imaging studies; and it is difficult to interpret the specific associations among images using based on a deep neural network model.

We organize this chapter as follows. We first describe our spatial Bayesian latent factor models in Section 2. In Section 3, we present the proposed Bayesian framework for estimation and prediction. In Section 4, we conduct a simulation study under different scenarios and discuss some criteria used for model evaluation and selection. Next, we illustrate the proposed method using the fMRI data from the Human Connectome Project database. We close with a discussion of future work.

3.2 Model

In this section, we present our spatial Bayesian latent factor model for image-on-image regression analysis. We extend the classical Bayesian latent factor model for functional and longitudinal data (*Montagna et al.*, 2012) to the case where the functional predictors are images with complex spatial dependence. Our goal of statistical modelling is fundamentally different from the one by *Montagna et al.* (2018) which focused on the meta-analysis of functional neuroimaging data.

Suppose the data consists of one outcome image and P predictor images from n subjects. For all subjects, we assume that both outcome and predictor images have been preprocessed and registered to the same brain region, denoted $\mathcal{R}$. Note that in practice $\mathcal{R}$ may refer to the whole brain region or one sub region of interest; and the following model and parameter settings are region-specific. For each subject $i(i = 1, \ldots, n)$ at voxel $v \in \mathcal{R}$, let $Z_i(v)$ and $X_{ip}(v)(p = 1, \ldots, P)$ represent the outcome image intensity and the pth predictor image intensity, respectively. To identify the association between $Z_i(v)$ and $X_{ip}(v)$ on brain region $\mathcal{R}$, we develop a spatial Bayesian latent factor model with three levels of hierarchy.

3.2.1 Level 1: Approximation of Outcome Images

At Level 1, we approximate the outcome image using a basis expansion approach. Let $\{b_m(v)\}_{m=1}^M$ be a set of M spatial knot-dependent basis functions that can well capture the variation of the outcome image in $\mathcal{R}$. For each subject i and any voxel $v \in \mathcal{R}$, we assume

$$Z_i(v) = U(v) + \sum_{m=1}^{M} \theta_{im} b_m(v) + e_i(v), \quad U(v) \sim \mathrm{N}(0, \sigma_u^2), \quad e_i(v) \sim \mathrm{N}(0, \sigma_e^2),$$

where $U(v)$ represents the population-level spatial-dependent intercept. As a prior specification, we assume $\{U(v)\}_{v\in\mathcal{R}}$ are independent and identically distributed as a normal distribution with mean zero and variance σ_u^2. The random errors $e_i(v)$ are assumed to be independent and identically distributed as a normal distribution with mean zero and variance σ_e^2 over all subjects across voxels in region $\mathcal{R}$. The parameter θ_{im} is the subject-specific basis coefficient of the mth spatial basis function $b_m(v)$. The term $\sum_{m=1}^M \theta_{im} b_{im}(v)$ captures the spatial dependence and smoothness of the subject-specific outcome images among voxels in parcel l.

3.2.2 Level 2: Sparse latent factor model for basis coefficients

At Level 2, we build a sparse latent factor model for the basis coefficient θ_{im}:

$$\theta_{im} = \sum_{k=1}^{K} \lambda_{mk} \eta_{ik} + \zeta_{im}, \qquad \zeta_{im} \sim \mathrm{N}(0, \sigma_\zeta^2),$$

where $\{\eta_{ik}\}_{k=1}^K$ represents a set of K latent factors for subject i and $\{\lambda_{mk}\}_{k=1}^K$ are the corresponding sparse loading coefficients, indicating the effects of the kth latent factors on the mth basis coefficient. The random error ζ_{im} explains the variation of the basis coefficient θ_{im} that cannot be explained by the latent factors. For the prior specification for the sparse loading coefficients λ_{mk}, we resort to inducing prior

distribution through the parameter-expansion (*Ghosh and Dunson*, 2009), leading to more efficient posterior computation. See more details in Section 3.

3.2.3 Level 3: Link to predictor images

At Level 3, we link latent factors to predictor images via a scalar-on-image regression:

$$\eta_{ik} = \sum_{v' \in \mathcal{R}} \widetilde{X}_i(v')\beta_k(v') + \epsilon_{ik}, \quad \widetilde{X}_i(v') = \sum_{p=1}^{P} \gamma_p X_{ip}(v'), \quad \epsilon_{ik} \sim \mathrm{N}(0, \sigma_\epsilon^2),$$

where the error term ϵ_{ik} follows a normal distribution with mean zero and variance σ_ϵ^2. The unit variance assumption on η_{ik} ensures latent factors are identifiable in the model. In brain imaging applications, we expect the effect of the predictor image from a voxel v' on the outcome image at voxel v are generally weak if not zero but similar across different predictors. Thus, to effectively reduce the dimension of parameter space, we consider a summarized predictor image $\widetilde{X}_i(v')$ as the average of selected predictor images from $\{X_{ip}(v')\}_{p=1}^{P}$. The latent selection indicator γ_p is assumed to follow a Bernoulli distribution with prior probability w, while π_p may include the prior knowledge on the proportion of important predictor images. To account for spatial dependence in predictors, we assign a Gaussian process (GP) prior to spatially-varying coefficient $\beta_k(v)$ and approximate it using a basis expansion approach: i.e. $\beta_k(v) = \sum_{m=1}^{M} \alpha_{km} b_m(v)$ with $\alpha_{km} \sim \mathrm{N}(0, \sigma_\alpha^2)$. The spatially-varying coefficient $\beta_k(v)$ is factor-specific and shared by all subjects. Typically, a small number of latent factors are needed to capture important feature information from the selected image predictors at the population level, contributing to predicting the outcome image.

3.2.4 Model Representation

Our proposed Bayesian hierarchical model has an equivalent model representation by integrating out θ_{im} and η_{ik} (see Appendix B.1). Specifically, we have

$$Z_i(v) = U(v) + f_i(v) + e_i(v),$$
$$f_i(v) = \sum_{p=1}^{P} \gamma_p \sum_{v' \in \mathcal{R}} \psi(v, v') X_{ip}(v') + \tilde{\epsilon}_i(v) + \tilde{\zeta}_i(v),$$

where $\tilde{\epsilon}_i(v) = \sum_{m=1}^{M} \sum_{k=1}^{K} \lambda_{mk} \epsilon_{ik} b_m(v)$, $\tilde{\zeta}_i(v) = \sum_{m=1}^{M} \zeta_{im} b_m(v)$ and

$$\psi(v, v') = \sum_{k=1}^{K} \left\{ \Big[\sum_{m=1}^{M} \lambda_{mk} b_m(v) \Big] \times \Big[\sum_{m'=1}^{M} \alpha_{km'} b_{m'}(v') \Big] \right\}. \tag{3.1}$$

From this representation, the outcome image $Z_i(v)$ has the subject-level component $f_i(v)$ which links to the predictor image $X_{ip}(v')$ using the spatially dependent weights $\psi(v, v')$ and predictor selection indicator γ_p. In particular, $\psi(v, v')$ represents the average change in the outcome image at voxel v per unit change in the value of any selected predictor image at voxel v'. Furthermore, equation (3.1) shows that this spatially varying prediction effect can be decomposed as the summation of K tensor products of spatially varying coefficients. This representation enables our model to retain complex spatial dependence structures in the outcome and predictors, respectively. Hence, our model is flexible to borrow strengths from the whole brain region to predict the outcome image at each voxel.

3.3 Posterior Computation

We resort to Markov chain Monte Carlo (MCMC) algorithms for posterior computation. For latent factor models, the performance of posterior computation may depend on prior specifications. In general, we can assign normal and inverse-gamma

prior distributions to factor loadings and residual variances respectively. Although those prior distributions produce conditionally conjugate posterior distributions and lead to straightforward computation by Gibbs sampler, such routine Bayesian implementations is poorly behaved (*Ghosh and Dunson*, 2009). To achieve efficient posterior computation for our model, we extend the parameter expansion (PX) method proposed by *Ghosh and Dunson* (2009). We construct a hierarchical model for the latent factors with different covariance structures.

3.3.1 Prior Specifications via Parameter Expansion

Our model needs additional constraints to ensure the K latent factors identifiable. Write $\boldsymbol{\theta}_i = (\theta_{im})_{M\times 1}$, $\boldsymbol{\Lambda} = (\lambda_{mk})_{M\times K}$, $\boldsymbol{\eta}_i = (\eta_{ik})_{K\times 1}$, $\boldsymbol{\zeta}_i = (\zeta_{im})_{M\times 1}$, $\tilde{\boldsymbol{X}}_i = \{\tilde{X}_i(v)\}_{|\mathcal{R}|\times 1}$, $\boldsymbol{\beta}_k = \{\beta_k(v)\}_{|\mathcal{R}|\times 1}$, $\boldsymbol{\beta} = \{\beta_k(v)\}_{|\mathcal{R}|\times K}$, $\boldsymbol{\epsilon}_i = (\epsilon_{ik})_{K\times 1}$, $\boldsymbol{\alpha}_k = (\alpha_{km})_{M\times 1}$ $\boldsymbol{\alpha} = (\alpha_{km})_{K\times M}$ and $\boldsymbol{b} = \{b_k(v)\}_{|\mathcal{R}|\times K}$, where $|\mathcal{R}|$ represents the number of voxels in $\mathcal{R}$. The matrix representations of the original inferential models in Levels 2 and 3 are shown in Table 3.1.

Following the PX approach, we develop the working models and the corresponding transformations between inferential and working parameters as shown in Table 3.1. We introduce parameters $\boldsymbol{\Phi} = \text{diag}\{\phi_1^2, \ldots, \phi_K^2\}$ and the sign function $S(x) = 1$ if $x \geq 1$ and -1 otherwise. An extra working intercept term $\boldsymbol{\mu}_i^*$ is included for more efficiently estimating working the latent factor $\boldsymbol{\eta}_i^* = (\eta_{ik}^*)$. The working factor loading $\boldsymbol{\Lambda}^* = (\lambda_{mk}^*)_{M\times K}$ is a lower triangular matrix without constraints on the elements. Instead of specifying a prior distribution for $\boldsymbol{\Lambda}$ directly, we induce a prior distribution for $\boldsymbol{\Lambda}^*$ and then transforms it to the prior distribution for $\boldsymbol{\Lambda}$. Specifically, those prior

distributions placed for working parameters are

$$\lambda^*_{mk} \sim \mathrm{N}(0, \sigma^2_\lambda), \quad m = 1, \cdots, M; k = 1, \cdots, \min(m, K)$$

$$\lambda^*_{mk} \sim \delta_0, \quad m = 1, \cdots, M; k = \min(m, K) + 1, \cdots K$$

$$\phi^2_k \sim \mathrm{Gamma}(a_\phi, b_\phi), \quad k = 1, \cdots, K$$

$$\boldsymbol{\mu}^*_i \sim \mathrm{N}_K(\mathbf{0}, \sigma^2_\mu \boldsymbol{I}_K), \quad i = 1, \cdots, N$$

$$\boldsymbol{\alpha}^*_k \sim \mathrm{N}_M(\mathbf{0}, \sigma^2_\alpha \boldsymbol{I}_M), \quad k = 1, \cdots, K$$

where δ_0 is a measure concentrated at zero. Hyperparameters $\sigma^2_\lambda, \sigma^2_\mu, \sigma^2_\alpha, a_\phi, b_\phi$ can be prespecified.

According to the above working model representation and prior specifications, we develop an efficient Gibbs sampler for posterior computation (see Appendix B.2). Of note, the PX approach leads to an over-parametrized working model and thus the posterior computation may exhibit a poor mixing due to the lack of identifiability (*Ghosh and Dunson*, 2009) for the working parameters. The parameters in the original inferential model are still identifiable and the Markov chains in the posterior computation usually show a much better mixing.

Gibbs samplers cycle through the simple steps to generate the MCMC samples of the parameters in the working models, on which after burn-in we take the corresponding transformations (see Table 3.1) to obtain posterior samples for the parameters in the original inferential model. By collecting those MCMC based posterior samples, we can make posterior inferences on any functions of the parameters in the original model, and we also can make predictions on the image outcomes based on their posterior predictive distributions. Details of estimation and prediction method are in Appendix B.3.

3.3.2 Basis Functions and Number of Latent Factors

It is challenging to choose the number of basis functions. In each parcel, the more basis functions are included in the model, the richer the spatially varying patterns of the outcome image the model can capture. On the other hand, to reduce computational costs, the bases should be locally concentrated and the number of basis functions should be much smaller than the number of voxels in the parcel. In addition, the appropriate basis functions are unknown in advance. Conceptually, any basis functions, like B-spline bases and Gaussian kernels, can be chosen for the smooth images. Here, we use a 3D isotropic Gaussian kernel.

$$b_m(v) = \exp\{-b\|v - \psi_m\|^2\}, \quad v \in \mathcal{R}, \;\; m = 1, 2, \ldots, M,$$

with kernel locations $\{\psi_m\}_{m=1}^M$ and parameter b controlling smoothness. Flexible approaches are available for estimating M, b and $\{\psi_m\}_{m=1}^M$ for basis functions. One approach is to perform the fully Bayesian inferences using the MCMC algorithm with appropriate prior specifications for those parameters. However, this approach suffers very large computational burden as the basis function has to be re-evaluated in each iteration of the MCMC algorithm. Hereafter, we adopt a relatively less computational-intense approach. We first choose a reasonable number of bases M as well as kernel locations $\{\psi_m\}_{m=1}^M$, and then we determine the smooth parameter b by minimizing the mean squared error (MSE) and mean squared prediction error (MSPE) of outcome images via cross validation (CV). The metrics MSE and MSPE are averaged over datasets, observations and voxels. Suppose we consider N^{folds} cross validation. For the jth fold, let $\mathcal{I}_j^{\text{ts}}$ and $\mathcal{I}_j^{\text{tr}}$ represent the indices of the subjects in the test set and training set, respectively. Let $\hat{Z}_i(v)$ for $i \in \mathcal{I}_j^{\text{tr}}$ and $\hat{Z}_k(v)$ for $k \in \mathcal{I}_j^{\text{ts}}$ represent the fitted and predicted outcome image at voxel v, respectively. We define

MSE and MSPE as

$$\text{MSE} = \frac{1}{N^{\text{folds}}} \times \frac{1}{N^{\text{tr}}} \times \frac{1}{N^{\text{voxel}}} \times \sum_{j=1}^{N^{\text{folds}}} \sum_{i \in \mathcal{I}_j^{\text{tr}}} \sum_{v \in \mathcal{R}} \left\{ Z_i(v) - \hat{Z}_i(v) \right\}^2$$

$$\text{MSPE} = \frac{1}{N^{\text{folds}}} \times \frac{1}{N^{\text{ts}}} \times \frac{1}{N^{\text{voxel}}} \times \sum_{j=1}^{N^{\text{folds}}} \sum_{k \in \mathcal{I}_j^{\text{ts}}} \sum_{v \in \mathcal{R}} \left\{ Z_k(v) - \hat{Z}_k(v) \right\}^2$$

where N^{folds}, N^{tr}, N^{ts} and N^{voxel} represent the number of folds, sample size of training set, sample size of the test set and the number of voxels used in the CV study, respectively.

To select the number of latent factors, a set of widely used model comparison criteria can be considered, such as deviance information criteria (DIC), Bayesian information criteria (BIC), Bayes factors (BF) and R-squared. One can fit the model multiple times with different values of K and choose the optimal value based on the above criteria. This approach is computationally intensive. We consider an alternative approach in light of latent factors. In our model, redundant latent factors may have very sparse, zero-concentrated or similar loading vectors. Hence, the optimal number of latent factors has a loading matrix with non zero-concentrated and distinguishable loading vectors. We extend the "elbow method" for clustering analysis to determine the optimal value of K. First, we fit our model with a relative large number K, for example $K = 20$, and then describe the distributions of each estimated loading vectors using varied metrics, including 1) range, 2) maximum absolute value, 3) standard deviation, 4) the number of values in each loading vector outside the 95%, 90% and 68% credible interval (CI) of the whole loading matrix. By plotting the sorted summary statistics (descending) to corresponding latent factors, the optimal choice of K is the value which give an angle in the graph. Finally, we need to refit the model with the choice of K for estimations and predictions.

3.4 Simulation Study

3.4.1 Data Generation and Method

In this section, we conduct simulation studies to compare the performance of our proposed method with another two methods, the linear regression model (*Tavor et al.*, 2016) and the voxel-wise regression. The three methods serve as the generating models in three different scenarios, respectively. In each scenario, we generate 10 data sets, each of which contains 100 observations as the training set and another 50 observations as the test set. Each simulated observation has a two-dimensional outcome image and a set of 20 two-dimensional predictor images on equally spaced grid points on $\{1, \ldots, 32\} \times \{1, \ldots, 32\}$. In the simulation study, we treat the whole image space as a single parcel. Specifically, we generate predictor images from a Gaussian Process with mean zero and a covariance function that $c(v_1, v_2) = 0.01 \exp\{-15 \times d^2_{v_1,v_2}\}$, where d_{v_1,v_2} is the Euclidean distance between any two grid points v_1 and v_2.

Scenario 1. We simulate outcome images from the following regression model

$$Z_i(v) = \beta_{i0} + \sum_{p=1}^{20} \beta_{ip} X_{ip}(v) + \epsilon_i(v), \quad \text{for} \quad i = 1, \ldots, 150, \tag{3.2}$$

where the linear coefficients β_{ip} are the same over the space but varied among observations. The error terms $\epsilon_i(v)$ are independently sampled from a normal distribution with mean zero and variance 0.1. The true linear coefficient β_{ip} is randomly sampled from a Normal distribution with mean μ_p and variance σ_p^2, where σ_p^2 is drawn from a gamma distribution with shape 0.1 and rate 0.1. The mean μ_p is sampled from the uniform distribution on interval $[-3.5, -1.5] \cup [1.5, 3.5]$ if $p \leq 5$, and μ_p is generated from a uniform distribution on $[-0.5, 0.5]$, otherwise.

Scenario 2. We generate outcome images from the following voxel-wise regression

model:

$$Z_i(v) = \beta_0(v) + \sum_{p=1}^{20} \beta_p(v) X_{ip}(v) + \epsilon_i(v), \quad \text{for} \quad i = 1, \ldots, 150, \tag{3.3}$$

where $\epsilon_i(v)$ are independently sampled from $\mathrm{N}(0, 0.1)$. We simulate $\beta_p(v)$ from a Gaussian process with mean zero and correlation kernel $\exp\{-15 d^2_{v_1,v_2}\}$. The marginal variance of $\beta_p(v)$ is 2.0 for $p \leq 5$ and 0.5 otherwise.

Scenario 3. We simulate data from our spatial Bayesian latent factor (SBLF) model (see Figure B.1 for an illustration). We first define a set of basis functions using Gaussian kernels with equally spaced kernel where the knots are defined on grid points $\{1, \cdots, 32\} \times \{1, \cdots, 32\}$. We follow the parameter expansion method to generate working parameters and then take the transformations to obtain the original parameters. We fix the true number of latent factors K to five. Loading elements are first generated from a normal distribution. Then, for each simulated loading vector, we replace those simulated values outside its 50% CI by zero's to maintain the sparsity of the loading matrix. . For the imaging predictor indicator $\gamma_p = 1.0$ if $p \leq 5$ and $\gamma_p = 0$ otherwise. See more details in Figure B.1.

We choose the values of parameters in above three scenarios in terms of the signal-to-noise ratio (SNR), a measure of signal strength relative to background noise. The definition of SNR used in our study is

$$\mathrm{SNR} = \frac{\mathrm{Var}\left[\mathrm{E}\{Z_i(v) \mid X_{i1}(v), \ldots, X_{iP}(v)\}\right]}{\mathrm{Var}\{\epsilon_i(v)\}}$$

To make sure that the results in different scenarios are comparable, we choose the parameter values so that the SNR's of simulated observations in the three scenarios have similar distributions with mean 30 and range [1, 100].

We run the MCMC algorithm for 25,000 iterations with 15,000 burn-in. We compute the posterior mean and credible intervals for the parameters of interest. For

all the parameters with Gamma priors in Section 3.3.1, we set both shape and scale parameters to 1.0. We fix $\sigma_\lambda^2 = \sigma_\mu^2 = \sigma_\alpha^2 = 1.0$. For other hyperprior specifications, $\omega \sim \text{Beta}(1.0, 1.0)$. All initial values are sampled from their corresponding prior distributions, except that the initial values of γ_p are 1.0. Further, we fit the model with 1, 5, 10 and 20 latent factors respectively.

3.4.2 Results

Table 3.2 shows estimation and prediction accuracy for the three scenarios, including MSE, MSPE and the proportion of observations for which our method produces smaller MSE or MSPE compared to the other methods. The method used as the true generating model in each scenario has the smallest MSE for test sets. In Scenario 1, when the data are generated from the linear regression model, SBLF outperforms the other methods for about 20% to 30% of observations in the test sets. With a similar SNR, in Scenario 2, when the data are generated from the voxel-wise regression model, SBLF achieves a smaller MSPE and over 90% better predictions than the voxel-wise regression method. In Scenario 3, when the data are generated from SBLF, SBLF with a correct number of latent factors K leads to the best performance and incorrect K can result in the estimation bias.

As we discussed in Section 3.3.2, it is of interest to evaluate different criteria for selecting the number of latent factors K. Our simulation study in Scenario 3 indicates that some of the widely used model comparison criteria, including DIC, BIC, BF and R-squared, could not help to identify the correct value of K. Specifically, BIC always prefers small K, while R-squared and BF are in favor of the largest K. The selection of K using DIC varied a lot from the 10 repeated data sets. However, the MSPE of outcomes for test set is a robust measure for choosing K. As shown in Table 3.2, in Scenario 3, SBLF with a correct value of K, $K = 5$, has the smallest MSPE and the largest proportion of outperformed observations than other methods in all ten

repeated studies. Further more, given $K = 20$ in Scenario 3. Figure 3.2 shows the summarized statistics of each of the 20 loading vectors, base on which we correctly determine the value of K ($K = 5$) using the "elbow method" mentioned in Section 3.3.2). Since the posterior inference on predictor selection are biased when $K = 20$, we fit the model with $K = 5$ and obtain the estimated posterior inclusion probability for the first five predictors are exactly 1.0 and zero for the rest predictors. This perfectly recovers the true parameter settings.

3.5 Application

3.5.1 The motivating HCP data

We apply our SBLF model to analyze a subset of neuroimaging data from the Human Connectome Project (HCP). Our goal is to make prediction on the individual task-evoked images using the corresponding task-independent images. *Tavor et al.* (2016) performed a similar analysis on the same data set using a simple linear regression approach ignoring the spatial dependence among voxels within parcels. Their analysis focused on the cortical surface imaging measurements, while our model is developed for analysis of the volumetric imaging data on 19 sub-cortical regions. The data set comprises 98 subjects functional and structure imaging data from the Q3 release. Details of all acquisition parameters and processing mechanisms are described in (*Barch et al.*, 2013).

In our analysis, we focus on the 19 sub-cortical regions consisting of 31,870 voxels. The outcome image is the faces-shapes contrast map derived from the EMOTION task fMRI data. The predictor images are 32 sub-cortical seed maps derived from the resting-state fMRI data. More details on the definition of the 32 sub-cortical seed maps can be found in (*Tavor et al.*, 2016). See examples of the outcome and predictor images shown in Figure B.2, B.3 and B.4 in the Appendix. It has been well known

that the amygdala complex as part of the neural circuitry consistently associates with emotional functioning (*Phan et al.*, 2002). Hence, we report the two amygdala regions on the left and right sides of the brain as examples to demonstrate our application analysis and results. There are 315 and 332 volumetric voxels within the left and right amygdala regions respectively and their corresponding example outcome and predictor maps are shown in Figure B.5 and B.6 in the Appendix.

3.5.2 Analysis

For SBLF, we run the proposed MCMC algorithm for 50,000 iterations with 25,000 burn-in. We adopt the same prior specifications as those used in simulation study. The initial values are randomly sampled from their prior distributions except that the initial values of the predictor selection indicators are set to one. We specify the basis functions for the left and right amygdala containing 51 and 58 knots respectively. To choose a good hyper-parameter b in the basis functions, We resort to a cross validation approach and consider three candidate values $\{1/10, 1/20, 1/30\}$. The basis functions with the three values are shown in Figure B.7, B.8 and B.9 in the Appendix. A smaller b results in more overlaps among basis functions and leads to smoother approximations to the outcome images. In contrast, basis functions with a large b do not overlap much resulting in a less smooth approximation to the outcome images. A very large b may lead to less flexibility in modeling the spatially-varying coefficients in the model and a bad approximation to outcome images. To choose the number of latent factors, we start with a large number $K = 20$ and applied the "elbow method" according to the sparsity of loading matrices to determine a smaller number of latent factors, given which we refit the model.

To the same data set, we also apply the other two simple alternatives: the linear regression approach (*Tavor et al.*, 2016), and the voxel wise analysis approach as we described in the simulation study. Given the optimal b and K, we re-fit our model

using imaging data from all 98 subjects and perform the same analysis for all 19 sub-cortical regions separately.

We check the convergence of all the MCMC simulations using the Gelman-Rubin diagnostics (*Gelman et al.*, 1992). Given each selected hyper-parameter b and number of latent factors K, we run five MCMC chains with different initial values. The potential scale reduction factors (PSRF) are estimated for each voxel point in the outcome images. The point estimates of PSRF range from 1.000 to 1.005 (median 1.000, mean 1.000) and the upper confidence limits have the maximum value 1.016 (median 1.000, mean 1.000), indicating the convergence of the MCMC simulations.

3.5.3 Results

Table 3.3 shows the 10-fold CV model fitting and prediction accuracy using SBLF with different values of b and K compared with the other two methods. For both left and right amygdala regions, when $b = 10$, SBLF has the smallest averaged MSPE and the largest proportions of better predicted outcomes for test sets than the other methods. Compared with the linear and voxel-wise regression methods, SBLF has the smallest MSE for fitting the outcome images in the training data set for all the combinations of K and b. For the left amygdala region, the optimal choice for the number latent factors is 9, with which the MSPE of SBLF is 1.168, smaller than that of linear regression (1.357) and voxel-wise regression (1.540). For over 69% and 65% of outcome images in the test set, SBLF produces a smaller MPSE compared to linear regression and voxel-wise regression, respectively. These proportions for the right amygdala are even larger (77.55% and 73.47%), as shown in Table 3.3.

Table B.2 in the Appendix shows the summarized results for all the 19 sub-cortical regions compared with the linear regression method. The squared errors of outcomes from our model are about 10.2 times smaller than the linear regression method on average for all parcel regions. SBLF has much higher R-Squared values than the

linear regression method across all parcels (0.934 v.s. 0.428 on average). Hence, our proposed SBLF model outperforms the linear regression model (*Tavor et al.*, 2016) for predicting the task-evoked functional brain activity from the task-free volumetric images in the sub-cortical regions.

From MCMC samples of the predictor selection indicators, $\boldsymbol{\gamma}$, we can estimate the posterior inclusion probability for each predictor image, indicating the uncertainty of including the corresponding predictor images into the model. For each amygdala region, by placing a threshold value on the posterior inclusion probability, we can obtain a set of predictor images that are associated with the outcome image with certain uncertainty level. We vary the threshold from 0.0 to 0.9 and list the corresponding set in Table B.1 in the Appendix. For the right amygdala region, the posterior probability of including the 28^{th} cortical seed map into the model is larger than 0.6. Among all the predictor images, this cortical seed map has the strongest association with the faces-shapes contrast image in the Emotion domain. Similarly, in the left amygdala region, the same predictor image also has a relative strong association (the posterior inclusion probability larger than 0.5) with the outcomes in the same task domain. However, in the left amygdala region, the 13^{th} and 15^{th} predictor images have more contributions to the outcome predictions given their estimations of their γ's over 0.8 in the left amygdala region. These strong associations do not appear in the right amygdala regions. These two sub-cortical seed maps are from the cerebellum sub-cortical seeds, indicating the significant associations between cerebellum structure and emotional functions in left amygdala.

It is of great interest to understand how the predictor images are associated with the outcome images. As presented in (3.2.4) and (3.1) in our SBLF model, $\psi(v, v')$ represents the prediction effect on a voxel v in the outcome image from any voxel v' in the predictor image. Figure 3.3 shows the estimated $\psi(v, v')$ on five outcome voxels v in the left amygdala region. The same predictor image have varied effects on different

outcome voxels. For example, the first outcome voxel (the first row in Figure 3.3) are negatively associated with closed voxels in predictor images, while the last two voxels (the forth and fifth rows) have more positive affects from voxels in the similar locations in the predictor images. In contrast, there are not significant effects from predictor images for the other two voxels (the second and third rows). Meanwhile, the significant associations exit within not only nearby voxels but also long-distance voxels. For example, for the first outcome voxel ($x = -20, y = -4, z = -30$) in Figure 3.3, some long-distance voxels in the image slice ($z = -18$) positively associated with it, while its nearby voxels have significant negative effects. Those estimated associations between predictor and outcome images from our proposed model can be implicit and significant for exploring brain functions, which cannot be provided by the other two methods.

3.6 Discussion

In this work, we propose a spatial Bayesian latent factor model for image-on-image regression. We use low-dimensional latent factors as bridge connecting the outcome image and predictor images in the same high dimensional imaging space. The proposed method is flexible to model the spatial dependence through pre-specified basis functions without imposing strong assumptions of spatial patterns. Our SBLF model can identify the associations between the outcome image and predictor images across the whole image space, not restricted to voxels from the same locations or nearby neighbors. The low-dimensional latent factors integrate information from predictor images through a regression model with spatially-varying coefficients. This regression model can include other clinical patient characteristics for integrative analysis. Our method can be applied to jointly analyze multimodality imaging data, such as the resting-state fMRI and the task fMRI and structural MRI.

We discuss the limitations and potential feature directions for our method. First,

the cross-validation approach to determine the number of basis functions and the number of latent factors is very computationally intensive. An alternative way is to treat those numbers as unknown parameters and assign a multinomial prior distribution (*Ghosh and Dunson*, 2009), then we can make fully Bayesian inference on the model. This approach often requires to develop trans-dimensional MCMC algorithms, which are challenging in practices. Second, we make a strong assumption that the spatially varying coefficients are common for all predictor images, while the spatial predictive effects of different predictor images can be different. This assumption might decrease the power to detect the important predictive effects and may inflate the false positive rate. We can relax this assumption by introducing the predictor specific spatially-varying coefficients, which will increase model complexity and thus more efficient computational algorithms are needed. Third, different subjects may have heterogeneous associations between the task related brain activity and the resting-state activity. Our current SBLF model cannot capture this heterogeneity. We can potentially extend our model by introducing subject-specific spatially-varying coefficients with clustering structures. Fourth, our current model focused on volumetric data and not applicable to the surface data, we can further extend our method to brain surface data by project cortex data in a sphere and then generate smooth basis functions based on kernels of spherical harmonics in the same sphere.

Table 3.1: Inferential and working models for parameter expansion approach with $i = 1, 2, \cdots, N,\ k = 1, \cdots, K,\ m = 1, \cdots, M$.

Inferential Model	Working Model	Transformations
$\boldsymbol{\theta}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i$	$\boldsymbol{\theta}_i = \boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^* + \boldsymbol{\zeta}_i$	$\lambda_{mk} = S(\lambda_{kk}^*)\phi_k^{-1}\lambda_{mk}^*$
$\boldsymbol{\eta}_i = \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}_i + \boldsymbol{\epsilon}_i$	$\boldsymbol{\eta}_i^* = \boldsymbol{\mu}_i^* + [\boldsymbol{\beta}^*]^T\tilde{\boldsymbol{X}}_i + \boldsymbol{\epsilon}^*$	$\eta_{ik} = S(\lambda_{kk}^*)\phi_k\big(\eta_{ik}^* - \mu_{ik}^*\big)$
$\boldsymbol{\beta} = \boldsymbol{b}\boldsymbol{\alpha}$	$\boldsymbol{\beta}^* = \boldsymbol{b}\boldsymbol{\alpha}^*$	$\boldsymbol{\beta}_k = S(\lambda_{kk}^*)\phi_k\boldsymbol{\beta}_k^*$
$\boldsymbol{\zeta}_i \sim N(0, \sigma_\zeta^2\boldsymbol{I})$	$\boldsymbol{\zeta}_i \sim N(\boldsymbol{0}, \sigma_\zeta^2\boldsymbol{I})$	$\boldsymbol{\alpha}_k = S(\lambda_{kk}^*)\phi_k\boldsymbol{\alpha}_k^*$
$\boldsymbol{\epsilon}_i \sim N(0, \sigma_\epsilon^2\boldsymbol{I})$	$\boldsymbol{\epsilon}_i^* \sim N(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{\Phi}^{-1})$	$\epsilon_{ik} = S(\lambda_{kk}^*)\phi_k\epsilon_{ik}^*$

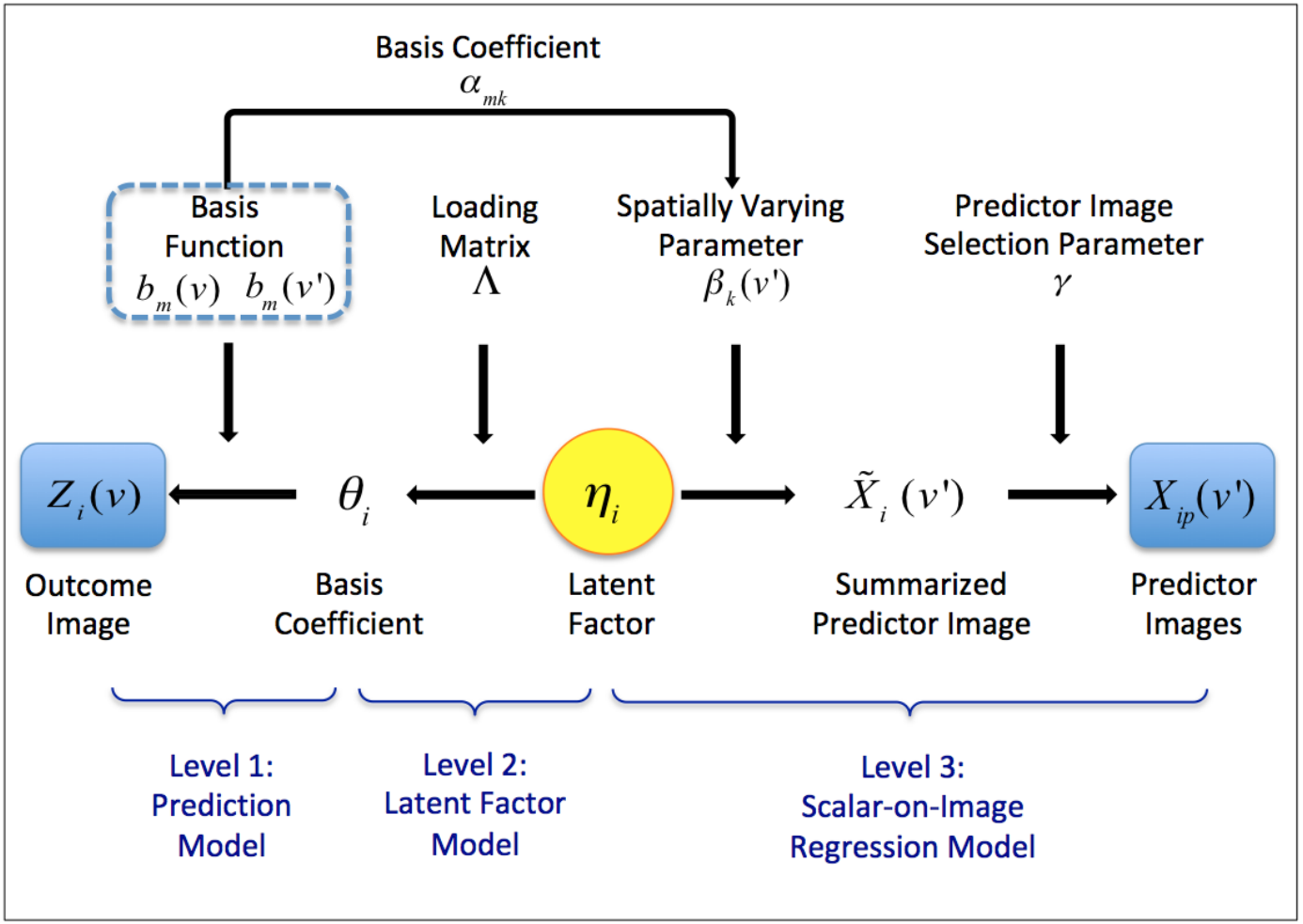


Figure 3.1: Graphical representation of the proposed spatial Bayesian latent factor model for image-on-image regression analysis.

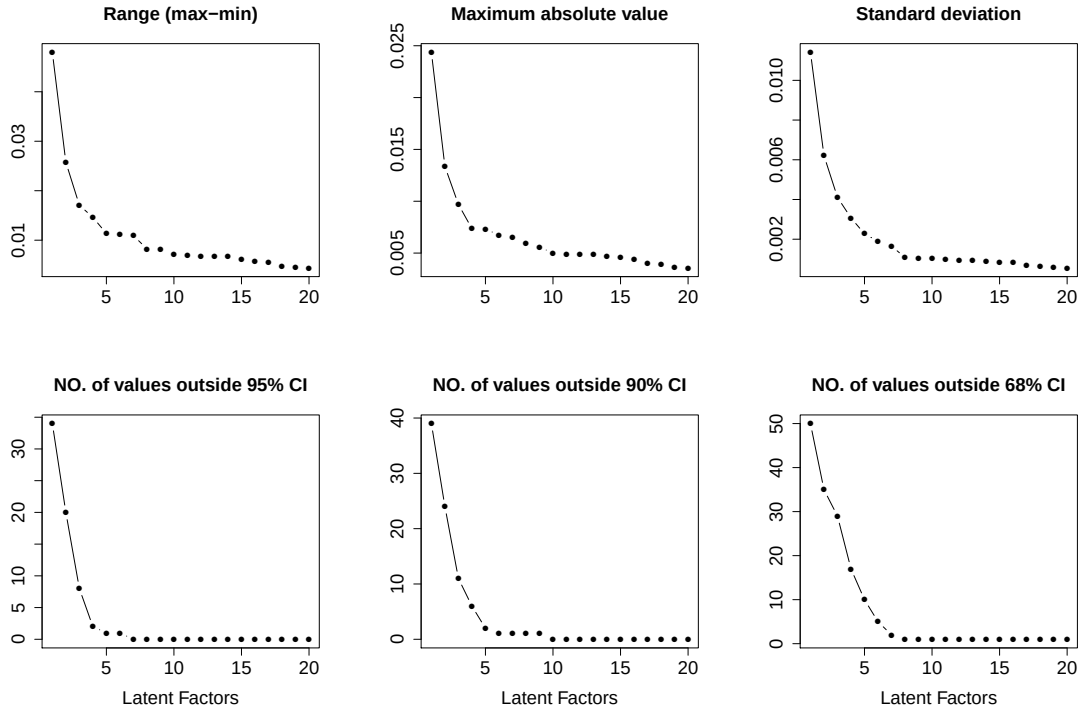


Figure 3.2: Statistical measures of the posterior mean estimations of loading matrix in simulation study Scenario 3, fitted with $K = 20$ (true $K = 5$). X-axis is the index of latent factors from 1 to 20. Figures on top are range (max value - min value), maximum absolute value and standard deviation of each loading vector, respectively. Figures on the bottom are number of values in each loading vector outside the 95%, 90% and 68% confidence interval of the whole loading matrix.

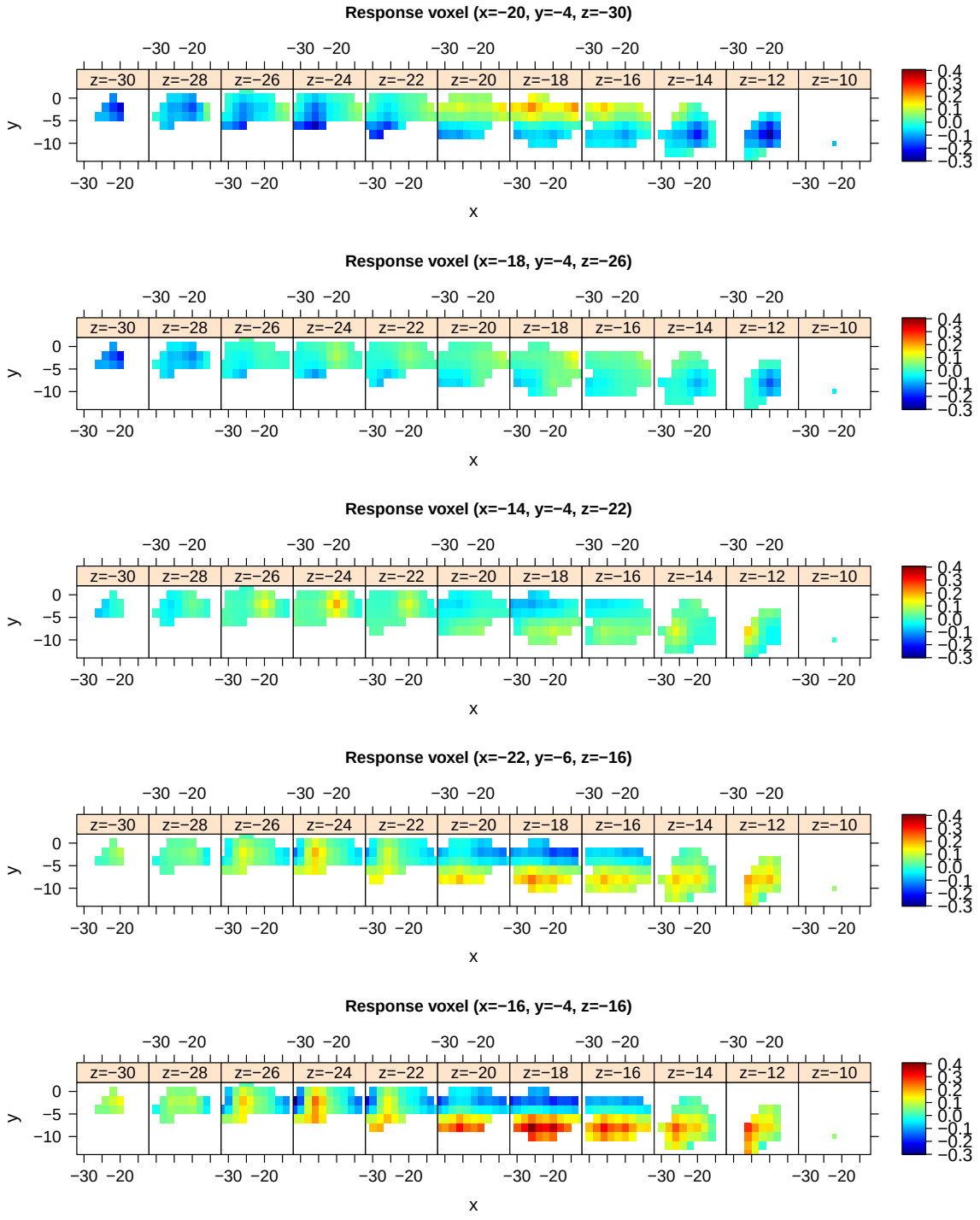


Figure 3.3: Spatially varying prediction effects $\psi(v, v')$ on five different response voxels v from all predictor voxels v'. Both v and v' are in the left amygdala maps. All maps are plotted on the same color scale.

Table 3.2: Simulation study results for Scenario 1-3. In each scenario, three models are fitted and their results compared in terms of 1) MSE, 2) MSPE and 3) the proportions of observations with smaller MSE/MSPE using SBLF than the linear (%) or voxel-wise regressions (%*). Results of multiple values of K used in our method are included and the true value of K used for simulations in Scenario 3 is $K = 5$. MSE/MSPE is reported as the averaged values over the total $32 \times 32 = 1024$ grid points, the 10 simulated datasets and 100/50 subjects for training/ test sets.

Generating Model	Analysis Method	K	Training			Test		
			MSE	%	%*	MSPE	%	%*
Scenario 1: Linear	Linear		0.010	-	-	1.024	-	-
	Voxel-wise		1.016	-	-	1.050	-	-
	SBLF	1	0.014	0.00	100.00	1.231	23.40	27.20
		5	0.014	0.00	100.00	1.395	31.60	34.20
		10	0.014	0.00	100.00	1.474	29.40	32.60
		20	0.014	0.00	100.00	1.509	26.00	27.80
Scenario 2: Voxel-wise	Linear		0.136	-	-	0.511	-	-
	Voxel-wise		0.008	-	-	0.496	-	-
	SBLF	1	0.023	100.00	0.00	0.314	94.40	92.80
		5	0.023	100.00	0.00	0.316	94.20	92.60
		10	0.023	100.00	0.00	0.322	93.40	91.80
		20	0.023	100.00	0.00	0.341	91.80	90.80
Scenario 3: SBLF	Linear		0.896	-	-	3.656	-	-
	Voxel-wise		2.510	-	-	3.642	-	-
	SBLF	1	0.149	100.00	100.00	3.436	59.80	56.40
		5	**0.149**	**100.00**	**100.00**	**1.633**	**94.80**	**94.40**
		10	0.149	100.00	100.00	1.923	89.40	89.00
		20	0.149	100.00	100.00	3.347	58.80	57.20

SBLF: our proposed spatial Bayesian latent factor model.
MSE: mean squared error of outcome estimations for training set.
MSPE: mean squared prediction error of outcome predictions for test set.
%, %*: proportion of simulated observations with smaller MSE/MSPE using SBLF than the linear and voxel-wise regression method, respectively.

Table 3.3: Application results of the left (1) and right (2) amygdala region. Performance of three different methods are compared in terms of 1) MSE, 2) MSPE and 3) the proportions of observations with smaller MSE/MSPE using SBLF than the linear (%) or voxel-wise regressions (%*). MSE/MSPE is reported as the averaged values over all voxels, subjects and 10-folds cross validation. Two tuning parameters, bandwidth value b for basis functions and the number of latent factors K, are tested to determine their optimal values b^* and K^*. The SBLF model is re-fitted with the value of K^* determined using the "Elbow" method with the loading matrix estimated with $K = 20$.

(1) Left amygdala region

Method	Bandwidth	NO. of Latents	Training			Test		
			MSE	%	%*	MSPE	%	%*
Linear	-	-	0.562	-	-	1.357	-	-
Voxel-wise	-	-	0.644	-	-	1.540	-	-
SBLF	$\boldsymbol{b* = 10}$	$K = 20$	0.060	100.00	100.00	1.198	66.33	63.37
		$\boldsymbol{K^* = 9}$	0.063	100.00	100.00	**1.168**	**69.39**	**66.30**
	$b = 20$	$K = 20$	0.009	100.00	100.00	1.304	54.08	60.02
		$K^* = 8$	0.009	100.00	100.00	1.348	51.02	58.16
	$b = 30$	$K = 20$	0.005	100.00	100.00	1.830	29.59	40.82
		$K^* = 5$	0.005	100.00	100.00	1.830	20.41	40.82

(2) Right amygdala region

Method	Bandwidth	NO. of Latents	Training			Test		
			MSE	%	%*	MSPE	%	%*
Linear	-	-	0.651	-	-	1.539	-	-
Voxel-wise	-	-	0.735	-	-	1.866	-	-
SBLF	$\boldsymbol{b* = 10}$	$K = 20$	0.066	100.00	100.00	1.359	70.41	73.47
		$\boldsymbol{K^* = 10}$	0.069	100.00	100.00	**1.260**	**77.55**	**73.47**
	$b = 20$	$K = 20$	0.010	100.00	100.00	1.398	68.37	74.49
		$K^* = 7$	0.010	100.00	100.00	1.941	38.78	50.00
	$b = 30$	$K = 20$	0.005	100.00	100.00	2.127	19.39	42.86
		$K^* = 5$	0.006	100.00	100.00	3.195	10.26	12.82

SBLF: our proposed spatial Bayesian latent factor model.
MSE: mean squared error of outcome estimations for training set.
MSPE: mean squared prediction error of outcome predictions for test set.
%, %*: proportion of simulated observations with smaller MSE/MSPE using SBLF than the linear and voxel-wise regression method, respectively.

CHAPTER IV

Extension of Spatial Bayesian Latent Factor Model to Cortical Surface Images

In this chapter, we extend the image-on-image regression model proposed in Chapter III to the case where outcome is a cortical surface image and predictor images are volumetric maps. The two types of spatial dependence are captured through the basis expansion approach. On one hand, we approximate the surface image using a set of spherical harmonics basis functions and link the basis coefficients to predictors through a latent factor model. On the other hand, we approximate the volumetric predictor images by the Gaussian kernels and assign GP priors to the spatially-varying regression coefficients. We perform simulation studies to evaluate the out-of-sample prediction performance of our method compared with ridge regression for different scenarios. We apply the proposed method to predict task-related spherical z-score maps using 32 sub-cortical volumetric seed maps from Human Connectome Project. Our method has a better prediction accuracy and can identify more important active brain regions compared to the ridge regression.

4.1 Introduction

The functional magnetic resonance imaging (fMRI) is a popular neuroimaging technique used to localize regions of brain activated by a task or stimulus (*Lindquist et al.*, 2008). fMRI uses Blood Oxygenation Level Dependent (BOLD) contrast to measure neuronal activity indirectly (*Ogawa et al.*, 1990, 1993). The traditional fMRI is volume data recorded at one single time point and measured in voxels (equally sized) filling the three-dimensional (3D) brain "native" space. Each voxel contains one value, representing the average signal measured at the given location. The general linear model (*Martin and Maes*, 1979) is the fundamental approach to model volumetric responses at each and every voxel, providing activation maps linked to a particular contrast (*Worsley and Friston*, 1995). It has been known that the neural activity occurs in gray matter. The volumetric fMRI data, however, consists of some other tissues involving white matter and cerebral spinal fluid (*Mejia et al.*, 2017).

In the past two decades, there is a growing popularity of surface-based fMRI data versus the traditional volumetric fMRI. The fMRISurface pipeline in the Human Connectome Project (HCP) has been developed to take a volume time series and map it to the standard Connectivity Informatics Technology Initiative (CIFTI) grayordinates space (gray-matter surface vertices or volume voxels) (*Glasser et al.*, 2013). The first step of this transforming procedure is to use a high-dimensional structural imaging to define which fMRI voxels are within the grey matter ribbon (*Dale et al.*, 1999). The second step is to form a smoothed 2D manifold within each hemisphere through a mesh applied to white matter surface. Third, the smoothed surface is inflated to a sphere, in which subjects brain are aligned to the template brain by aligning cortical folding patterns. Finally, the same transformation can be applied to each fMRI volume to obtain a cortical surface fMRI time series. The fMRISurface HCP pipeline produces a triangular mesh with approximately 30,000 surface vertices in both left and right cerebral cortices.

There are many benefits of cortical surface fMRI over volumetric fMRI data, including (1) easier imaging visualization, (2) reduced image dimension (less vertices than voxels), (3) only consisting gray matter tissue and (4) improved across-subject alignment (*Mejia et al.*, 2017; *Glasser et al.*, 2013; *Gao et al.*, 2015; *Coalson et al.*, 2018). The most important one is that the distance between two vertices on the cortical surface is neurologically meaningful. The neighborhood of voxels in Euclidean distance may be neurologically misspecified, such as two voxels with short straight-line distance in Euclidean space but from different surface areas or even two types of tissues. In contrast, neighbour or nearby vertices in cortical surface fMRI are defined based on the geodesic distance along the surface so that they tend to present similar neuronal activities.

For cortical surface imaging data obtained along curved non-Euclidean surfaces, traditional statistical analysis and smoothing techniques based on the Euclidean metric structure are inefficient. For example, spherical wavelets or basis functions are preferred to approximate signals on the sphere than Gaussian kernels, the non-linear function of Euclidean distance. In particular, a set of spherical harmonics (SH) functions (*Kennedy and Sadeghi*, 2013) is a natural choice of basis functions, widely used for cerebral surface parameterization (*Chung et al.*, 2008; *Gerig et al.*, 2001; *Gu et al.*, 2004; *Kelemen et al.*, 1999; *Yotter et al.*, 2010). The SH approximation represents the coordinates of mesh vertices as a linear combination of the SH functions. Each SH is an orthonormal basis of functions defined on the surface of a unit sphere, encoding the main global geometric features. Its power lies in the fact that it only requires a small number of linear coefficients to represent general functions and large data sets accurately (*Schröder and Sweldens*, 1995). Those SH linear coefficients can be estimated in a least-squares fashion, however, it might be difficult to implement due to the difficulty of high-dimensional matrix inversion given a large number of vertices (*Chung et al.*, 2008). Alternatively, we can use the iterative residual fitting algo-

rithm, creating a less accurate reconstruction due space partition (*Elahi et al.*, 2017). Bayesian methods, however, have been sucecessfully applied for making inference on the SH coefficients (*Das et al.*, 2015; *Muir and Tkalčić*, 2015; *Pinaud et al.*, 2015; *Angers et al.*, 2005).

In Chapter III, we have developed a spatial Bayesian latent factor (SBLF) model to predict task-related maps using task-free images, where both the outcome and predictors are volumetric data from the same region of interest (ROI). Considering that the surface-based fMRI may have better signals of neural activities than volumetric fMRI, we expect some benefits from applying our SBLF model to predict task-evoked cortical surface image with task-free volumetric maps. We approximate the volumetric outcome images using a set of linearly weighted basis functions, which are simulated from isotropic Gaussian kernels based on Euclidean distance and not suitable for modelling surface-based imaging data. Therefore, we naturally choose the SH functions as bases for the analysis of cortical surface images mapped to a sphere. The linear coefficients of SH functions are modeled using latent factor model and estimated under the hierarchical Bayesian framework proposed for SBLF model.

The remainder of this chapter is organized as follows. Section 4.2 presents the SBLF model which has been developed in Chapter III with the SH functions defined in the spherical surface domain. The details of the SH functions are described in Section 4.2.1. We assess the model performance and compare the proposed method with ridge regression via simulation studies in Section 4.3. We then apply the model to the analyses of the HCP data in Section 4.4. Finally, we draw conclusions with a brief discussion in Section 4.5.

4.2 Methods

In this section, we represent the SBLF model developed in Chapter III with an extension for the case where the outcome is a cortical surface image and predictors

are sub-cortical volumetric images. Suppose each of the N observation consists of one outcome image and P predictor images. Outcome images are cortical surface data mapped to a unit sphere, denoted $\mathbb{S}^2$, while predictor images are registered to a volumetric space $\mathcal{R}$. Note that $(\mathbb{S}^2, \mathcal{R})$ may practically refer to either the whole brain region or some regions of interest; and the following model and parameter settings are region-specific. For subject i (i=1, $\cdots$, N), let $Z_i(s)$ represent the outcome imaging value at surface vertex $s \in \mathbb{S}^2$ and $X_{ip}(v)$ $(p = 1, \ldots, P)$ be the p^{th} predictor image value at voxel $v \in \mathcal{R}$. Next, we define a set of SH functions to approximate spherical surface outcome images.

4.2.1 Spherical Harmonics

We denote by $\mathbb{S}^2$ a unit sphere embedded into $\mathbb{R}^3$, that is,

$$\mathbb{S}^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\},$$

where $\|\cdot\|$ represents the Euclidean norm. Consider the parameterization of $\mathbb{S}^2$ by $y(\vartheta, \psi) = (\sin\vartheta\cos\psi, \sin\vartheta\sin\psi, \cos\psi)$ where $\vartheta \in [0, \pi]$ and $\psi \in [0, 2\pi)$ are the pole angle and azimuthal angle, respectively, in the spherical coordinate system. In practice, it is sufficient to only use the real-valued SH functions. It is more convenient for the real-valued stochastic model (*Homeier and Steinborn*, 1996; *Courant and Hilbert*, 2008). Specifically, the real-valued SH of degree ℓ and order m is

$$Y_{\ell m}(\vartheta, \psi) := \begin{cases} d_{\ell m} P_\ell^{|m|}(\cos\vartheta)\sin(|m|\psi), & -\ell \le m \le -1 \\ \frac{d_{\ell m}}{\sqrt{2}} P_\ell^{|m|}(\cos\vartheta), & m = 0 \\ d_{\ell m} P_\ell^{|m|}(\cos\vartheta)\cos(|m|\psi), & 1 \ge m \ge \ell \end{cases}$$

where $d_{\ell m} = \sqrt{\frac{2\ell+1}{2\pi}\frac{(\ell-|m|)!}{(\ell+|m|)!}}$. $P_\ell^{|m|}$ is the associated Legendre functions of order m (*Courant and Hilbert*, 2008), defined as

$$P_\ell^m(x) = \frac{(1-x^2)^{m/2}}{2^\ell \ell!}\frac{\partial^{\ell+m}}{\partial x^{\ell+m}}(x^2-1)^\ell, \quad x \in [-1,1]$$

for $\ell \in \mathbb{N}_0, m = 0, \ldots, \ell$. The SH functions are orthonormal basis on $L^2(\mathbb{S}^2, \mathbb{C})$ and every real-valued function f in $L^2(\mathbb{S}^2, \mathbb{C})$ can be represented by the SH series expansion

$$f = \sum_{\ell=1}^{\infty}\sum_{m=-\ell}^{\ell} c_{\ell m} Y_{\ell m} \tag{4.1}$$

The coefficients $c_{\ell m}$ for $\ell \in N_0, m = -\ell, \ldots, \ell)$ are a sequence of centred, Gaussian random variables (*Müller*, 2006; *Lang et al.*, 2015). The expansion (4.1) is commonly referred as the SH representation.

4.2.2 Spatial Bayesian Latent Factor Model

For the sub-model at Level 1, We first define a set of SH functions $\{Y_{lm}(s)\}_{|\mathbb{S}^2|\times(L+1)^2}$ with degree $l = 0, \ldots, L$ and order $m = -l, \cdots, l$, where $|\mathbb{S}^2|$ is the number of vertices in $\mathbb{S}^2$ and $(L+1)^2$ is the total number of SH functions. For each subject i and any vertex $s \in \mathbb{S}^2$, we assume

$$Z_i(s) = U(s) + \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} \theta_{i\ell m} Y_{\ell m}(s) + e_i(s), \qquad e_i(s) \sim \mathrm{N}(0, \sigma_e^2)$$

where $U(v)$ represents the population-level averaged outcome image. As a prior specification, we assume $\{U(s)\}_{s\in\mathbb{S}^2}$ are independent and identically distributed as a normal distribution with mean zero and variance σ_u^2. The residuals $e_i(s)$, $\forall i \in \{1, \ldots, N\}$ and $s \in \mathbb{S}^2$, are independently and identically distributed random variables following $\mathrm{N}(0, \sigma_e^2)$. The term $\sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell}\theta_{i\ell m}Y_{i\ell m}(s)$ captures the spatial dependence and

smoothness of the subject-specific outcome images among vertices in $\mathbb{S}^2$.

Similar to the models in Chapter III, at Level 2, subject-specific SH coefficient $\theta_{i\ell m}$ are separated into a set of K latent factors $\{\eta_{ik}\}_{k=1}^{K}$ for subject i and sparse loading elements $\{\lambda_{\ell mk}\}_{k=1}^{K}$ common for all subjects. This sub-model is defined as

$$\theta_{i\ell m} = \sum_{k=1}^{K} \lambda_{\ell mk}\eta_{ik} + \zeta_{i\ell m}, \qquad \zeta_{i\ell m} \sim \mathrm{N}(0, \sigma_\zeta^2)$$

The error term $\zeta_{i\ell m}$ explains the variation of the basis coefficients $\theta_{i\ell m}$ that cannot be explained by the latent factors. Same as Chapter III, for the prior specification for the sparse loading coefficients λ_{mk}, we resort to inducing prior distribution through the parameter-expansion (*Ghosh and Dunson*, 2009), leading to more efficient posterior computation.

At Level 3, we link each subject-specific latent factor η_{ik} via the scalar-on-image regression:

$$\eta_{ik} = \sum_{v\in\mathcal{R}} \widetilde{X}_i(v)\beta_k(v) + \epsilon_{ik}, \quad \widetilde{X}_i(v) = \sum_{p=1}^{P} \gamma_p X_{ip}(v), \quad \epsilon_{ik} \sim \mathrm{N}(0, \sigma_\epsilon^2),$$

where the error term ϵ_{ik} follows the normal distribution with mean zero and unit variance σ_ϵ^2 to ensure the identifiability of latent factors. We effectively reduce the dimension of coefficients by introducing a summarized predictor image $\widetilde{X}_i(v)$, the average of the selected predictor images from $\{X_{ip}(v)\}_{p=1}^{P}$. Each predictor image $\{X_{ip}(v')\}_{v'\in\mathcal{R}}$ has a corresponding selection indication γ_p. This binary indicator is assumed to follow a Bernoulli distribution with prior probability w, including the prior knowledge on the proportion of important predictor images. To account for spatial dependence in predictors, we assign a GP prior to spatially-varying coefficient $\beta_k(v)$, approximated using a basis expansion approach, i.e. $\beta_k(v) = \sum_{m'=1}^{M} \alpha_{km'} b_{m'}(v)$ with $\alpha_{km'} \sim \mathrm{N}(0, \sigma_\alpha^2)$ and a set of M bases $\{b_{m'}(v)\}_{m'=1}^{M}$. We use a small number of latent factors (K) to capture important feature information than other predictors,

and achieve efficient dimension reduction.

In summary, our proposed Bayesian hierarchical model is

$$\begin{aligned}
\text{Level 1}: \quad & Z_i(s) = U(s) + \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \theta_{i\ell m} Y_{\ell m}(s) + e_i(s), \\
\text{Level 2}: \quad & \theta_{i\ell m} = \sum_{k=1}^{K} \lambda_{\ell m k} \eta_{ik} + \zeta_{i\ell m}, \\
\text{Level 3}: \quad & \eta_{ik} = \sum_{v \in \mathcal{R}} \widetilde{X}_i(v) \beta_k(v) + \epsilon_{ik},
\end{aligned}$$

where $\widetilde{X}_i(v) = \sum_{p=1}^{P} \gamma_p X_{ip}(v)$ and $\beta_k(v) = \sum_{m'=1}^{M} \alpha_{km'} b_{m'}(v)$.

4.2.3 Model Representation

By integrating out the SH coefficient $\theta_{i\ell m}$ and latent factors η_{ik}, the conditional expectation of outcome

$$E[Z_i(s)|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iP}] = U(s) + f_i(s) = U(s) + \sum_{p=1}^{P} \gamma_p \sum_{v \in \mathcal{R}} \psi(s, v) X_{ip}(v),$$

where

$$\psi(s, v) = \sum_{k=1}^{K} \left\{ \Big[\sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \lambda_{\ell m k} Y_{\ell m}(s) \Big] \times \Big[\sum_{m'=1}^{M} \alpha_{km'} b_{m'}(v) \Big] \right\}. \tag{4.2}$$

We denote $|\mathcal{R}|$ by the number of volume voxels v in $\mathcal{R}$, $|\mathbb{S}^2|$ by the number of surface vertices s in $\mathbb{S}^2$ and write $\boldsymbol{X}_{ip} = \{X_{ip}(v)\}_{|\mathcal{R}|}$. This conditional expectation consists of the population-level component $U(s)$ and subject-level component $f_i(s)$. In detail, $X_{ip}(v)$, the subject-specific predictor image value at voxel $v \in \mathcal{R}$, contributes to the prediction of $Z_i(s)$ using the spatially dependent weight $\psi(s, v)$ and predictor selection indicator γ_p. Each weight $\psi(s, v)$ measures the average change in the outcome image at any surface vertex $s \in \mathbb{S}^2$ per unit change in value of any selected predictor image at

voxel $v \in \mathcal{R}$. Furthermore, equation (4.2) suggests that the spatially varying linear effects $\psi(s, v)$ can be decomposed as the summation of K tensor products of two linearly weighted basis expansions. Therefore, the weight matrix $\mathbf{\Psi} = \{\psi(s, v)\}_{|\mathbb{S}^2|\times|\mathcal{R}|}$ from our model not only retains complex spatial dependence structures in both the outcome and predictors but also borrows strengths from the whole brain region to predict the outcome image at each voxel.

4.2.4 Posterior Computation

As demonstrated in Section 3.3, we extend the parameter expansion (PX) method proposed by *Ghosh and Dunson* (2009) to achieve efficient posterior computation for our model. This posterior computation approach as well as the corresponding prior distributions assigned to parameters in Section 3.3.1 work in the same way for the model with cortical surface outcome images. Therefore, we skip the details of posterior computation here for avoiding redundant demonstrations.

4.3 Simulation Study

In this section, we conduct simulation studies for model assessment. The linear regression (*Tavor et al.*, 2016) and voxel-wise regression approaches used for data generation and method comparison in Chapter IV do not work in the case where the outcome and predictor images are from different image space. We then compare the performance of our proposed SBLF model with ridge regression method, which is fitted to each surface vertex $s \in \mathbb{S}$ independently.

4.3.1 Data Generation and Method

We simulated 50 datasets, each of which contains 100 observations as the training set and another 100 observations as the test set. Each simulated observation has 20 two-dimensional predictor images $\{\mathbf{X}_{i1}, \cdots, \mathbf{X}_{i20}\}$ on the $\{1, \ldots, 32\} \times \{1, \ldots, 32\}$

grid, denoted $\mathcal{R}$. Specifically, we independently generate each predictor image $\mathbf{X}_{ip} = \{X_{ip}(v)\}_{|\mathcal{R}|}$ from a Gaussian Process with mean zero function and a covariance function that $c(v_1, v_2) = 0.1^2 \times \exp\{-15 \times d^2_{v_1,v_2}\}$, where d_{v_1,v_2} is the Euclidean distance between any two grid points v_1 and v_2 in $\mathcal{R}$. Every outcome image consists of 1,250 spherical surface vertices, denoted $s \in \mathbb{S}^2$, whose coordinates covering 25 ϑ's (pole angle) and 50 ψ's (azimuthal angle) which equally distribute in [0, π) and [0, 2π], respectively.

We design two scenarios to generate data for the simulation study. In Scenario 1, our proposed SBLF model serves as the generating model. Except the basis functions used for approximating outcome images, the model and parameter settings are the same as the settings in Chapter III Section 3.4.1 with details presented in Figure B.1. Meanwhile, we generate a set of 121 SH bases with degree $L = 10$, covering all 1,250 vertices s on the surface of a unit sphere $\mathbb{S}^2$.

The generating model of Scenario 2 is based on Equation (4.2) in Section 4.2.2, where the spatially varying linear coefficient $\psi(s, v)$ is the summation of K tensor products of two linearly weighted basis expansions. We define a generating model which only captures the spatial dependence of spherical outcome images, while entirely ignoring the spatial correlation structure in predictor images. In this way, we can evaluate the performance of SBLF model when it is misspecified to the simulated data in Scenario 2. The generating model is

$$Z_i(s) = \sum_{p=1}^{P=20} \gamma_p \sum_{v \in \mathcal{R}} \Big[\psi(s, v) X_{ip}(v)\Big] + e_i(s), \qquad e_i(s) \sim N(0, 0.5)$$

where the first 10 out of $P = 20$ predictor images serve as important predictors in the generating model with $\gamma_p = 1$ and $\gamma_p = 0$ otherwise. The error term $e_i(s)$ of every spherical vertex $s \in \mathbb{S}^2$ follows independent normal distribution with mean zero and variance 0.5^2. The linear coefficient $\psi(s, v)$ only consists of a set of linearly weighted

SH bases functions to account the spatial dependence in outcome images. Specifically,

$$\psi(s,v) = \sum_{k=1}^{K=5} \sum_{\ell=0}^{L=10} \sum_{m=-\ell}^{\ell} \lambda_{i\ell mk} Y_{\ell m}(s),$$

$$\text{with } \lambda_{i\ell mk} = \bar{\lambda}_{\ell mk} + \epsilon_{i\ell mk},\ \bar{\lambda}_{\ell mk} \sim N(0, 0.1),\ \epsilon_{i\ell mk} \sim N(0, 0.05),$$

where the basis coefficients $\lambda_{i\ell mk}$ has a mean term $\bar{\lambda}_{\ell mk}$ common for all observations and subject-specific residuals $\epsilon_{i\ell mk}$. Given above linear coefficients, the generating model can be written as

$$Z_i(s) = \Big[\sum_{k=1}^{K=5} \sum_{\ell=0}^{L=10} \sum_{m=-\ell}^{\ell} \lambda_{i\ell mk} Y_{\ell m}(s) \Big] \times \Big[\sum_{p=1}^{P=20} \gamma_p \sum_{v \in \mathcal{R}} X_{ip}(v) \Big] + e_i(s) \tag{4.3}$$

In above equation 4.3, the term $\sum_{p=1}^{P=20} \gamma_p \sum_{v \in \mathcal{R}} X_{ip}(v)$ indicates that the information of the all important predictor images is summed as a scalar value across all voxels $v \in \mathcal{R}$ equally without accounting for spatial dependence in $\mathcal{R}$.

We run the Markov chain Monte Carlo (MCMC) algorithm for 50,000 iterations with 25,000 burn-in to compute the posterior mean estimations and predictions. We use Gamma(1.0, 1.0) as the prior distribution for $\sigma_e^{-2}, \sigma_\epsilon^{-2}, \sigma_\zeta^{-2}$. We also fix the values of hyper variance parameter $\sigma_\lambda^2, \sigma_\mu^2, \sigma_\alpha^2$ at 1.0. The hyper-parameter ω has a prior distribution Beta(1.0, 1.0). All initial values are sampled from their corresponding prior distributions, except that all initial values of $\{\gamma_p\}_{p=1}^{P=20}$ are set to be 1.0. We fit ridge regression model using R package "glmnet" and the tuning parameter λ is selected from 100 candidate values using 10-folds cross-validation (CV) approach.

4.3.2 Results

We use two types of statistical measures for model evaluation, (1) the root mean squared errors of estimations (RMSE) and predictions (RMSPE) and (2) R-Squared values, averaged over all spatial vertices $s \in \mathbb{S}^2$, subjects and 10-folds CV. Table 4.1

summarizes the two measures under SBLF and ridge regression models fitted to the simulated datasets in Scenario 1 and 2. In both of the two scenarios, our proposed SBLF model achieves smaller RMSE/RMSPE and higher R-Squared values for the training and validation set. Meanwhile, in Scenario 1, we find the large variations in RMSE/RMSPE and R-Squared values from the results of ridge regression method from their density plots shown in Figure 4.1. In Scenario 2, both SBLF and ridge regression model are misspecified. In particular, SBLF in Scenario 2 accounts the spatially dependence in outcome images but incorrectly model the spatial correlations in predictors, while ridge regression method captures the independent linear associations of predictor voxels but ignores the spatial structure of outcome images. Therefore, for both of the two misspecified models in Scenario 2, their values of RMSE/RMSPE are higher and R-Squared values are smaller than their corresponding values in Scenario 1. However, our SBLF model still has better performance in prediction and model fitting than ridge regression in Scenario 2 since we can correctly capture the importance spatial dependence in outcome images.

To determine an optimal number of latent factors, we first fit the SBLF model with a relatively large enough value K, e.g. $K = 20$, and then summarize estimated loading vectors in terms of six statistics to evaluate the sparseness and necessary of each loading vector. The six statistical measures are 1)standard deviation, 2) range (max value minus min value), 3) maximum absolute value, 4) the number of values in each loading vector outside the 95%, 90% and 68% credible interval (CI) of the whole loading matrix. Figure 4.2 show the plots of six summary statistics corresponding to 20 estimated loading vectors. Based on "Elbow" method, the optimal choice of K is 5 or 6, given the truth that $K = 5$ latent factors are used in the generating model of Scenario 1. Therefore, the simulation study confirms that the number of latent factors can be correctly identified based on the summary statistics of the estimated loading matrix and "Elbow" method.

4.4 Application

4.4.1 The motivating HCP data

We apply our method to analyze the HCP data consisting of 98 subjects using their functional and structural MRI scans in the Q3 release. See the details of the preprocessing procedure in *Barch et al.* (2013). In this study, the task-evoked fMRI data are some faces-shapes contrast maps from the EMOTION task domain. Their corresponding z-score maps are then derived and have been registered to both volumetric and cortical surface system. The CIFIT image consists of 32,492 vertices on each left and right cerebral surface and can be mapped to a sphere with a radius of 100.

Based on a multimodal parcellation approach of human cerebral cortex (*Glasser et al.*, 2016), there are 180 distinct cortical surface-based parcels in each left or right side of the brain. The sizes of parcels vary a lot since the total number of vertices in those surface-based parcels widely range from 31 to 810. The averaged z-score outcome values across vertices within a single parcel vary from -1.625 to 7.904. Only a few of parcels contain the relatively large z-score values, while most parcels have small z-score values concentrated around zero. It is more likely that those parcels with many large absolute values indicate strong signals of emotion activity. Therefore, instead of using all 180 surface parcels, we select 29 surface-based nested parcels on the right side of the brain as our outcome image space to reduce computation task. We show some example figures of outcome images of all 180 parcels and the 29 selected parcels separately in Figure 4.3, displayed on both cortical surface and spherical surface.

For predictor images, we use a set of 32 volumetric sub-cortical seed maps from resting-state fMRI data. These predictor images consist of 19 sub-cortical regions of interests (ROIs) and 31,870 volumetric voxels in total. It has been well known that the amygdala complex as part of the neural circuitry consistently associates with the

emotion functions (*Phan et al.*, 2002). Therefore, we further reduce the predictor image space to right amygdala ROI (332 voxels) to predict the outcome image within the 29 selected parcels in the right hemisphere of the brain.

4.4.2 Methods

We run the MCMC algorithm with a total of 50,000 iterations with 25,000 burn-in. The priors placed on hyper-parameters are the same as those for simulation studies. Initial values are all randomly generated from corresponding prior distributions. Also, the SBLF model is fitted multiple times with a set of candidate SH degree values ($L \in \{5, 10, 15, 20\}$) and the number of latent factors ($K \in \{10, 20, 30\}$). We determine the first-step choices of L and K in terms of RMSE/RMSPE and R-Squared values for the training and test sets in a 10-fold CV study. We then use the same "Elbow" method as for simulation studies to choose the optimal number of latent factors by fitting our model with selected L and a relatively large enough value of K, for example, $K = 20$. Given optimal L and K from CV study, the SBLF model is finally re-fitted to the whole data set. We also fit the ridge regression for comparison and tune the tuning parameter λ in ridge regression via a 10-folds CV study from 100 candidate values, implemented in R (version 3.3.3).

4.4.3 Results

We first determine the optimal degree of SH L and the number of latent factors K based on the results in Table 4.2. As the increase in L and K, the proposed models achieve reduced RMSE and R-Squared values for the training set because of extra parameters introduced in the model. To avoid the heavy computation and serious over-fitting issue with too many parameters, the combination ($L = 10, K = 20$) is preferred as the first-step choice, with which the fitted model has relative small RMSPE for test sets and reasonable RMSE as well as R2-Squared values for the

training data. Further, we refit the SBLF model with $L = 10, K = 20$ and describe the distributions of estimated loading matrix in Figure 4.4. As shown in the three plots on top, the magnitudes and variations within every loading vector keep decreasing to zero, suggesting some potential redundant loading vectors concentrated around zero. The plots on the bottom of Figure 4.4 indicate that about $K = 10$ latent factors is enough to conserve most of the significant loading elements (e.g., top 5% absolute values). Therefore, the optimal number of latent factor is $K^* = 10$ and the SBLF model is refitted given $L = 10$ and $K = 10$.

We further compare the SBLF model ($L = 10, K = 20$) with ridge regression method from two aspects: (1) model fitting and predictions and (2) brain active region identification.On one hand, in Table 4.1, the SBLF model fitted to this real dataset reaches much higher averaged R-Squared values (0.670) and smaller RMSE (1.705) than the ridge regression method (0.208, 2.380) for the training set. For the test set, our model achieves an R-Squared value 0.127, higher than that of the ridge regression method (0.102). When it comes to prediction performance, however, both of the ridge regression and our proposed method return predicted outcome values with smaller scales than their truly observed values in the test sets. This issue might be solved from our method by providing more informative prior distributions for some model parameters, e.g., α_k, the basis coefficients of spatially varying parameters $\boldsymbol{\beta}_k$, which can control the scale of outcomes.

On the other hand, the issue of predictions with smaller scales might not affect the performance of models in identifying task-related active brain regions. We assume that large magnitude values in the z-score maps of task-evoked images represent signals of brain activity. We define subject-specific signal regions by placing some proper thresholds on the outcome images of every subject. A set of threshold values, the top 1%, 5%, 10%, 20% and 50% of the absolute values, are placed on the observed images, estimations of the training set, and predictions of test sets from SBLF and ridge re-

gression methods, respectively. In Table 4.3, we present the number and proportions of active vertices correctly identified by SBLF and ridge regression methods, averaged over subjects and the ten folds in CV study. Meanwhile, we include the number and proportion of subjects, whose active vertices are more accurately identified by SBLF model than ridge regression method. Based on the results in Table 4.3, the SBLF model almost always achieve more correctly identified active vertices for training set than ridge regression method, although the two models have similar performance for the predictions of test sets. We also provide example figures from a randomly selected subject to show his/her actual and estimated outcome images and their corresponding active vertices for model comparisons, as shown in Figure 4.5. The threshold used in this case is the top 10% of absolute values, resulting in 489 active vertices in his/her observed outcome image (restricted to the 29 selected parcels). Among the 489 true active vertices, there are 386 (78.94%) vertices are correctly found by SBLF model, while just 282 (57.67%) vertices found by ridge regression. Meanwhile, we show the false active vertices (non-active vertices which are incorrectly identified as active vertices by fitted models) in blue color and the missed vertices (active vertices which are not identified as active vertices by fitted models) in red color in the last two plots on the second and third row in Figure 4.5. The two plots indicate that most of the false active vertices by SBLF model lie in the border of the major active region (e.g., the region in green). In contrast, the active vertices identified by ridge regression method shift away from the actual active region, considering the large red region on top and blue region on the bottom (the plot in the bottom corner in Figure 4.5). Hence, we come up with the conclusion that our proposed SBLF model has better performance in identify active regions than ridge regression method.

The posterior mean estimations of predictor selection variables $\{\gamma_p\}_{p=1}^{32}$ from the re-fitted model with $L = 10$ and $K = 10$ indicate three sub-cortical seed maps restricted to the right amygdala region, which significantly contribute to the prediction

of the task-evoked surface outcomes within the 29 selected parcels involving brain activities related to emotion function. The estimated posterior probabilities of the three predictor images are 0.978, 0.997, and 0.999, respectively, while other predictor indicators have probabilities less than 0.1. The underlying task-free brain connectivity and structure presented by the three ROIs (seeds of the selected seed maps) in the right amygdala region might have important association with the human emotion function.

4.5 Discussion

In this Chapter, we extend the image-on-image regression model proposed in Chapter III to the case where the outcome is a cortical surface image, and predictor images are volumetric maps. This method captures the complex and different spatial correlation structures in the outcome map and predictor images. We approximated the subject-level cortical surface image via the expansion on a set of spherical harmonics functions, with basis coefficients linking to image predictors through a latent factor model. Meanwhile, we assign GP priors to the spatially varying regression coefficients of the volumetric predictor images. In the simulation study, the proposed method achieves better performance in prediction accuracy than ridge regression model in two different scenarios. We also apply our method to predict spherical z-score maps derived from face-shapes contrast maps with 32 sub-cortical seed maps from resting-state fMRI. The real data analyses suggest the higher accuracy of identifying active brain regions using the SBLF method than ridge regression.

The proposed SBLF method has several contributions to general image-on-image regression. First, this method captures the complex and different spatial correlation structures in the outcome and predictor images. Linear combinations of SH basis functions and Gaussian kernels are used to account for spatial dependences of spherical and volumetric imaging data, respectively. Therefore, this image-on-image

regression model doesn't have to be restricted the same imaging space of the outcome and predictor images and will benefit the analysis of multimodal neuroimaging data and the associations of surface-based parcels and ROIs of volumetric data. Meanwhile, the proposed method implements imaging dimension reduction in two aspects. First, it reduces the dimension of outcome image to a relatively small number through a linear expansion of basis functions. Second, the proposed method link the two types of high-dimensional outcome and predictor images via a few latent factors, where the optimal number of latent factors is usually smaller than 10 in the application study.

There are several future directions. Some of the other basis functions or parametric models can be considered for approximating cortical surface images mapped on a sphere, such as spherical wavelets, in order to better capture the global and local spatial dependence. In addition, spatially varying parameters can be different among multiple predictors, considering the various spatial correlation structures of multimodal predictor images, although introducing predictor-specific spatially varying parameters will bring additional computational task and potential identifiable issues.

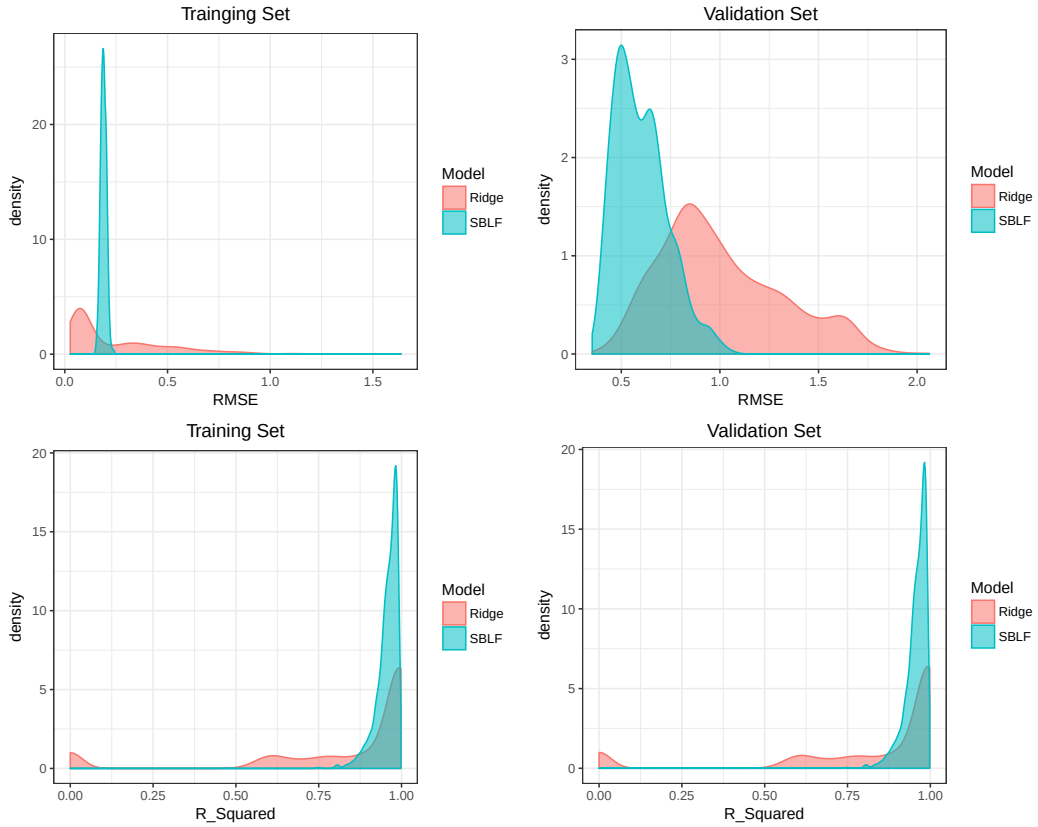


Figure 4.1: Density plots of RMSE/RMPE (1st row) and R-Squared (2nd row) values for the training (1st column) and validation (2nd column) sets in simulation study Scenario 1. Two models, SBLF (blue) and ridge regression model (red), are fitted and compared.

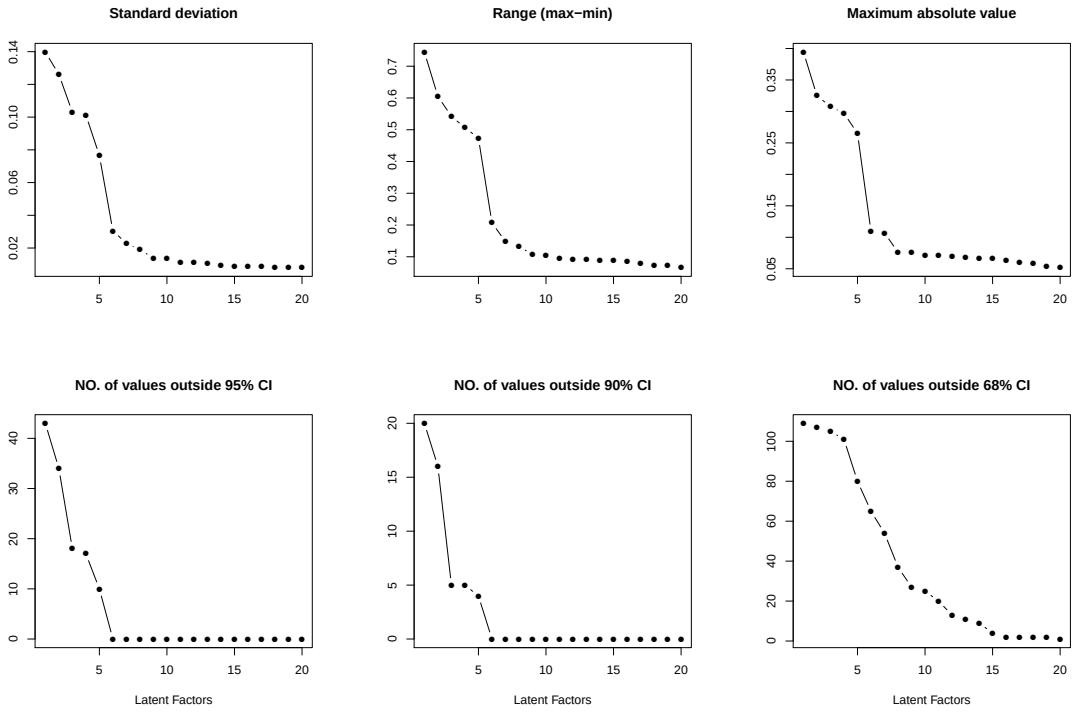
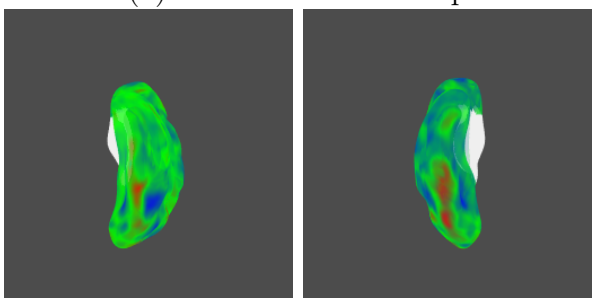
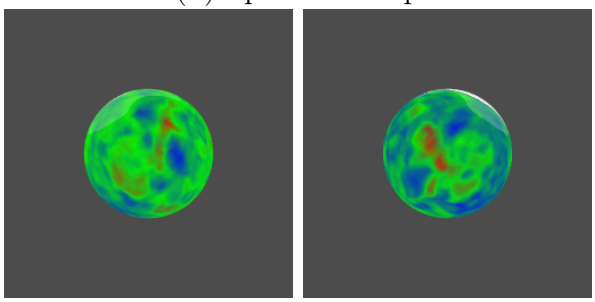


Figure 4.2: Summary statistics of the posterior mean estimations of loading matrix in simulation study Scenario 1 with $K = 20$ (true $K = 5$). X-axis is the index of latent factors from 1 to 20. Figures on top row are for statistics including standard deviation, range (max value - min value) and maximum absolute value of each estimated loading vector, respectively. Figures on the bottom show the number of values in each loading vector outside the 95%, 90% and 68% confidence interval (CI) of the whole estimated loading matrix. The determined optimal value of latent factors is $K = 5$ or 6.

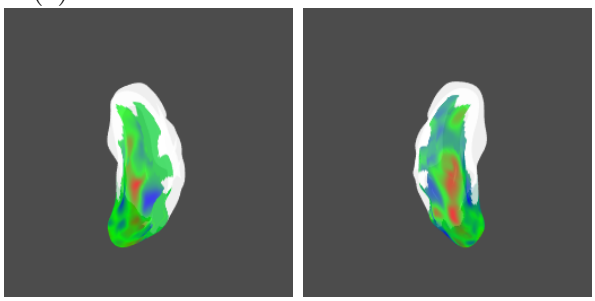
(a) Cerebral surface map



(b) Spherical map



(c) Selected Parcels on cerebral surface



(d) Selected Parcels on spherical map

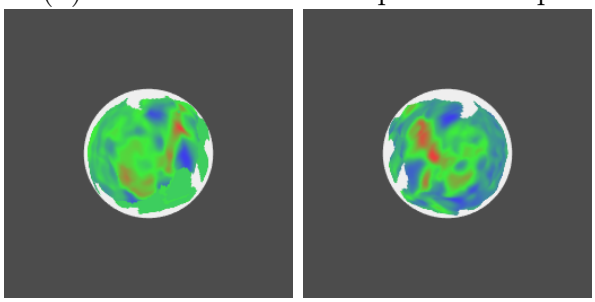


Figure 4.3: Example outcome images from a subject (id=151627) shown on cerebral surface (a) and spherical surface (b) of the whole left (left column) and right brain (right column). The cerebral (c) and spherical (d) surface images show the observed outcome images within 29 selected parcels in the right brain (left column) and another 18 parcels in the left brain (right column).

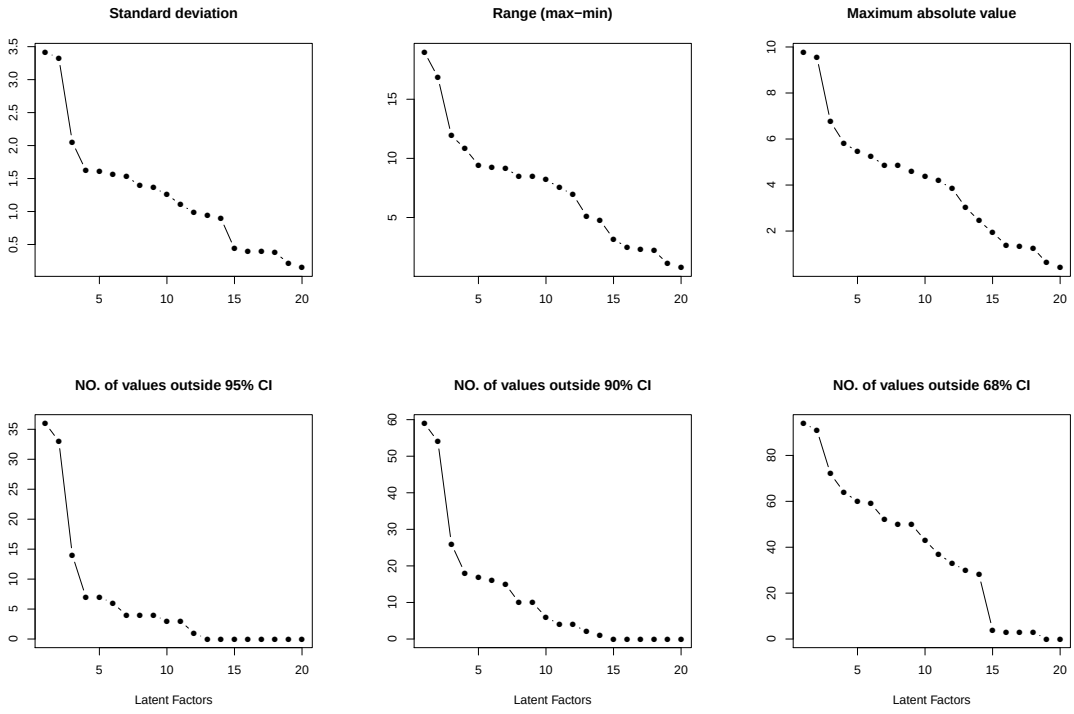
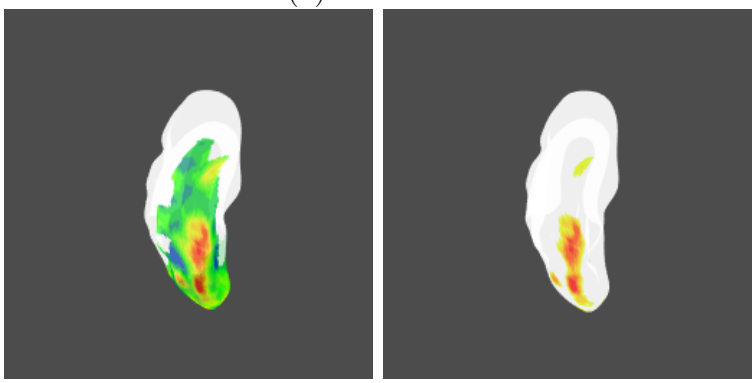
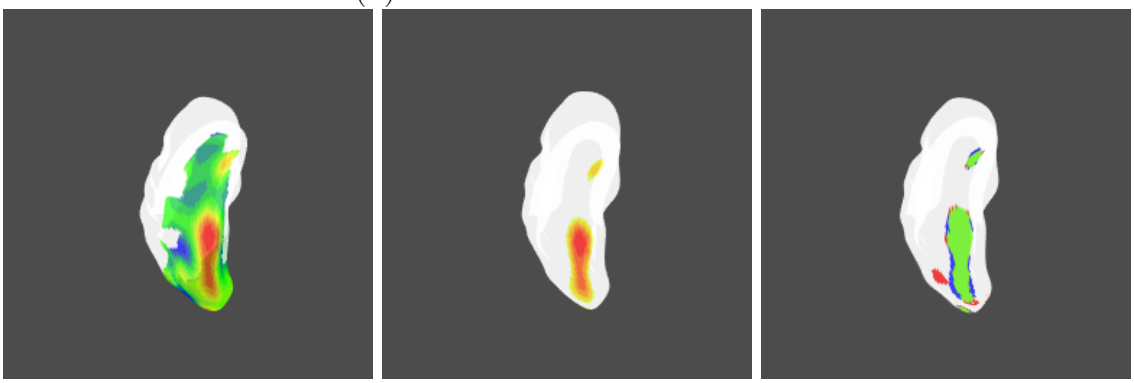


Figure 4.4: Summary statistics of the posterior mean of loading matrix estimated from SBLF with $K = 20$ in HCP application study. X-axis is the index of latent factors from 1 to 20. Figures on top row are for statistics including standard deviation, range (max value - min value) and maximum absolute value of each estimated loading vector, respectively. Figures on the bottom show the number of values in each loading vector outside the 95%, 90% and 68% confidence interval (CI) of the whole estimated loading matrix. The determined optimal value of latent factors is $K = 10$.

(a) Real data



(b) estimations of SBLF model



(c) Estimations of ridge regression model

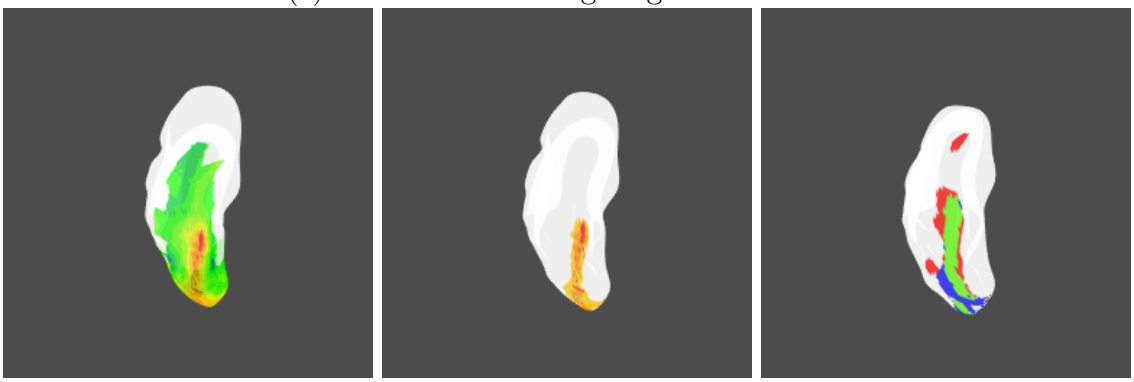


Figure 4.5: Examples of the observed outcome (subject id = 151627) images and his/her estimated outcome images by SBLF and ridge regression models fitted within the 29 selected surface parcels. The first and second plots in each row are the outcome images shown in 29 parcels and only selected active vertices (with the top 10% of the absolute values as the threshold, resulting in 489 active vertices), respectively. The last image in the second and third row, contains three types of selected active vertices, (1) correctly selected active vertices (green), (2) false active vertices (blue, non-active vertices in observed image but identified as active vertices by fitted models), and (3) misspecified active vertices (red, active vertices in observed image but not identified as active vertices by fitted models). The number of the three types of vertices are (386, 103, 103) and (282, 207, 207) using SBLF and ridge regression model, respectively.

Table 4.1: Results of simulation Scenario 1 and 2, including mean value and range [min, max] of RMSE/RMSPE and R-Squared values (averaged across subjects and simulations for each vertex in outcome image) for training and validation sets, respectively.

Generating Model	Analysis Method	Training Set		Validation Set	
		RMSE	R-Squared	RMSPE	R-Squared
Scenario 1	SBLF	0.189 [0.146, 0.249]	0.957 [0.749, 0.996]	0.600 [0.353, 1.069]	0.715 [0.047, 0.948]
	Ridge	0.251 [0.370, 1.636]	0.829 [0.000, 0.998]	1.005 [0.370, 2.062]	0.342 [0.011, 0.647]
Scenario 2	SBLF	0.471 [0.363, 0.590]	0.904 [0.768, 0.987]	1.356 [1.008, 2.143]	0.303 [0.000, 0.863]
	Ridge	0.756 [0.044, 2.086]	0.614 [0.000, 0.999]	1.577 [1.066, 3.028]	0.148 [0.000, 0.600]

SBLF: our proposed spatial Bayesian latent factor model.
RMSE/RMPE: root mean squared errors of estimations of training sets and predictions of validation set, respectively.

Table 4.2: Model performance of our SBLF and ridge regression models for HCP real data analysis. Our proposed SBLF model are fitted multiple times with candidate values of SH degrees (L = 5, 10, 15 or 20) and the number of latent factors (K=10, 20, 30). For each model fitting, results include the averaged values and range [min, max] of RMSE/RMSPE and R-Squared values at every vertex $s \in \mathbb{S}^2$ for the training and validation sets, averaged across subjects and 10-folds CV . The optimal choice is (L=10, K=10).

Model	SH Degree	Latent Factors	Training Set		Validation Set	
			RMSE	R-Squared	RMSPE	R-Squared
Ridge Regression	-	-	2.380 [0.060, 5.290]	0.208 [0.000, 0.998]	2.647 [0.565, 5.928]	0.102 [0.000, 0.870]
SBLF	L=5	K=10	2.224 [0.677, 6.225]	0.517 [0.044, 0.869]	3.444 [0.589, 10.674]	0.125 [0.000, 0.848]
		K=20	2.242 [0.677, 6.334]	0.516 [0.043, 0.891]	3.371 [0.635, 10.426]	0.124 [0.000, 0.827]
		K=30	2.228 [0.639, 6.276]	0.519 [0.042, 0.878]	3.493 [0.553, 10.862]	0.117 [0.000, 0.790]
	L=10	**K=10**	**1.705 [0.559, 4.022]**	**0.670 [0.141, 0.936]**	**3.299 [0.569, 9.756]**	**0.127 [0.000, 0.871]**
		K=20	1.674 [0.533, 4.007]	0.685 [0.156, 0.950]	3.273 [0.543, 9.535]	0.109 [0.000, 0.816]
		K=30	1.674 [0.540, 4.042]	0.687 [0.161, 0.964]	3.293 [0.578, 9.589]	0.100 [0.000, 0.716]
	L=15	K=10	1.371 [0.387, 3.213]	0.772 [0.187, 0.963]	3.281 [0.566, 9.432]	0.120 [0.000, 0.772]
		K=20	1.339 [0.372, 3.163]	0.784 [0.206, 0.967]	3.627 [0.536, 10.907]	0.099 [0.000, 0.749]
		K=30	1.328 [0.363, 3.162]	0.788 [0.201, 0.969]	3.429 [0.571, 10.273]	0.107 [0.000, 0.868]
	L=20	K=10	1.098 [0.295, 2.777]	0.847 [0.378, 0.973]	3.333 [0.637, 9.299]	0.113 [0.000, 0.923]
		K=20	1.066 [0.306, 2.721]	0.856 [0.397, 0.976]	3.783 [0.540, 12.032]	0.074 [0.000, 0.707]
		K=30	1.051 [0.277, 2.706]	0.860 [0.409, 0.977]	3.488 [0.549, 10.384]	0.110 [0.000, 0.833]

SBLF: our proposed spatial Bayesian latent factor model.
RMSE/RMPE: root mean squared errors of estimations of training sets and predictions of validation set, respectively.

Table 4.3: Results of HCP data analysis using SBLF and ridge regression models to identify active vertices. A sequence of thresholds, (top 1%, 5%, 10%, 20% and 50% of the absolute values), are used to define the active points in the observed, estimated and predicted outcome images, respectively. The number and proportion of active vertices correctly identified by SBLF and ridge regression methods are averaged over subjects and 10-folds CV, respectively. The table also include the metric N^{subj}(SBLF), the number and proportion of subjects who have more correctly identified active vertices by our SBLF method than ridge regression model.

Threshold	Training Set			test Set		
	SBLF	Ridge	N^{subj}(SBLF)	SBLF	Ridge	N^{subj}(SBLF)
Top 1%	11	12	37	3	4	4
(N^s=49)	(22.49%)	(24.35%)	(42.05%)	(6.94%)	(7.75%)	(40.00%)
Top 5%	136	115	64	90	102	3
(N^s=245)	(55.41%)	(46.84%)	(72.72%)	(36.82%)	(41.71%)	(30.00%)
Top 10%	344	303	74	269	283	1
(N^s=489)	(70.37%)	(62.00%)	(84.09%)	(54.91%)	(57.96%)	(10.00%)
Top 20%	750	684	81	620	614	6
(N^s=977)	(76.77%)	(70.09%)	(92.05%)	(63.46%)	(62.86%)	(60.00%)
Top 50%	1940	1740	87	1628	1622	5
(N^s=2441)	(79.46%)	(71.32%)	(98.86%)	(66.69%)	(66.46%)	(50.00%)

SBLF: our proposed spatial Bayesian latent factor model.
Ridge: ridge regression method.
Threshold: the top percentage of the absolute values.
N^s: the number of vertices with absolute values above the threshold placed on the observed outcome images.
N^{subj}(SBLF): the number of subjects who have more active vertices correctly identified by SBLF model than ridge regression method.

CHAPTER V

Conclusion

In this dissertation, we develop two spatial Bayesian models for scalar-on-image and image-on-image regression, respectively. We account for complex spatial correlations in high-dimensional images and use efficient sampling algorithms, fast computational techniques and dimension reduction methods.

In Chapter II, we propose a Bayesian scalar-on-image regression model to predict clinical subtypes of MS using MRI lesion images. Both non-spatially and spatially varying variables can be used as predictors in the logit link function for multinomial logistic regression model. The spatial correlations existing in brain image space are measured through a GP prior distribution and estimated under the Bayesian framework by using HMC algorithm. In addition, the application of FFT algorithm make it feasible and efficient to manipulate the high-dimensional correlation matrix. Both simulation study and the application analysis of MS data prove the high prediction accuracy of our method. In practical applications, our method is able to help doctors to determine the subtypes for MS patients with their MRI lesion images at a single time point instead of monitoring the disease progression for a very long time.

In Chapter III and IV, we develop a fully Bayesian hierarchical spatial model for image-on-image regression. We introduce low-dimensional latent factors as a bridge linking high-dimensional outcome and predictor images. Those latent factors serve as

various scalar measures summarized from predictor images. Furthermore, we assign flexible prior distributions to capture the complex and various spatial dependence within multimodal outcome and predictor images. In particular, we use isotropic Gaussian kernels and SH basis functions to approximate cortical surface-based and volumetric images. In addition, we resort to inducing prior distributions for the sparse loading elements through the parameter expansion approach, leading to more efficient posterior computation. In multiple simulation studies, our proposed spatial Bayesian latent factor models achieve better performance in prediction accuracy than linear regression and voxel-wise regression methods (in cases where outcome and predictors share the same imaging space), and ridge-regression method (in the case of various imaging spaces). We also apply our model to predict task-evoked images in Emotion task domain from subcortical seed-based maps (within amygdala region) from resting-state fMRI given data from HCP.

Given multiple neuroimaging modalities, our methods are limited by assuming the same spatial correlation structure for all imaging predictors. In the future, we plan to expand the model involving a set of various spatially varying coefficients $\{\beta_{kq}(v)\}$, where q is the index of different spatial dependence. Furthermore, we can also include corresponding latent indicator of important predictors $\{\boldsymbol{\gamma}_{pq}\}$ to each spatial structure. These expansions may lead to more explicit measures of predictor images and the selection of grouped predictor images with similar structure or function connectivity. However, the newly introduced parameters will result in extra computational task so that other efficient methods or sampling algorithm are in need for estimating spatially varying parameters.

Our proposed methods can be used to either the whole image space or any single region/parcel of interest, however, ignore the spatial correlations among regions for independent region-wise analyses. Hence, another promising future direction is to expand the spatial Bayesian latent factor model to a hierarchical model for the

integration analysis of multiple regions with both within-region and region-wise spatial effects or correlation structures. We will benefit from this future work in the analysis of brain activity and functional connectivity among multiple neuroimaging modalities.

APPENDICES

APPENDIX A

Appendices of Chapter III

A.1 Gradient of Log-Joint-Posterior

To implement HMC algorithm using Leapfrog method for our model, we need to calculate the log-joint-posterior density and its gradient. By Brook's lemma *Besag* (1974b), the joint distribution (2.3), up to a constant of proportionality, can be written as

$$
\begin{aligned}
P(\boldsymbol{\theta}_k|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\theta}_{-k},\boldsymbol{\Omega}) \propto & \Big[\prod_{i=1}^{N}\prod_{l=0}^{K-1}\pi_{il}^{I(y_i=l)}\Big] \times \exp\left(-\frac{1}{2}\alpha_k^2\right) \times \exp\Big(-\frac{1}{2}\boldsymbol{\zeta}_k^T\boldsymbol{\zeta}_k\Big) \\
& \times |\boldsymbol{\Sigma}_k|^{-1/2}\exp\Big(-\frac{1}{2}(\boldsymbol{\gamma}_k-\boldsymbol{U}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\gamma}_k-\boldsymbol{U}_k)\Big) \\
& \times \sigma_k^{a_\sigma-1}\exp\left(-b_\sigma\sigma_k\right) \times \rho_k^{a_\rho-1}\exp\left(-b_\rho\rho_k\right) \\
& \times \exp\Big(-\frac{1}{2}(\boldsymbol{U}_k-\boldsymbol{\mu}_0)^T\boldsymbol{\Lambda}_0^{-1}(\boldsymbol{U}_k-\boldsymbol{\mu}_0)\Big) \\
& \times |\boldsymbol{\Sigma}_k^{-1}|^{-\frac{\nu_0+p+1}{2}}\exp\Big(-\frac{1}{2}\mathrm{Tr}(\boldsymbol{\Phi}_0\boldsymbol{\Sigma}_k)\Big) \qquad \text{(A.1)}
\end{aligned}
$$

where

$$\pi_{ik} = P(y_i = k) = \frac{\exp\left[\alpha_k + \mathbf{z}_i^T\gamma_k + \sigma_k \boldsymbol{X}_i^T \boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{\zeta}_k\right]}{1 + \sum_{l=1}^{K-1}\exp\left[\alpha_l + \mathbf{z}_i^T\gamma_l + \sigma_l \boldsymbol{X}_i^T \boldsymbol{C}^{\frac{1}{2}}(\rho_l)\boldsymbol{\zeta}_l\right]} \tag{A.2}$$

$$I(y_i = k) = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{if } y_i \neq k \end{cases}$$

Then, the log-joint-posterior density is

$$\begin{aligned}\log P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) \propto & \left[\sum_{i=1}^{N}\sum_{l=0}^{K-1} I(y_i = l)\log(\pi_{il})\right] - \frac{1}{2}\alpha_k^2 - \frac{1}{2}\boldsymbol{\zeta}_k^T\boldsymbol{\zeta}_k \\ & - \frac{1}{2}(\boldsymbol{\gamma}_k - \boldsymbol{U}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\gamma}_k - \boldsymbol{U}_k) \\ & + (a_\sigma - 1)\log(\sigma_k) - b_\sigma\sigma_k + (a_\rho - 1)\log(\rho_k) - b_\rho\rho_k.\end{aligned}$$

Then, the gradients of the log-joint-posterior density are derived as:

$$\nabla_{\alpha_k}\log P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) \propto \left[\sum_{i=1}^{N}\big(I(y_i = k) - \pi_{ik}\big)\right] - \alpha_k$$

$$\nabla_{\boldsymbol{\zeta}_k}\log P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) \propto \left[\sum_{i=1}^{N}\big(I(y_i = k) - \pi_{ik}\big)\sigma_k\boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{X}_i\right] - \boldsymbol{\zeta}_k$$

$$\nabla_{\gamma_k}\log P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) \propto \left[\sum_{i=1}^{N}\big(I(y_i = k) - \pi_{ik}\big)\boldsymbol{Z}_i\right] - \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{\gamma_k} - \boldsymbol{U}_k)$$

$$\nabla_{\sigma_k}\log P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) \propto \left[\sum_{i=1}^{N}\big(I(y_i = k) - \pi_{ik}\big)\boldsymbol{X}_i^T\boldsymbol{C}^{\frac{1}{2}}(\rho_k)\boldsymbol{\zeta}_k\right] + \frac{a_\sigma - 1}{\sigma_k}$$

$$\nabla_{\rho_k}\log P(\boldsymbol{\theta}_k|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}_{-k}, \boldsymbol{\Omega}) \propto \left[\sum_{i=1}^{N}\big(I(y_i = k) - \pi_{ik}\big)\sigma_k\boldsymbol{X}_i^T\boldsymbol{\zeta}_k\right] + \frac{a_\rho - 1}{\rho_k}$$

A.2 Constraints on σ and ρ

Since σ_k and ρ_k have to be positive, some constraints must be considered while updating the state at each time in (**??**). For example, given upper and lower bounds

of $\boldsymbol{\theta}_k$ denoted as $\boldsymbol{u}$ and $\boldsymbol{l}$, we repeat following steps until $\boldsymbol{u} \leq \boldsymbol{\theta}_k(t+\delta) \leq \boldsymbol{l}$

- if $\boldsymbol{\theta}_k(t+\delta) > \boldsymbol{u}$, then

$$\boldsymbol{\theta}_k(t+\delta) = \boldsymbol{u} - \big(\boldsymbol{\theta}_k(t+\delta) - \boldsymbol{u}\big) \text{ and } \boldsymbol{\xi}_k(t+\delta/2) = -\boldsymbol{\xi}_k(t+\delta/2)$$

- if $\boldsymbol{\theta}_k(t+\delta) < \boldsymbol{l}$, then

$$\boldsymbol{\theta}_k(t+\delta) = \boldsymbol{l} + \big(\boldsymbol{l} - \boldsymbol{\theta}_k(t+\delta)\big) \text{ and } \boldsymbol{\xi}_k(t+\delta/2) = -\boldsymbol{\xi}_k(t+\delta/2)$$

A.3 Posterior Estimates of ρ, σ and α

Table A.1: Posterior estimates (mean, standard deviation (SD) and 95% Confidence Interval (CI)) of ρ, σ and α.

Parameter	MS subtype	Mean	SD	95% CI
ρ	PRP MS	4.50	1.69	[1.19, 7.81]
	SCP MS	7.98	2.11	[3.85, 12.12]
σ	PRP MS	2.81	0.60	[1.64, 3.99]
	SCP MS	2.92	0.60	[1.74, 4.10]
α	PRP MS	-0.31	1.04	[-2.35, 1.73]
	SCP MS	-0.39	0.96	[-2.28, 1.49]

A.4 Trace plots and ACF plots of ρ, σ, α and β

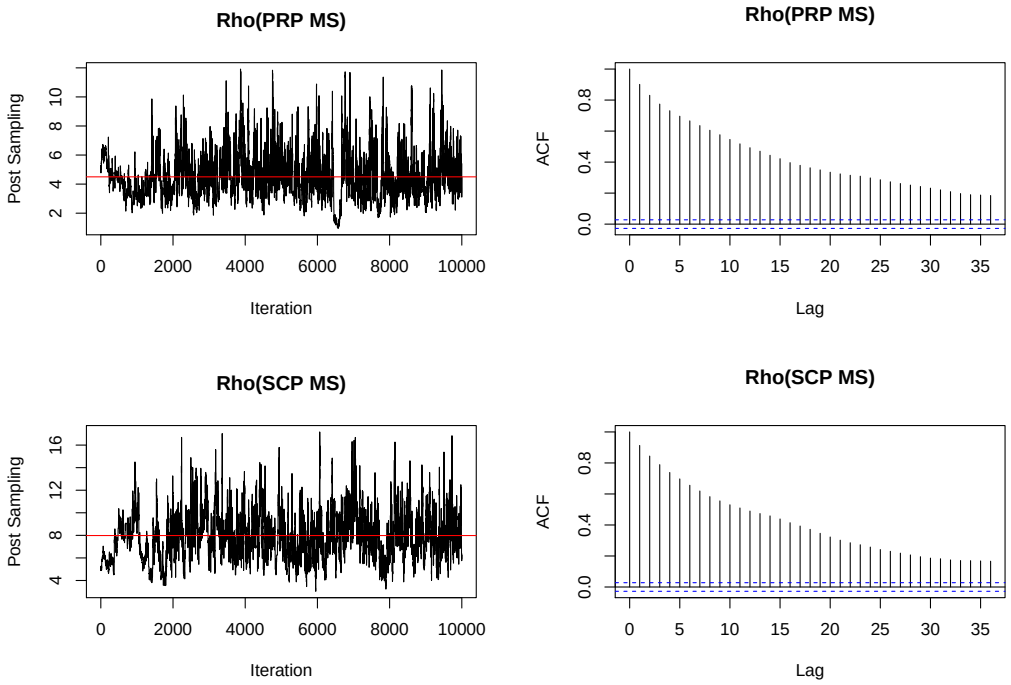


Figure A.1: Trace plots and ACF plots of ρ for PRP (first row) and SCP (second row) MS subtype groups, respectively.

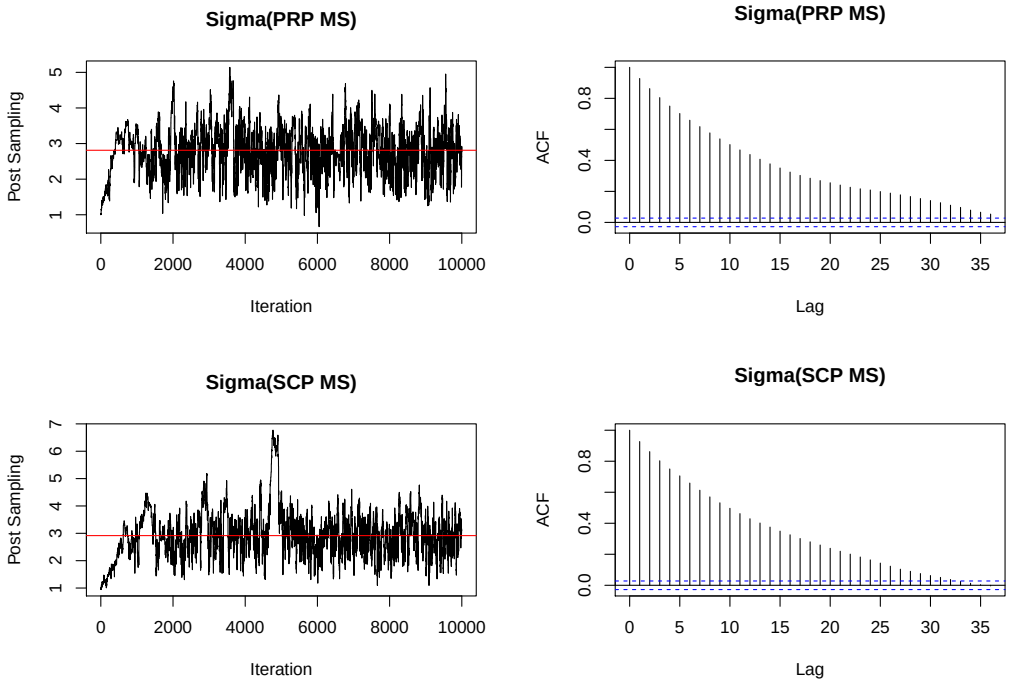


Figure A.2: Trace plots and ACF plots of σ for PRP (first row) and SCP (second row) MS subtype groups, respectively.

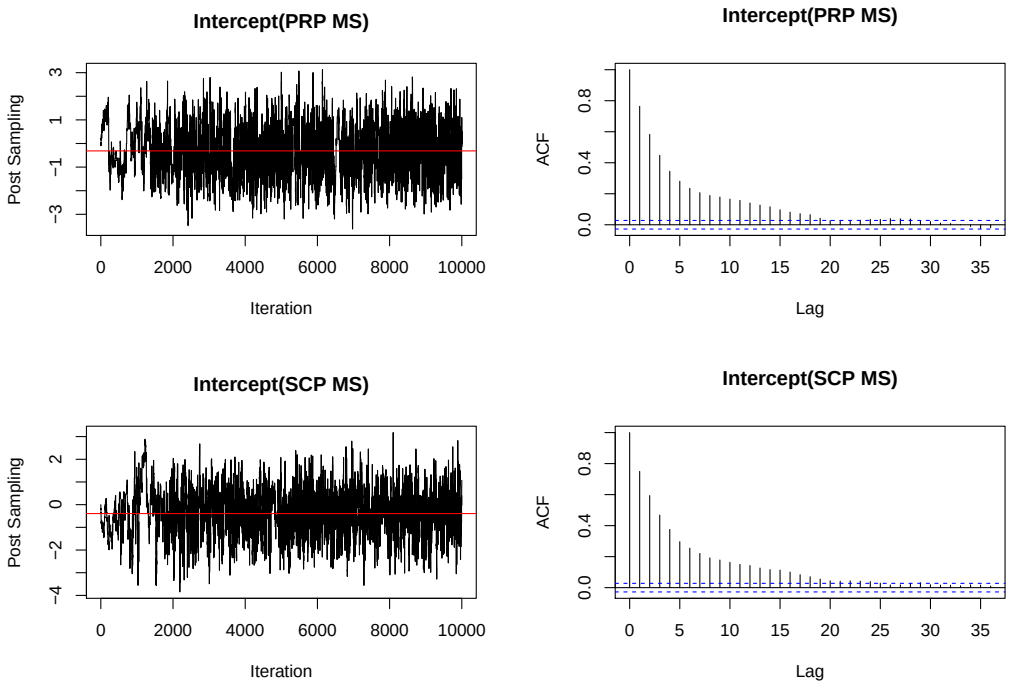


Figure A.3: Trace plots and ACF plots of α for PRP (first row) and SCP (second row) MS subtype groups, respectively.

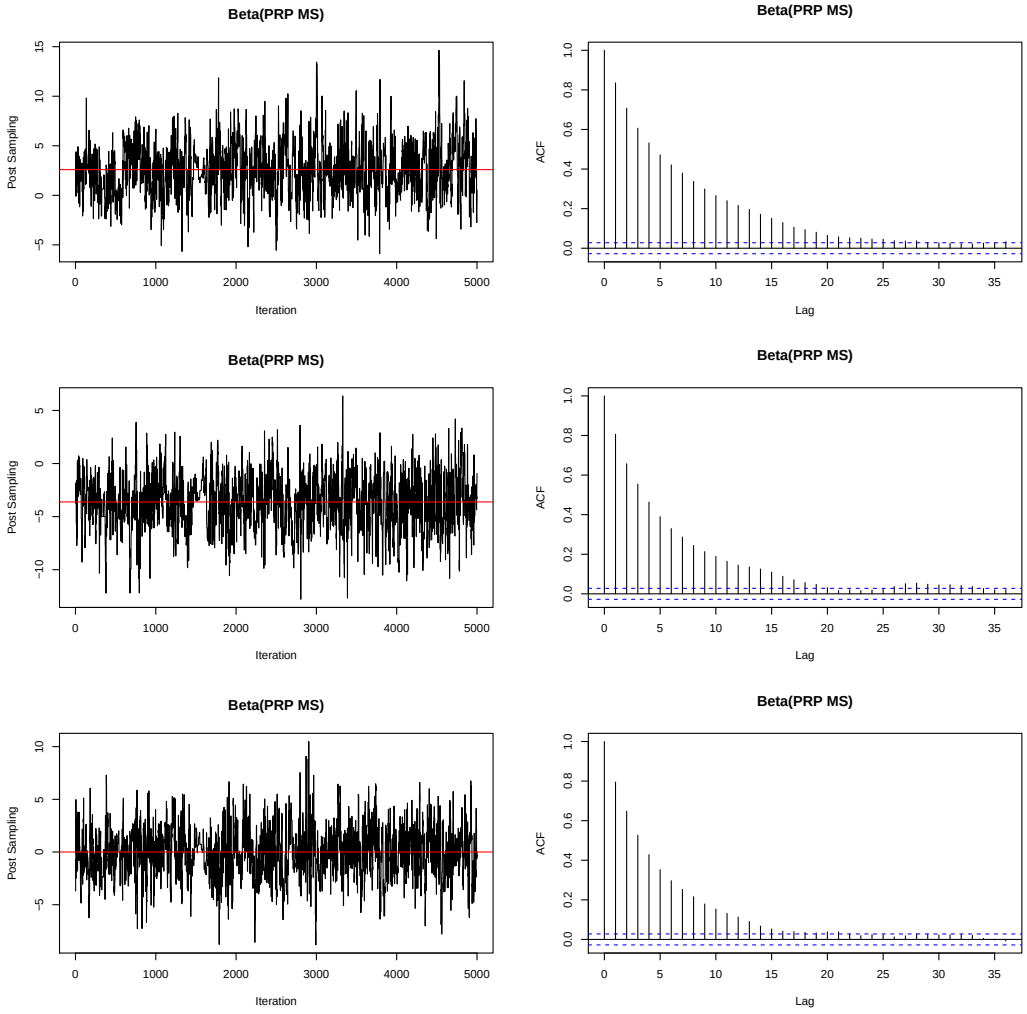


Figure A.4: Trace plots and ACF plots of α for PRP MS subtype group.

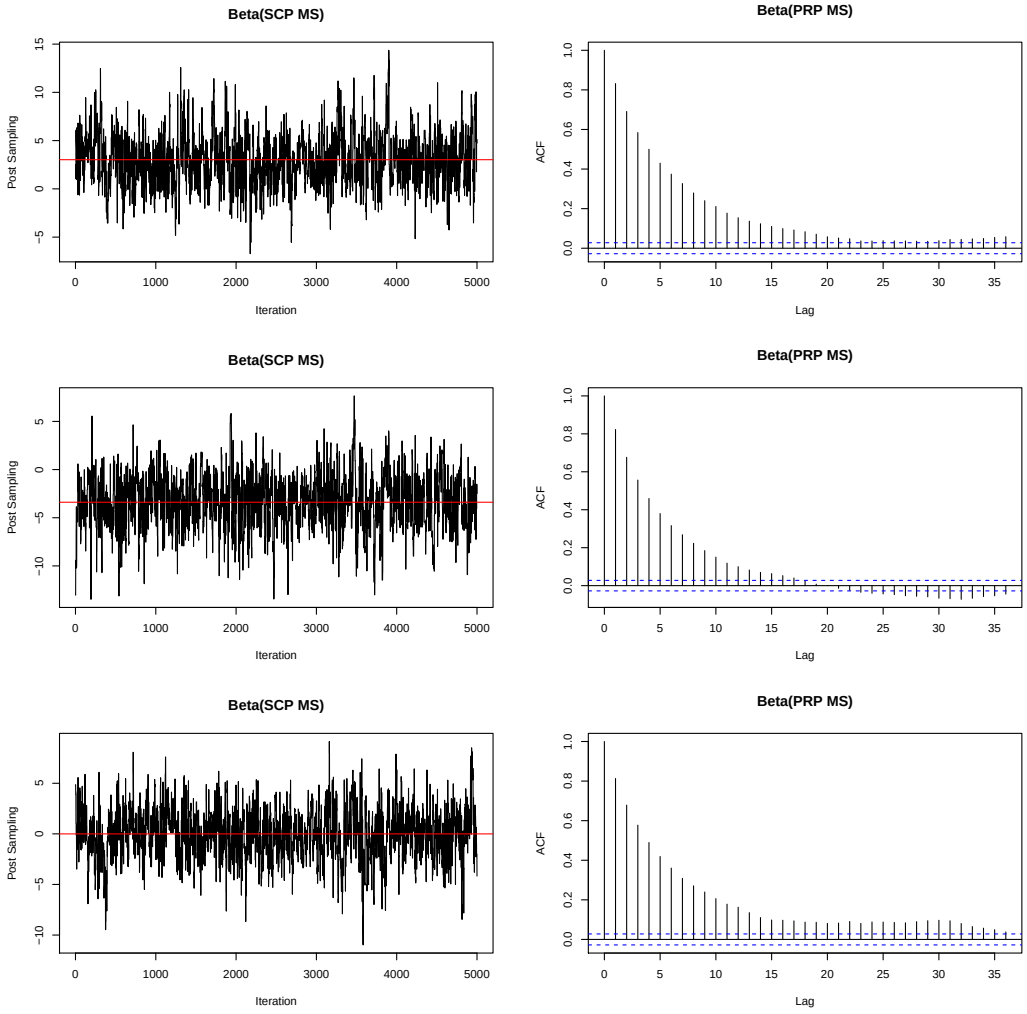


Figure A.5: Trace plots and ACF plots of α for SCP MS subtype group.

APPENDIX B

Appendices of Chapter IV

B.1 Full Model

Our hierarchical full model is:

$$\text{Level 1}: \quad Z_i(v) = U(v) + \sum_{m=1}^{M} \theta_{im} b_m(v) + e_i(v)$$

$$\text{Level 2}: \quad \theta_{im} = \sum_{k=1}^{K} \lambda_{mk} \eta_{ik} + \zeta_{im}$$

$$\text{Level 3}: \quad \eta_{ik} = \sum_{v' \in \mathcal{R}} \widetilde{X_i}(v') \beta_k(v') + \epsilon_{ik}$$

$$\text{where} \quad \widetilde{X_i}(v') = \sum_{p=1}^{P} \gamma_p X_{ip}(v'), \quad \beta_k(v') = \sum_{m=1}^{M} \alpha_{km} b_m(v')$$

Combining the two sub-models on Level 1 and 2, we get

$$\begin{aligned}
Z_i(v) =& U(v) + \sum_{m=1}^{M} \Big[\sum_{k=1}^{K} \lambda_{mk}\eta_{ik} + \zeta_{im} \Big] b_m(v) + e_i(v) \\
=& U(v) + \sum_{m=1}^{M} \sum_{k=1}^{K} \lambda_{mk}\eta_{ik} b_m(v) + \sum_{m=1}^{M} \zeta_{im} b_m(v) + e_i(v) \\
=& U(v) + \sum_{k=1}^{K} \Big[\sum_{m=1}^{M} \lambda_{mk} b_m(v) \Big] \eta_{ik} + \tilde{\zeta}_i(v) + e_i(v) \\
=& U(v) + \sum_{k=1}^{K} \tilde{\lambda}_k(v)\eta_{ik} + \tilde{\zeta}_i(v) + e_i(v) \qquad \text{(B.1)}
\end{aligned}$$

where the spatially varying prediction effect on response voxel v from predictor voxel v' is

$$\tilde{\lambda}_k(v) = \sum_{m=1}^{M} \lambda_{mk} b_m(v), \qquad \tilde{\zeta}_i(v) = \sum_{m=1}^{M} \zeta_{im} b_m(v)$$

By plugging equations on Level 3 in equation (B.1), we get

$$\begin{aligned}
Z_i(v) =& U(v) + \sum_{k=1}^{K} \tilde{\lambda}_k(v)\eta_{ik} + \tilde{\zeta}_i(v) + e_i(v) \\
=& U(v) + \sum_{k=1}^{K} \tilde{\lambda}_k(v) \Big[\sum_{v' \in \mathcal{R}} \Big(\sum_{p=1}^{P} \gamma_p X_{ip}(v') \Big) \beta_k(v') \Big] + \\
& \sum_{k=1}^{K} \tilde{\lambda}_k(v)\epsilon_{ik} + \tilde{\zeta}_i(v) + e_i(v) \\
=& U(v) + \sum_{p=1}^{P} \gamma_p \sum_{v' \in \mathcal{R}} \Big[\sum_{k=1}^{K} \tilde{\lambda}_k(v)\beta_k(v') \Big] X_{ip}(v') + \tilde{\epsilon}_i(v) + \tilde{\zeta}_i(v) + e_i(v) \\
=& U(v) + \sum_{p=1}^{P} \gamma_p \sum_{v' \in \mathcal{R}} \psi(v, v') X_{ip}(v') + \tilde{\epsilon}_i(v) + \tilde{\zeta}_i(v) + e_i(v)
\end{aligned}$$

where

$$
\begin{aligned}
\boldsymbol{\psi}(v, v') &= \sum_{k=1}^{K} \tilde{\lambda}_k(v)\beta_k(v') \\
&= \sum_{k=1}^{K}\sum_{m=1}^{M} \lambda_{mk} b_m(v)\beta_k(v') \\
&= \sum_{m=1}^{M} \Big[\sum_{k=1}^{K} \lambda_{mk}\beta_k(v') \Big] b_m(v) \\
&= \sum_{m=1}^{M} \Big[\sum_{k=1}^{K} \lambda_{mk} \Big(\sum_{m'=1}^{M} \alpha_{km'} b_{m'}(v') \Big) \Big] b_m(v) \\
&= \sum_{m=1}^{M}\sum_{m'=1}^{M} \Big[\sum_{k=1}^{K} \lambda_{mk}\alpha_{km'} \Big] b_{m'}(v') b_m(v) \\
&= \sum_{k=1}^{K} \left\{ \Big[\sum_{m=1}^{M} \lambda_{mk} b_m(v) \Big] \times \Big[\sum_{m'=1}^{M} \alpha_{km'} b_{m'}(v') \Big] \right\}
\end{aligned}
$$

and

$$
\tilde{\epsilon}_i(v) = \sum_{k=1}^{K} \tilde{\lambda}_k(v)\epsilon_{ik} = \sum_{m=1}^{M}\sum_{k=1}^{K} \lambda_{mk}\epsilon_{ik} b_m(v)
$$

B.2 Full Conditional Posterior Distributions

Elements in the lower triangular part of the working loading matrix, denoted $\boldsymbol{\lambda}^*_m = \{\lambda^*_{m1}, \lambda^*_{m2}, \ldots, \lambda^*_{mq}\}^T$ with $q = \min(m, K)$, are sampled from a multivariate normal distribution that

$$
\begin{aligned}
\pi(\boldsymbol{\lambda}^*_m|\cdot) &\sim \mathrm{N}_q(\mathrm{Mean}, \mathrm{Cov}) \\
\mathrm{Mean} &= \mathrm{Cov} \times \sigma_\zeta^{-2} \sum_{i=1}^{N} \theta_{im} \boldsymbol{\eta}^*_{i_m} \\
\mathrm{Cov} &= \Big[\sum_{i=1}^{N} \big(\sigma_\zeta^{-2} \boldsymbol{\eta}^*_{i_m} \boldsymbol{\eta}^{*^T}_{i_m} \big) + \sigma_\lambda^{-2} \boldsymbol{I}_q \Big]^{-1}
\end{aligned}
$$

where $\boldsymbol{\eta}^*_{i_m} = (\eta^*_{i1}, \eta^*_{i2}, \ldots, \eta^*_{iq})^T$, the corresponding latent vector to $\boldsymbol{\lambda}^*_m$. Other working loading matrix elements in the upper triangular part are set to be 0.

The subject-specific working latent vector $\boldsymbol{\eta}^*_i = (\eta^*_{ik})_{K\times 1} = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iK})^T$ is sampled from a posterior multivariate normal distribution that

$$\begin{aligned}
\pi(\boldsymbol{\eta}^*_i|\cdot) &\sim \mathrm{N}_K\big(\mathrm{Mean}, \mathrm{Cov}\big),\\
\mathrm{Mean} = \mathrm{Cov} &\times \Big\{\sigma_\zeta^{-2}\boldsymbol{\Lambda}^{*^T}\boldsymbol{\theta}_i + \sigma_\epsilon^{-2}\boldsymbol{\Phi}\big(\boldsymbol{\mu}^*_i + \boldsymbol{\beta}^{*^T}\tilde{\boldsymbol{X}}_i\big)\Big\}\\
\mathrm{Cov} &= \Big(\sigma_\zeta^{-2}\boldsymbol{\Lambda}^{*^T}\boldsymbol{\Lambda}^* + \sigma_\epsilon^{-2}\boldsymbol{\Phi}\Big)^{-1}
\end{aligned}$$

where we write $\boldsymbol{\Lambda}^* = (\lambda^*_{mk})_{M\times K}$, $\boldsymbol{\theta}_i = (\theta_{im})_{M\times 1}$, $\boldsymbol{\Phi} = \mathrm{diag}\{\phi^2_1, \phi^2_2, \ldots, \phi^2_K\}$, $\boldsymbol{\mu}^*_i = (\mu^*_{ik})_{K\times 1}$, $\boldsymbol{\beta}^* = \{\beta^*_k(v)\}_{|\mathcal{R}|\times K}$, $\tilde{\boldsymbol{X}}_i = \{\tilde{X}_i(v)\}_{|\mathcal{R}|\times 1}$ and let $|\mathcal{R}|$ represent the number of voxels in $\mathcal{R}$. Similarly, the extra K-length intercept vector $\boldsymbol{\mu}^*_i$ is sampled from a multivariate normal distribution that

$$\pi(\boldsymbol{\mu}^*_i|\cdot) \sim \mathrm{N}_K\Big\{\sigma_\epsilon^{-2}\boldsymbol{\Phi}\big(\sigma_\epsilon^{-2}\boldsymbol{\Phi} + \sigma_\mu^{-2}\boldsymbol{I}\big)^{-1}\big(\boldsymbol{\eta}^*_i - \boldsymbol{\beta}^{*^T}\tilde{\boldsymbol{X}}_i\big),\quad \big(\sigma_\epsilon^{-2}\boldsymbol{\Phi} + \sigma_\mu^{-2}\boldsymbol{I}_M\big)^{-1}\Big\}$$

The working basis coefficient vector $\boldsymbol{\alpha}^*_k$ for approximating $\boldsymbol{\beta}^*_k$ has the following full conditional posterior distribution that

$$\begin{aligned}
\pi(\boldsymbol{\alpha}^*_k|\cdot) &\sim \mathrm{N}_M(\mathrm{Mean}, \mathrm{Cov})\\
\mathrm{Mean} = \mathrm{Cov} &\times \sigma_\epsilon^{-2}\phi^2_k\boldsymbol{b}^T\Big(\sum_{i=1}^N\big(\eta^*_{ik} - \mu^*_{ik}\big)\tilde{\boldsymbol{X}}_i\Big)\\
\mathrm{Cov} &= \Big[\sigma_\epsilon^{-2}\phi^2_k\boldsymbol{b}^T\big(\sum_{i=1}^N\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T\big)\boldsymbol{b} + \sigma_\alpha^{-2}\boldsymbol{I}_M\Big]^{-1}
\end{aligned}$$

where $\boldsymbol{\alpha}^*_k = (\alpha_{mk})_{M\times 1}$ and $\boldsymbol{b} = \{b_m(v)\}_{|\mathcal{R}|\times M}$.

The reciprocal of diagonal element ϕ^2_k in the working matrix $\boldsymbol{\Phi}$ is sampled from a

Gamma full conditional posterior distribution that

$$\pi(\phi_k^{-2}|\cdot) \sim \text{Gamma}\left(a_\phi + \frac{N}{2}, \quad b_\phi + \frac{1}{2}\sigma_\epsilon^{-2}\sum_{i=1}^{N}\left(\eta_{ik}^* - \mu_{ik}^* - \tilde{\boldsymbol{X}}_i^T\boldsymbol{\beta}_k^*\right)^2\right)$$

Here we list the full conditional posterior distributions for other parameters outside the inferential models.

$$\pi(\boldsymbol{U}|\cdot) \sim \text{N}_{|\mathcal{R}|}\left(\frac{\sigma_e^{-2}}{N\sigma_e^{-2}+\sigma_u^{-2}}\sum_{i=1}^{N}\left(\boldsymbol{Z}_i - \boldsymbol{b}\boldsymbol{\theta}_i\right), \ \left(N\sigma_e^{-2}+\sigma_u^{-2}\right)^{-1}\boldsymbol{I}_{|\mathcal{R}|}\right)$$

$$\pi(\sigma_u^{-2}|\cdot) \sim \text{Gamma}\left(a_u + \frac{1}{2}, \ b_u + \frac{1}{2}\boldsymbol{U}^T\boldsymbol{U}\right)$$

$$\pi(\boldsymbol{\theta}_i|\cdot) \sim \text{N}_M\left(\left[\sigma_e^{-2}\boldsymbol{b}^T\boldsymbol{b} + \sigma_\zeta^{-2}\boldsymbol{I}_M\right]^{-1}\left(\sigma_e^{-2}\boldsymbol{b}^T(\boldsymbol{Z}_i - \boldsymbol{U}) + \sigma_\zeta^{-2}\boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^*\right), \left[\sigma_e^{-2}\boldsymbol{b}^T\boldsymbol{b} + \sigma_\zeta^{-2}\boldsymbol{I}_M\right]^{-1}\right)$$

$$\pi(\sigma_e^{-2}|\cdot) \sim \text{Gamma}\left(a_e + \frac{QN}{2}, \quad b_e + \frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{Z}_i - \boldsymbol{U} - \boldsymbol{b}\boldsymbol{\theta}_i\right)^T\left(\boldsymbol{Z}_i - \boldsymbol{U} - \boldsymbol{b}\boldsymbol{\theta}_i\right)\right)$$

$$\pi(\sigma_\zeta^{-2}|\cdot) \sim \text{Gamma}\left(a_\zeta + \frac{MN}{2}, \quad b_\zeta + \frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\theta}_i - \boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^*\right)^T\left(\boldsymbol{\theta}_i - \boldsymbol{\Lambda}^*\boldsymbol{\eta}_i^*\right)\right)$$

$$\pi(\sigma_\epsilon^{-2}|\cdot) \sim \text{Gamma}\left(a_\epsilon + \frac{KN}{2}, \quad b_\epsilon + \frac{1}{2}\sum_{i=1}^{N}\left[\boldsymbol{\eta}_i^* - \boldsymbol{\mu}_i^* - \boldsymbol{\beta}^{*^T}\tilde{\boldsymbol{X}}_i\right]^T\boldsymbol{\Phi}\left[\boldsymbol{\eta}_i^* - \boldsymbol{\mu}_i^* - \boldsymbol{\beta}^{*^T}\tilde{\boldsymbol{X}}_i\right]\right)$$

$$\pi(\gamma_p|\cdot) \sim \text{Bernoulli}\left(\frac{c_1}{c_0 + c_1}\right)$$

$$c_1 = \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sigma_\epsilon^{-2}\phi_k^2\left(\tau_{ik,-p} - \boldsymbol{\beta}_k^{*^T}\boldsymbol{X}_{ip}\right)^2\right\} \times w$$

$$c_0 = \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sigma_\epsilon^{-2}\phi_k^2\tau_{ik,-p}^2\right\} \times (1-w)$$

$$\tau_{ik,-p} = \eta_{ik}^* - \mu_{ik}^* - \boldsymbol{\beta}_k^{*^T}\sum_{h=1,h\neq p}^{P}\gamma_h\boldsymbol{X}_{ih}$$

$$\pi(w|\cdot) \sim \text{Beta}\left(a_w + \sum_{p=1}^{P}\gamma_p, \ \ b_w + P - \sum_{p=1}^{P}\gamma_p\right)$$

where $\boldsymbol{U} = \{U(v)\}_{|\mathcal{R}\times 1|}$, $\boldsymbol{Z}_i = \{Z_i(v)\}_{|\mathcal{R}\times 1|}$, $\boldsymbol{\beta}_k^* = \{\beta_k^*(v)\}_{|\mathcal{R}|\times 1}$, $\boldsymbol{X}_{ip} = \{X_{ip}(v)\}_{|\mathcal{R}|\times 1}$

and $\boldsymbol{\gamma} = (\gamma_p)_{P\times 1}$.

B.3 Estimations and Predictions from Gibbs Sampler

Let $\hat{\boldsymbol{\Theta}}^{(t)}$ be the set of posterior samples of all parameters in every iteration t given the dataset $(\boldsymbol{Z}, \boldsymbol{X})$, that $\hat{\boldsymbol{\Theta}}^{(t)} = (\hat{\boldsymbol{\alpha}}^{(t)}, \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\gamma}}^{(t)}, \hat{\boldsymbol{\eta}}^{(t)}, \hat{\boldsymbol{\Lambda}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{U}}^{(t)}, \hat{\sigma}_{\epsilon}^{(t)}, \hat{\sigma}_{\zeta}^{(t)}, \hat{\sigma}_{e}^{(t)})$.

In every iteration t, the posterior estimations of outcome image at voxel v for subject i in training set that $i \in \mathcal{I}_i^{\text{tr}}$ is defined as

$$[\hat{Z}_i(v)]^{(t)} = \hat{U}^{(t)}(v) + \sum_{m=1}^{M} \hat{\theta}_{im}^{(t)} b_m(v) + \hat{e}_i^{(t)}(v), \;\; i \in \mathcal{I}_i^{\text{tr}}$$

where

$$\hat{\theta}_{im}^{(t)} = \sum_{k=1}^{K} \hat{\lambda}_{mk}^{(t)} \hat{\eta}_{ik}^{(t)} + \hat{\zeta}_{im}^{(t)}$$

$$\hat{\eta}_{ik}^{(t)} = \sum_{v\in\mathcal{R}} \Big[\sum_{p=1}^{P} \hat{\gamma}_p^{(t)} X_{ip}(v) \Big] \hat{\beta}_k^{(t)}(v) + \hat{\epsilon}_{ik}^{(t)}$$

$$\hat{e}_i^{(t)}(v) \sim N(0, \hat{\sigma}_e^{(t)}), \;\; \hat{\zeta}_{im}^{(t)}(v) \sim N(0, \hat{\sigma}_{\zeta}^{(t)}), \;\; \hat{\epsilon}_i^{(t)}(v) \sim N(0, \hat{\sigma}_{\epsilon}^{(t)})$$

For a total of T iterations, the posterior mean estimation of outcome image at voxel v for subject $i \in \mathcal{I}_i^{\text{tr}}$ from the training set is

$$\hat{Z}_i(v) = \sum_{t=T/2+1}^{T} [\hat{Z}_i(v)]^{(t)}, \quad i \in \mathcal{I}_i^{\text{tr}}$$

Given the new predictor images $\{\boldsymbol{X}_{j1}, \boldsymbol{X}_{j2}, \ldots, \boldsymbol{X}_{jP}\}$ ($\boldsymbol{X}_{jp} = \{X_{jp}(v)\}_{|\mathcal{R}|\times 1}$) of subject $j \in \mathcal{I}_j^{\text{ts}}$ in the test set, the prediction of the corresponding outcome image $\boldsymbol{Z}_j = \{Z_j(v)\}_{|\mathcal{R}|\times 1}$ at voxel v in each iteration t is defined as

$$[\hat{Z}_j(v)]^{(t)} = \hat{U}^{(t)}(v) + \sum_{m=1}^{M} \hat{\theta}_{jm}^{(t)} b_m(v) + \hat{e}_j(v)^{(t)}, \quad j \in \mathcal{I}_j^{\text{ts}}$$

where

$$\hat{\theta}_{jm}^{(t)} = \sum_{k=1}^{K} \hat{\lambda}_{mk}^{(t)} \hat{\eta}_{jk}^{(t)} + \hat{\zeta}_{jm}^{(t)}$$

$$\hat{\eta}_{jk}^{(t)} = \sum_{v \in \mathcal{R}} \Big[\sum_{p=1}^{P} \hat{\gamma}_{p}^{(t)} X_{jp}(v) \Big] \hat{\beta}_{k}^{(t)}(v) + \hat{\epsilon}_{jk}^{(t)}$$

$$\hat{e}_{j}^{(t)}(v) \sim N(0, \hat{\sigma}_{e}^{(t)}), \;\; \hat{\zeta}_{jm}^{(t)}(v) \sim N(0, \hat{\sigma}_{\zeta}^{(t)}), \;\; \hat{\epsilon}_{j}^{(t)}(v) \sim N(0, \hat{\sigma}_{\epsilon}^{(t)})$$

In particular, $\hat{U}^{(t)}(v)$, $\hat{\lambda}_{mk}^{(t)}$, $\gamma_{p}^{(t)}$, $\hat{\beta}_{k}^{(t)}(v)$, $\hat{\sigma}_{e}^{(t)}$, $\hat{\sigma}_{\zeta}^{(t)}$ and $\hat{\sigma}_{\epsilon}^{(t)}$ are posterior samples from training set in iteration t. Therefore, the final prediction of outcome image at voxel v for a subject j in test test ($j \in \mathcal{I}_{j}^{\text{ts}}$) is calculated as

$$\hat{Z}_{j}(v) = \sum_{t=T/2+1}^{T} [\hat{Z}_{j}(v)]^{(t)}$$

B.4 Graphical Representation of The Generating Models Used For Simulation Study

B.5 Figures and Tables of Real Data Analysis

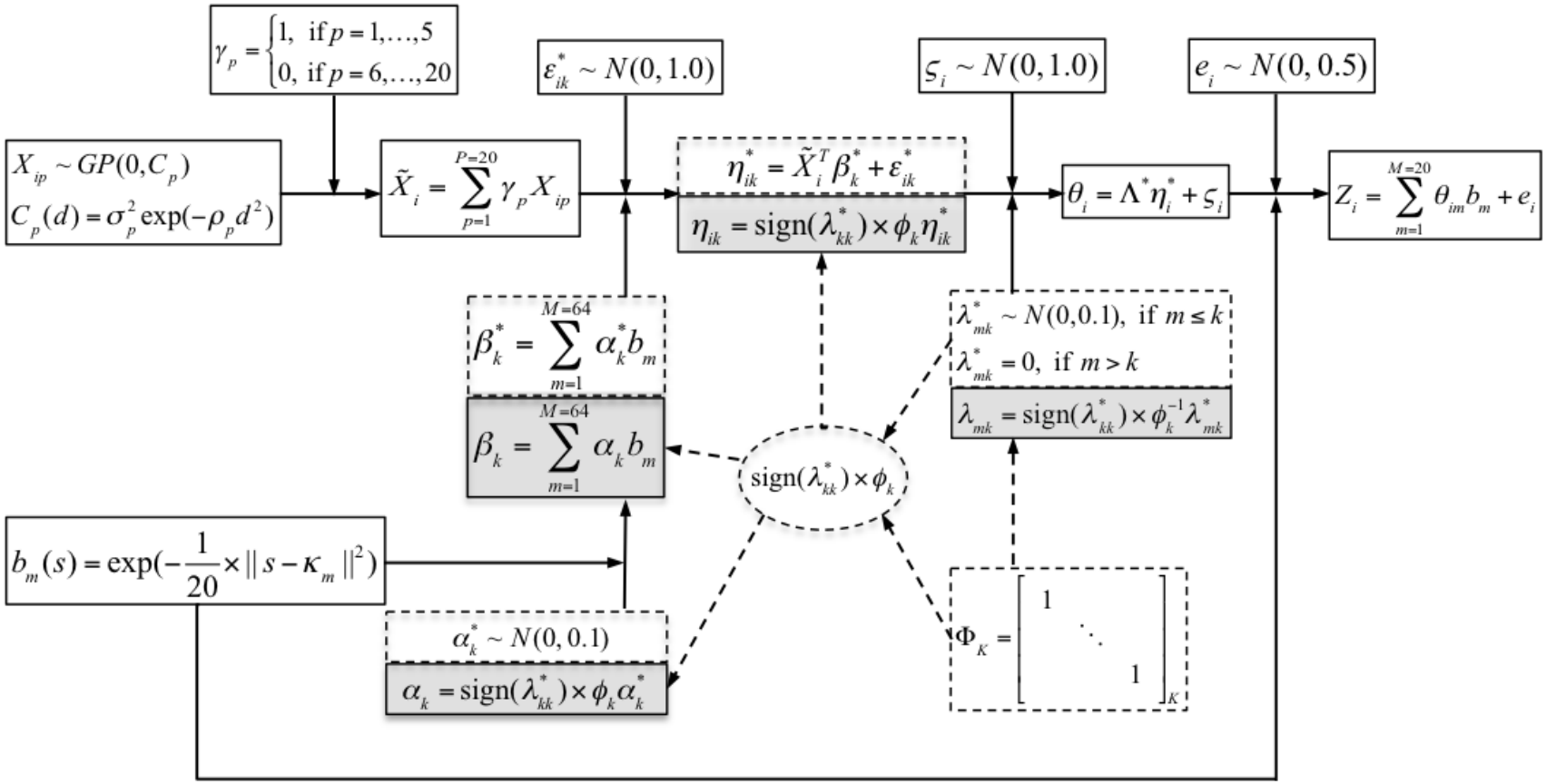


Figure B.1: Graphical representation of the generating models used in simulation study scenario 3. Dashed squares contain working parameters used in parameter expansion method. Shaded squares represent the transformation mechanism from working to original inferential parameters. The true number of latent factors K is set to be 5.

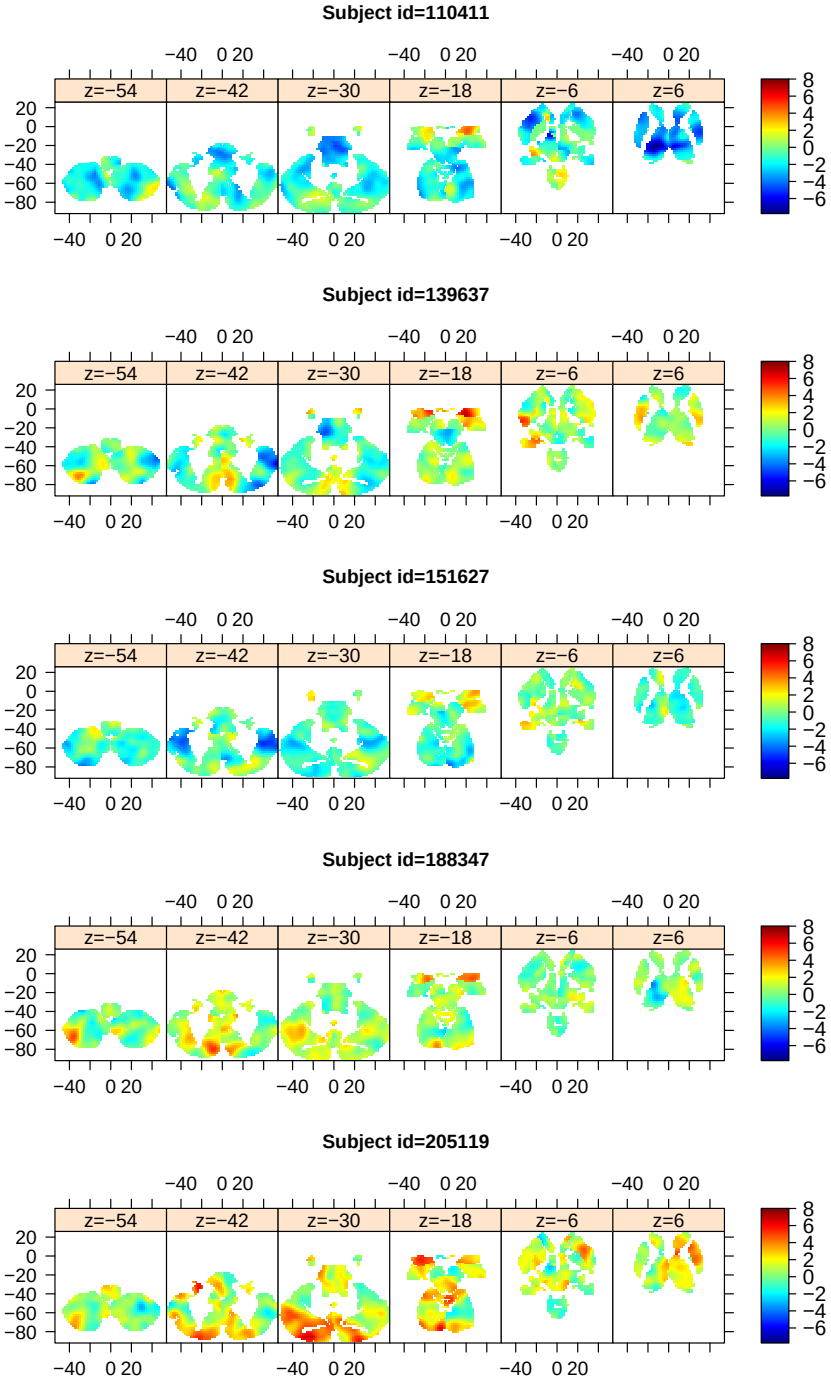


Figure B.2: Examples of outcome images (whole-brain faces-shapes contrast maps in EMOTION domain) from 5 subjects. Maps are shown at six different axial slices. All maps are plotted on the same color scale.

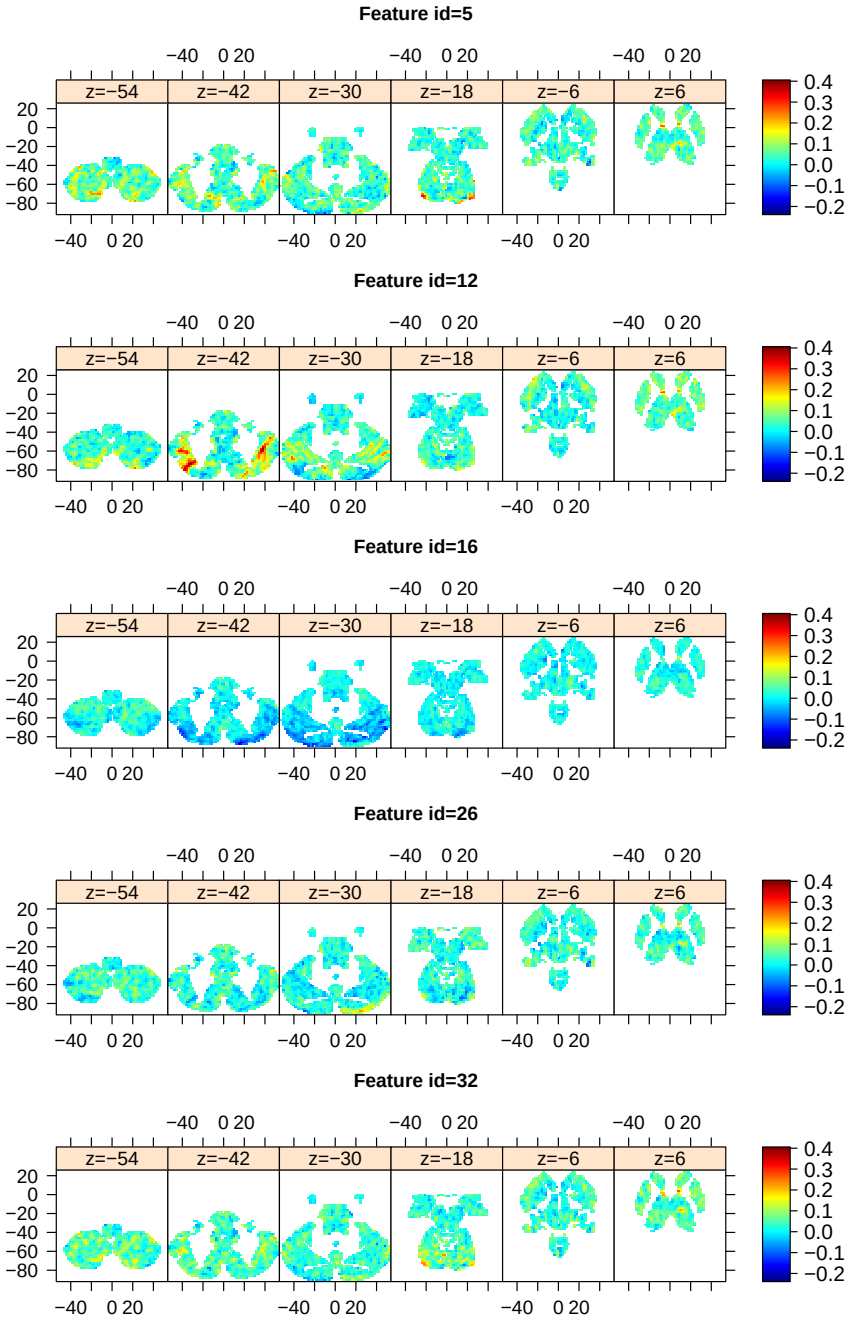


Figure B.3: Examples of six predictor maps (whole-brain sub-cortical seed maps) from a single subject (id = 110411), shown at six different axial slices. All maps are plotted on the same color scale.

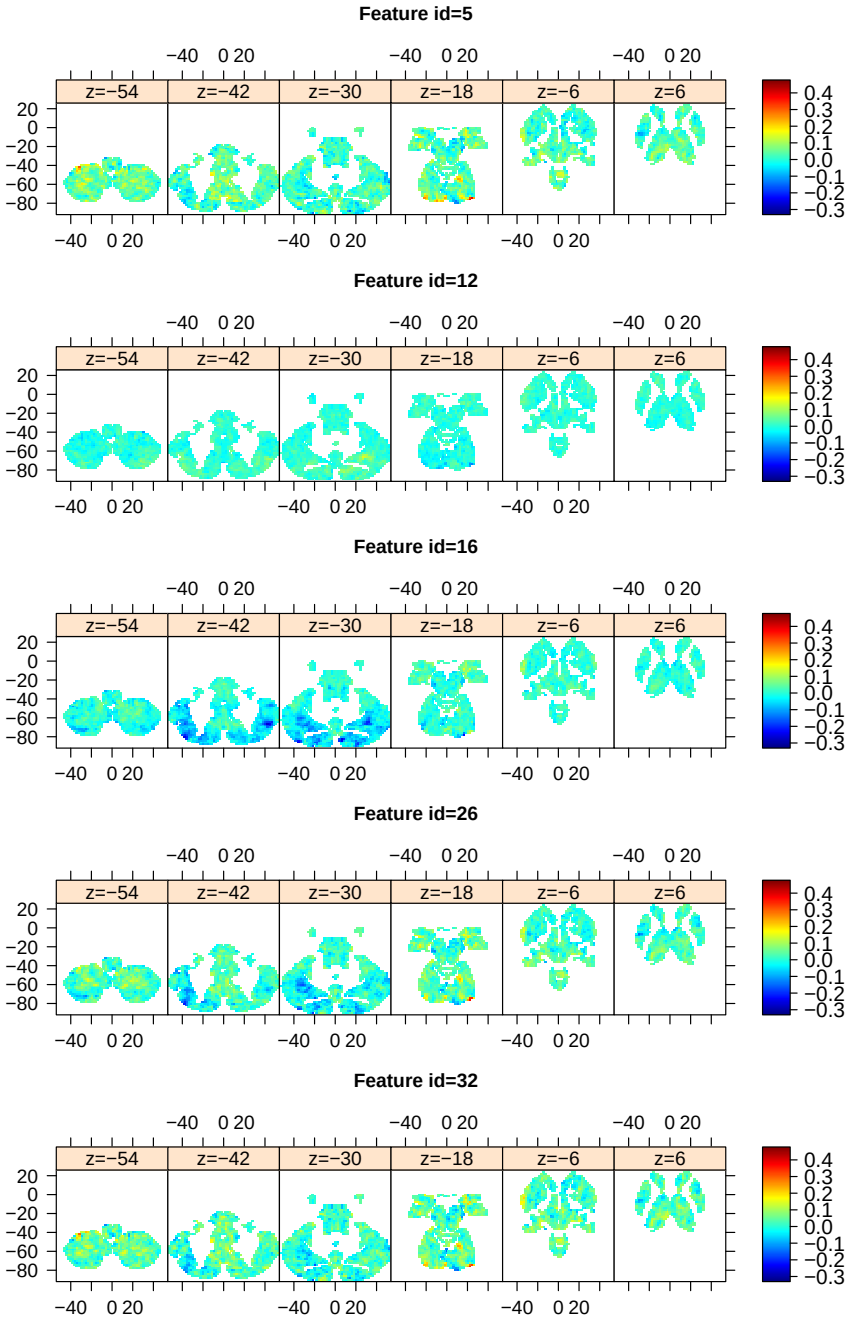


Figure B.4: Examples of six feature whole-brain maps (sub-cortical seed maps) from a single subject (id = 139637), shown at six different axial slices. All maps are plotted on the same color scale.

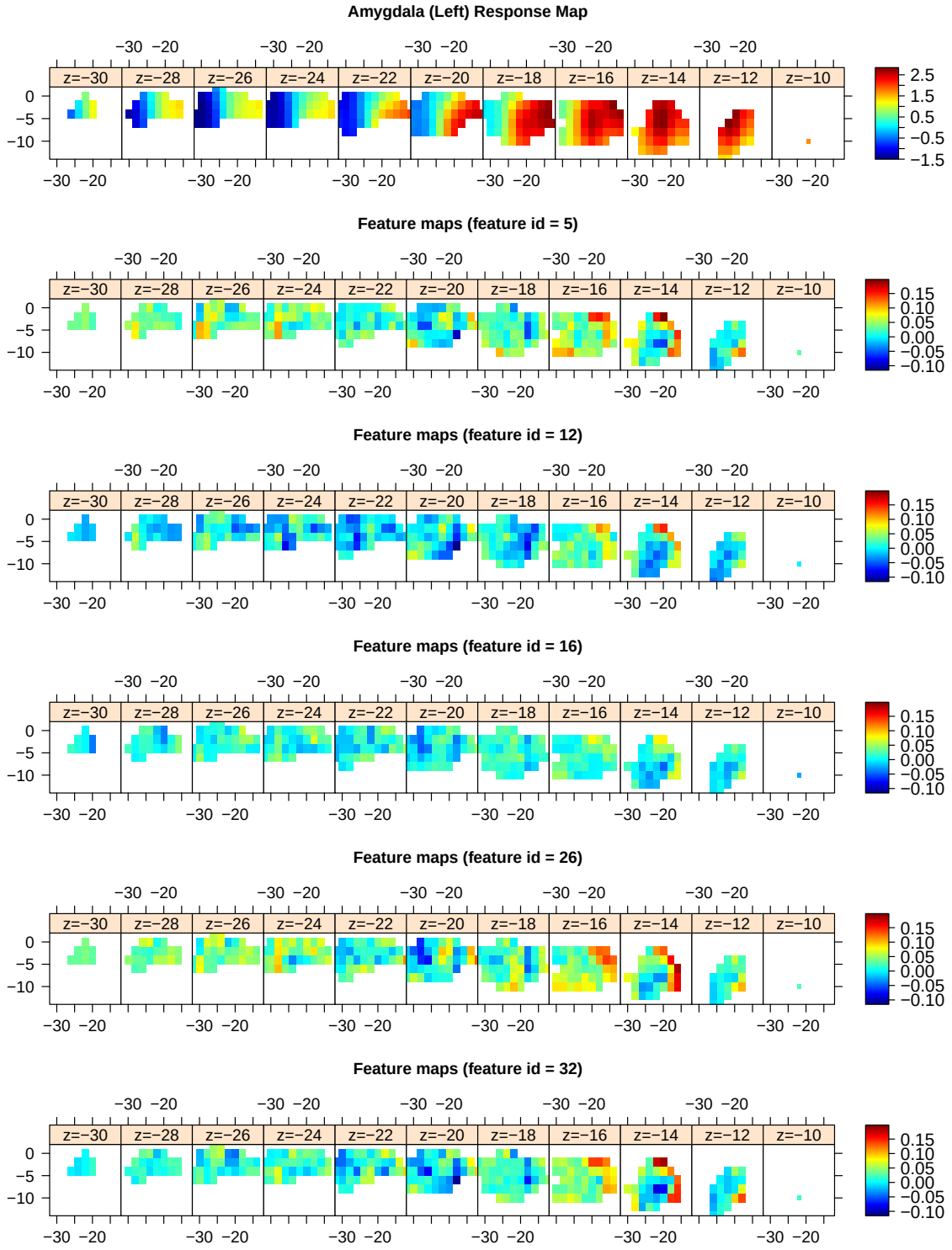


Figure B.5: Example images of outcome (faces-shapes contrast maps in EMOTION domain) and predictor (five sub-cortial seed maps) within the left amygdala region from a single subject (id = 110411), shown at axial slices.

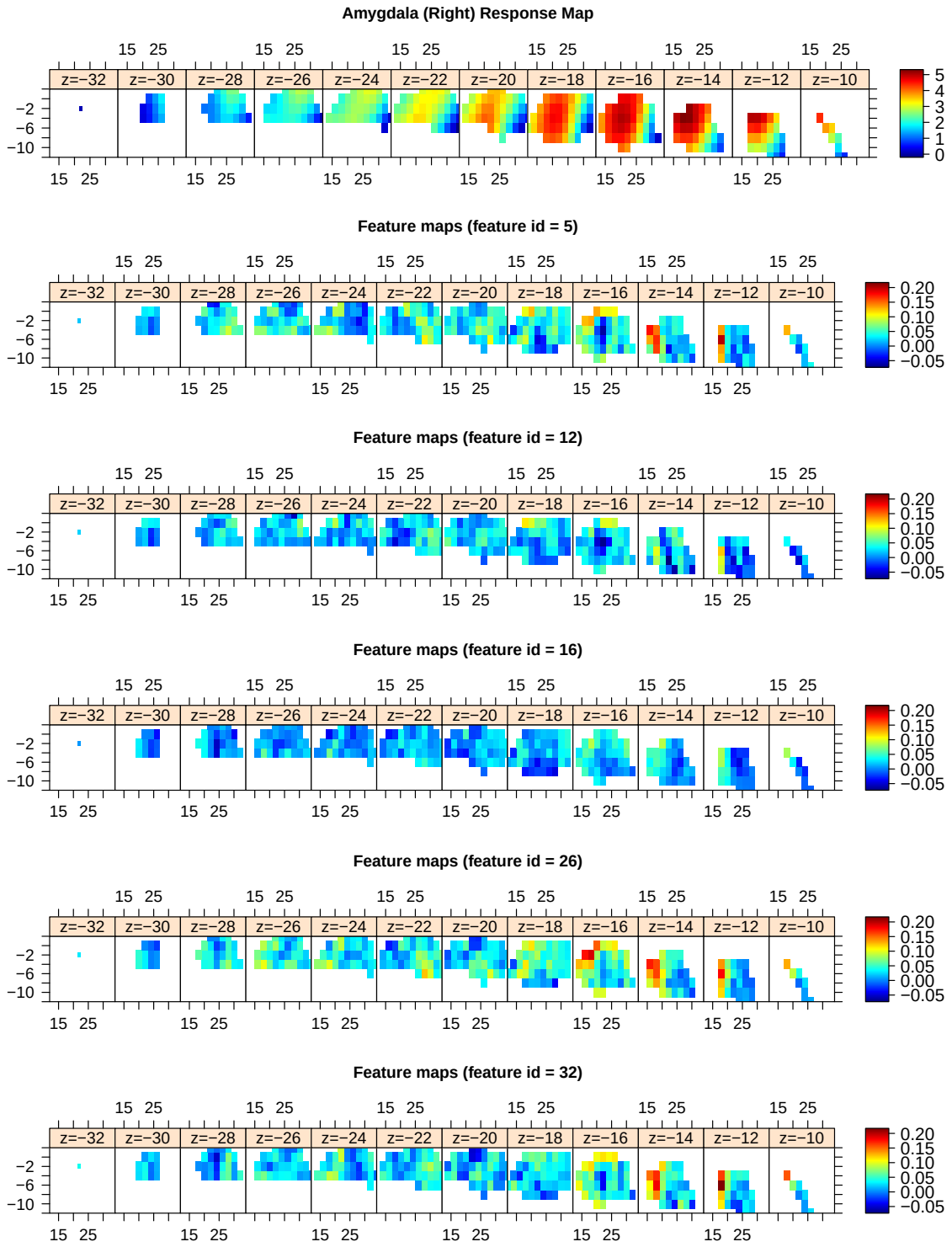


Figure B.6: Example images of outcome (faces-shapes contrast maps in EMOTION domain) and predictor (five sub-cortial seed maps) within the right amygdala region from a single subject (id = 110411), shown at axial slices.

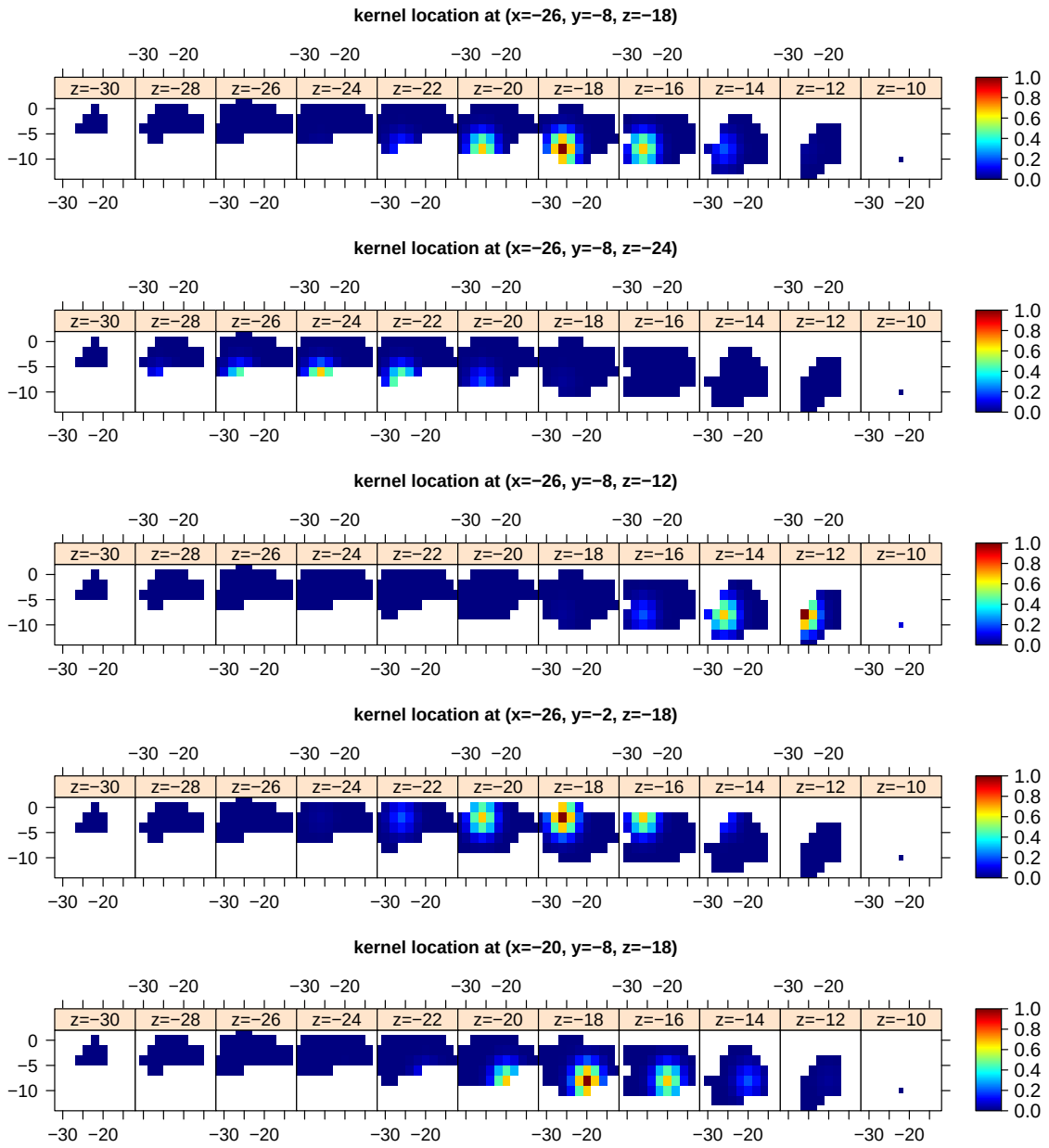


Figure B.7: Maps of basis functions with bandwidth value $b = 1/10$ and five different kernel locations for application analysis within the left amygdala region, shown at axial slices.

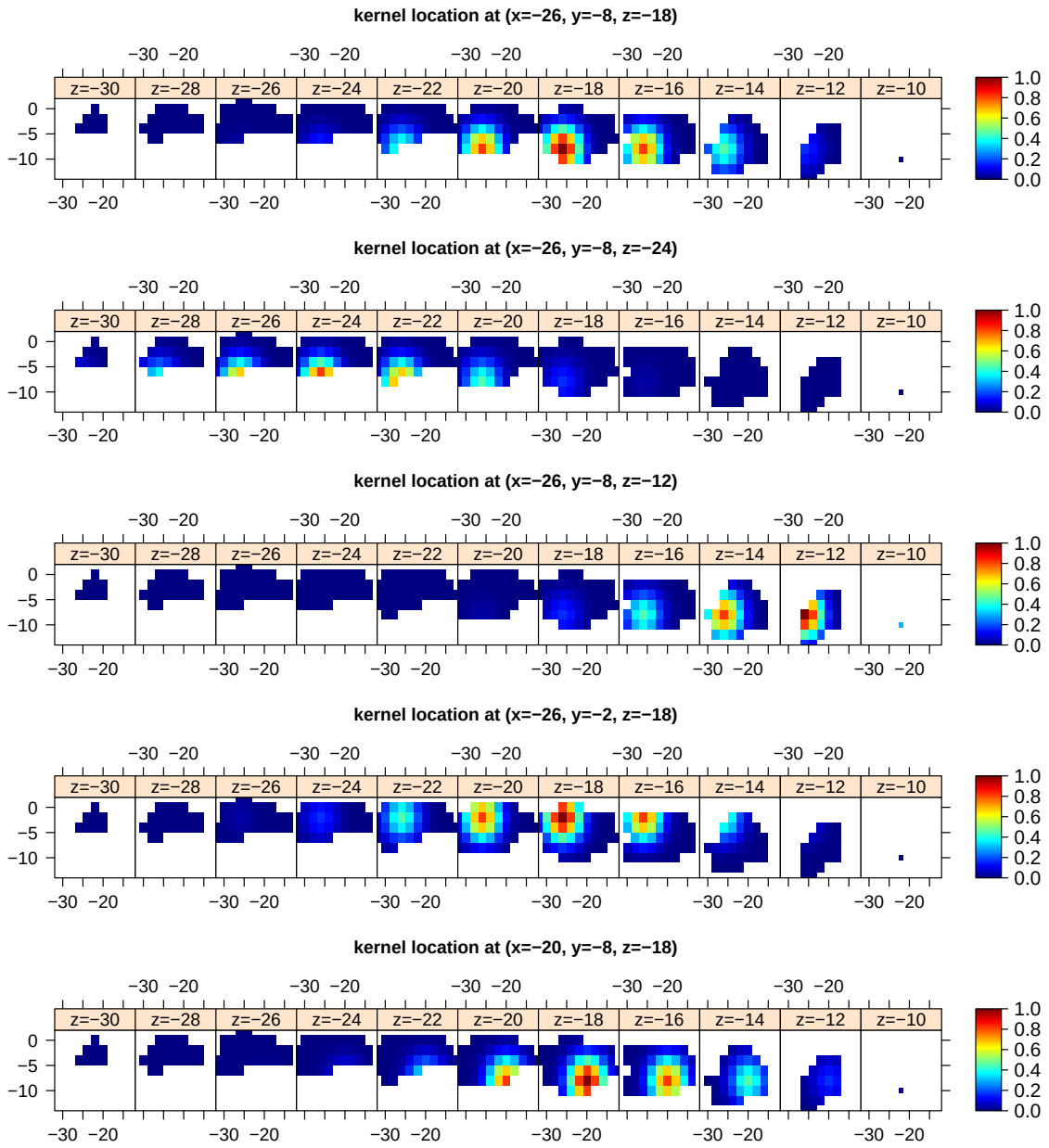


Figure B.8: Maps of basis functions with bandwidth value $b = 1/20$ and five different kernel locations for application analysis within the left amygdala region, shown at axial slices.

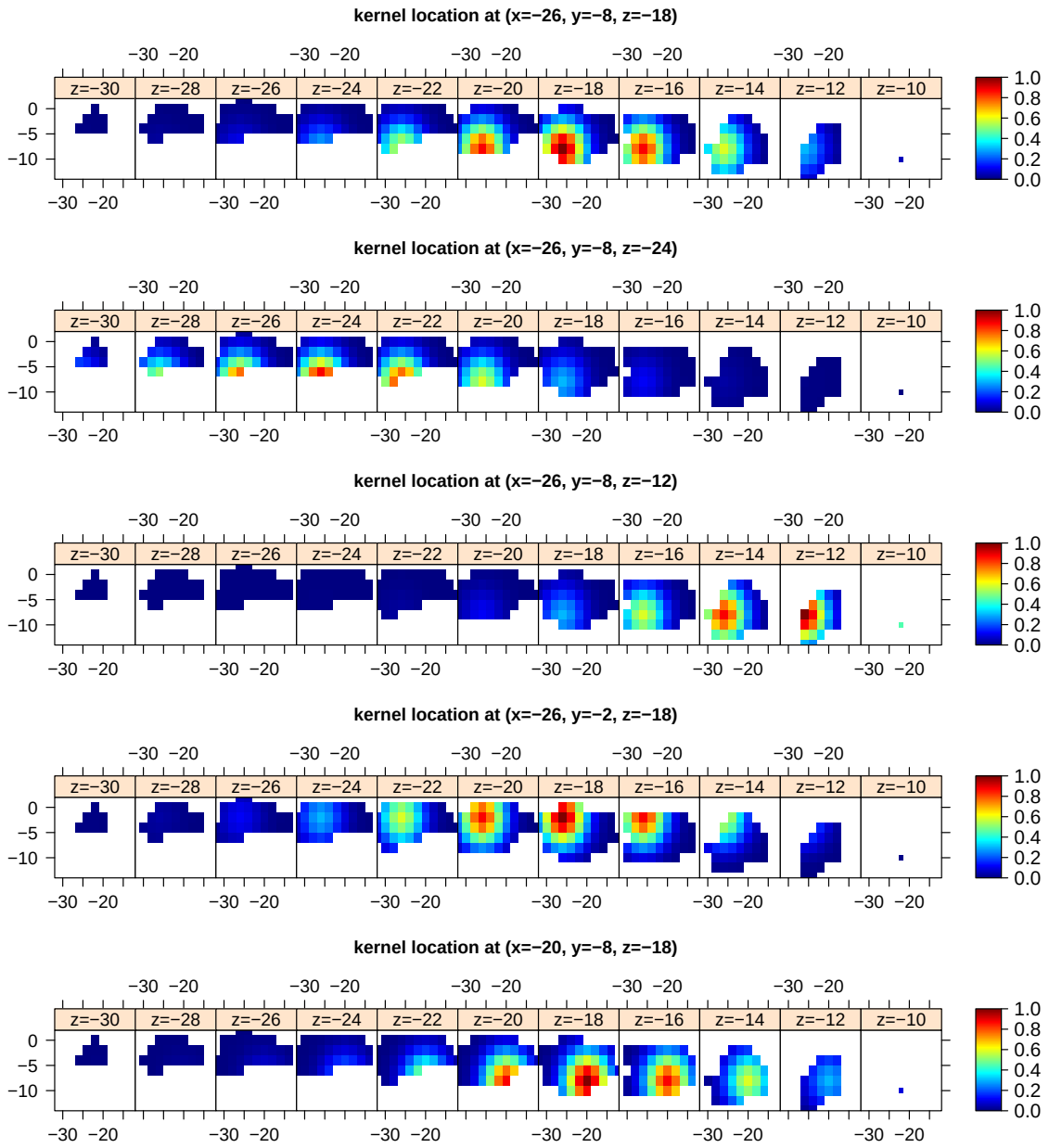


Figure B.9: Maps of basis functions with bandwidth value $b = 1/30$ and five different kernel locations for application analysis within the left amygdala region, shown at axial slices.

Table B.1: Selected feature images in application analysis based on posterior estimations of $\boldsymbol{\gamma}$ and varied thresholds for the left (1) and right (2) amygdala regions, respectively.

(1) Left amygdala region

Threshold	NO. of Features	Selected Feature (Index)
0.0	10	1, 2, 6, 13, 15, 21, 22, 28, 29, 31
0.1	8	1, 2, 13, 15, 21, 22, 28, 29
0.3	7	1, 2, 13, 15, 21, 22, 28
0.4	5	1, 2, 13, 15, 28
0.5	4	**2**, **13**, **15**, **28**
0.8	2	**13**, **15**
0.9	0	None

(2) Right amygdala region

Threshold	NO. of Features	Selected Feature (Index)
0.0	10	1, 2, 4, 13, 15, 21, 22, 28, 29, 31
0.1	9	1, 2, 4, 13, 15, 21, 22, 28, 31
0.2	6	1, 13, 15, 21, 22, 28
0.3	5	1, 13, 21, 22, 28
0.4	4	1, 21, 22, 28
0.5	2	**21**, **28**
0.6	1	**28**

Table B.2: Results of independent parcels analysis in Application study using SBLF and linear regression models.

Parcel Name	NO. of Voxels	Optimal K	Mean Squared Error		R-Squared	
			SBLF	Linear	SBLF	Linear
ACCUMBENS LEFT	135	8	0.031	0.151	0.941	0.759
ACCUMBENS RIGHT	140	6	0.037	0.194	0.926	0.692
AMYGDALA LEFT	315	9	0.063	0.562	0.932	0.600
AMYGDALA RIGHT	332	10	0.069	0.651	0.935	0.621
BRAIN STEM	3472	2	0.107	1.349	0.921	0.068
CAUDATE LEFT	728	5	0.055	0.660	0.949	0.405
CAUDATE RIGHT	766	8	0.061	0.691	0.946	0.412
DIENCEPHALON VENTRAL LEFT	706	5	0.085	1.169	0.935	0.257
DIENCEPHALON VENTRAL RIGHT	712	10	0.088	1.117	0.930	0.246
HIPPOCAMPUS LEFT	764	5	0.060	0.818	0.949	0.380
HIPPOCAMPUS RIGHT	795	5	0.063	0.828	0.948	0.371
PALLIDUM LEFT	297	5	0.061	0.370	0.917	0.563
PALLIDUM RIGHT	260	6	0.062	0.389	0.913	0.495
PUTAMEN LEFT	1060	6	0.071	0.776	0.943	0.376
PUTAMEN RIGHT	1010	7	0.066	0.730	0.940	0.363
THALAMUS LEFT	1288	4	0.076	0.829	0.931	0.324
THALAMUS RIGHT	1248	6	0.078	0.764	0.927	0.346
Averaged Values Across Parcels			**0.067**	**0.709**	**0.934**	**0.428**

SBLF: our proposed spatial Bayesian latent factor model.

BIBLIOGRAPHY

BIBLIOGRAPHY

Allen, G. I. (2013), Multi-way functional principal components analysis, in *Computational advances in multi-sensor adaptive processing (CAMSAP), 2013 IEEE 5th International Workshop on*, pp. 220–223, IEEE.

Angers, J.-F., P. T. Kim, et al. (2005), Multivariate bayesian function estimation, *The Annals of Statistics*, *33*(6), 2967–2999.

Ardran, G. (1979), The application and limitation of the use of x-rays in medical diagnosis, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *292*(1390), 147–156.

Atlas, S. W. (2009), *Magnetic resonance imaging of the brain and spine*, vol. 1, Lippincott Williams & Wilkins.

Bakshi, R., A. Minagar, Z. Jaisani, and J. S. Wolinsky (2005), Imaging of multiple sclerosis: role in neurotherapeutics, *NeuroRx*, *2*(2), 277–303.

Balafar, M. A., A. R. Ramli, M. I. Saripan, and S. Mashohor (2010), Review of brain mri image segmentation methods, *Artificial Intelligence Review*, *33*(3), 261–274.

Barch, D. M., et al. (2013), Function in the human connectome: task-fmri and individual differences in behavior, *Neuroimage*, *80*, 169–189.

Bates, E., S. M. Wilson, A. P. Saygin, F. Dick, M. I. Sereno, R. T. Knight, and N. F. Dronkers (2003), Voxel-based lesion–symptom mapping, *Nature neuroscience*, *6*(5), 448–450.

Besag, J. (1974a), Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236.

Besag, J. (1974b), Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236.

Bowman, F. D. (2005), Spatio-temporal modeling of localized brain activity, *Biostatistics*, *6*(4), 558–575.

Bowman, F. D. (2007), Spatiotemporal models for region of interest analyses of functional neuroimaging data, *Journal of the American Statistical Association*, *102*(478), 442–453.

Bowman, F. D. (2014), Brain imaging analysis, *Annual Review of Statistics and Its Application*, *1*, 61–85.

Bowman, F. D., B. Caffo, S. S. Bassett, and C. Kilts (2008), A bayesian hierarchical framework for spatial modeling of fmri data, *NeuroImage*, *39*(1), 146–156.

Bracewell, R. N., and R. N. Bracewell (1986), *The Fourier transform and its applications*, vol. 31999, McGraw-Hill New York.

Brezger, A., L. Fahrmeir, and A. Hennerfeind (2007), Adaptive gaussian markov random fields with applications in human brain mapping, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *56*(3), 327–345.

Calabrese, M., V. Poretto, A. Favaretto, S. Alessio, V. Bernardi, C. Romualdi, F. Rinaldi, P. Perini, and P. Gallo (2012), Cortical lesion load associates with progression of disability in multiple sclerosis, *Brain*, *135*(10), 2952–2961.

Charil, A., A. P. Zijdenbos, J. Taylor, C. Boelman, K. J. Worsley, A. C. Evans, and A. Dagher (2003), Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: application to 452 patient data sets, *Neuroimage*, *19*(3), 532–544.

Charil, A., A. Dagher, J. P. Lerch, A. P. Zijdenbos, K. J. Worsley, and A. C. Evans (2007), Focal cortical atrophy in multiple sclerosis: relation to lesion load and disability, *Neuroimage*, *34*(2), 509–517.

Chen, Y., X. Wang, L. Kong, and H. Zhu (2016), Local region sparse learning for image-on-scalar regression, *arXiv preprint arXiv:1605.08501*.

Chung, M. K., R. Hartley, K. M. Dalton, and R. J. Davidson (2008), Encoding cortical surface by spherical harmonics, *Statistica Sinica*, pp. 1269–1291.

Coalson, T. S., D. C. Van Essen, and M. Glasser (2018), Lost in space: The impact of traditional neuroimaging methods on the spatial localization of cortical areas, *bioRxiv*, p. 255620.

Compston, A., and A. Coles (2002), Multiple sclerosis, *The Lancet*, *359*(9313), 1221–1231.

Courant, R., and D. Hilbert (2008), *Methods of Mathematical Physics: Partial Differential Equations*, John Wiley & Sons.

Daghighian, F., R. Sumida, and M. E. Phelps (1990), Pet imaging: an overview and instrumentation, *Journal of nuclear medicine technology*, *18*(1), 5–13.

Dale, A. M. (1999), Optimal experimental design for event-related fmri, *Human brain mapping*, *8*(2-3), 109–114.

Dale, A. M., B. Fischl, and M. I. Sereno (1999), Cortical surface-based analysis: I. segmentation and surface reconstruction, *Neuroimage*, *9*(2), 179–194.

Das, S., B. D. Wandelt, and T. Souradeep (2015), Bayesian inference on the sphere beyond statistical isotropy, *Journal of Cosmology and Astroparticle Physics*, *2015*(10), 050.

Donya, M., M. Radford, A. ElGuindy, D. Firmin, and M. H. Yacoub (2015), Radiation in medicine: Origins, risks and aspirations, *Global Cardiology Science and Practice*, p. 57.

Duff, E. P., et al. (2015), Learning to identify cns drug action and efficacy using multistudy fmri data, *Science Translational Medicine*, *7*(274), 274ra16–274ra16.

Elahi, U., Z. Khalid, R. A. Kennedy, and J. D. McEwen (2017), Iterative residual fitting for spherical harmonic transform of band-limited signals on the sphere: Generalization and analysis, in *2017 International Conference on Sampling Theory and Applications (SampTA)*, pp. 470–474, IEEE.

Epstein, C. L. (2007), *Introduction to the mathematics of medical imaging*, SIAM.

Filli, L., et al. (2012), Spatiotemporal distribution of white matter lesions in relapsing–remitting and secondary progressive multiple sclerosis, *Multiple sclerosis journal*, *18*(11), 1577–1584.

Gao, J. S., A. G. Huth, M. D. Lescroart, and J. L. Gallant (2015), Pycortex: an interactive surface visualizer for fmri, *Frontiers in neuroinformatics*, *9*, 23.

Ge, T., N. Müller-Lenke, K. Bendfeldt, T. E. Nichols, and T. D. Johnson (2014a), Analysis of multiple sclerosis lesions via spatially varying coefficients, *The annals of applied statistics*, *8*(2), 1095.

Ge, T., N. Müller-Lenke, K. Bendfeldt, T. E. Nichols, and T. D. Johnson (2014b), Analysis of multiple sclerosis lesions via spatially varying coefficients, *The annals of applied statistics*, *8*(2), 1095.

Gelfand, A. E., H.-J. Kim, C. Sirmans, and S. Banerjee (2003), Spatial modeling with spatially varying coefficient processes, *Journal of the American Statistical Association*, *98*(462), 387–396.

Gelman, A., D. B. Rubin, et al. (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, *7*(4), 457–472.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014), *Bayesian data analysis*, vol. 2, Chapman & Hall/CRC Boca Raton, FL, USA.

Gerig, G., M. Styner, D. Jones, D. Weinberger, and J. Lieberman (2001), Shape analysis of brain ventricles using spharm, in *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*, pp. 171–178, IEEE.

Ghosh, J., and D. B. Dunson (2009), Default prior distributions and efficient posterior computation in Bayesian factor analysis, *Journal of Computational and Graphical Statistics*, *18*(2), 306–320.

Gibbs, M., and D. MacKay (1996), Efficient implementation of gaussian processes.

Glasser, M. F., et al. (2013), The minimal preprocessing pipelines for the human connectome project, *Neuroimage*, *80*, 105–124.

Glasser, M. F., et al. (2016), A multi-modal parcellation of human cerebral cortex, *Nature*, *536*(7615), 171.

Glasser, O. (1993), *Wilhelm Conrad Röntgen and the early history of the Roentgen rays*, 1, Norman Publishing.

Glover, G. H. (2011), Overview of functional magnetic resonance imaging, *Neurosurgery Clinics*, *22*(2), 133–139.

Goldsmith, J., and T. Kitago (2016), Assessing systematic effects of stroke on motor control by using hierarchical function-on-scalar regression, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *65*(2), 215–236.

Goldsmith, J., M. P. Wand, and C. Crainiceanu (2011), Functional regression via variational bayes, *Electronic Journal of Statistics*, *5*, 572.

Goldsmith, J., L. Huang, and C. M. Crainiceanu (2014), Smooth scalar-on-image regression via spatial bayesian variable selection, *Journal of Computational and Graphical Statistics*, *23*(1), 46–64.

Golub, G. H., and C. F. Van Loan (2012), *Matrix computations*, vol. 3, JHU Press.

Gössl, C., D. P. Auer, and L. Fahrmeir (2001), Bayesian spatiotemporal inference in functional magnetic resonance imaging, *Biometrics*, *57*(2), 554–562.

Groves, A. R., M. A. Chappell, and M. W. Woolrich (2009), Combined spatial and non-spatial prior for inference on mri time-series, *NeuroImage*, *45*(3), 795–809.

Gu, X., Y. Wang, T. F. Chan, P. M. Thompson, and S.-T. Yau (2004), Genus zero surface conformal mapping and its application to brain surface mapping, *IEEE transactions on medical imaging*, *23*(8), 949–958.

Hand, D. J. (2006), Classifier technology and the illusion of progress, *Statistical science*, pp. 1–14.

Hashemi, R. H., W. G. Bradley, and C. J. Lisanti (), Mri: The basics. 2004.

Hazra, A., B. J. Reich, D. S. Reich, R. T. Shinohara, and A.-M. Staicu (2017), A spatio-temporal model for longitudinal image-on-image regression, *Statistics in Biosciences*, pp. 1–25.

Heggie, J. C. (2001), Magnetic resonance imaging: Principles, methods and techniques by perry sprawls, *Australasian Physical & Engineering Science in Medicine*, *24*(1), 57–57.

Hoffman, M. D., and A. Gelman (2014), The no-u-turn sampler: adaptively setting path lengths in Hamiltonian monte carlo., *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holland, C. M., A. Charil, I. Csapo, Z. Liptak, M. Ichise, S. J. Khoury, R. Bakshi, H. L. Weiner, and C. R. Guttmann (2012), The relationship between normal cerebral perfusion patterns and white matter lesion distribution in 1,249 patients with multiple sclerosis, *Journal of Neuroimaging*, *22*(2), 129–136.

Homeier, H. H., and E. O. Steinborn (1996), Some properties of the coupling coefficients of real spherical harmonics and their relation to gaunt coefficients, *Journal of Molecular Structure: THEOCHEM*, *368*, 31–37.

Huang, H., P. S. Yu, and C. Wang (2018), An introduction to image synthesis with generative adversarial nets, *arXiv preprint arXiv:1803.04469*.

Huang, L., J. Goldsmith, P. T. Reiss, D. S. Reich, and C. M. Crainiceanu (2013), Bayesian scalar-on-image regression with application to association between intracranial dti and cognitive outcomes, *NeuroImage*, *83*, 210–223.

Hyun, J. W., Y. Li, J. H. Gilmore, Z. Lu, M. Styner, and H. Zhu (2014), Sgpp: spatial gaussian predictive process models for neuroimaging data, *NeuroImage*, *89*, 70–80.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros (2017), Image-to-image translation with conditional adversarial networks, *arXiv preprint*.

Kang, J., B. J. Reich, and A.-M. Staicu (2016), Scalar-on-image regression via the soft-thresholded gaussian process, *arXiv preprint arXiv:1604.03192*.

Kang, J., B. J. Reich, and A.-M. Staicu (2018), Scalar-on-image regression via the soft-thresholded gaussian process, *Biometrika*, *105*(1), 165–184.

Katanoda, K., Y. Matsuda, and M. Sugishita (2002), A spatio-temporal regression model for the analysis of functional mri data, *NeuroImage*, *17*(3), 1415–1428.

Kelemen, A., G. Székely, and G. Gerig (1999), Elastic model-based segmentation of 3-d neuroradiological data sets, *IEEE Transactions on medical imaging*, *18*(10), 828–839.

Kennedy, R. A., and P. Sadeghi (2013), *Hilbert space methods in signal processing*, Cambridge University Press.

Kwong, K. K., et al. (1992), Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation., *Proceedings of the National Academy of Sciences*, *89*(12), 5675–5679.

Lang, A., C. Schwab, et al. (2015), Isotropic gaussian random fields on the sphere: regularity, fast simulation and stochastic partial differential equations, *The Annals of Applied Probability*, *25*(6), 3047–3094.

Lazar, N. (2008), *The statistical analysis of functional MRI data*, Springer Science & Business Media.

Li, F., T. Zhang, Q. Wang, M. Z. Gonzalez, E. L. Maresh, J. A. Coan, et al. (2015), Spatial bayesian variable selection and grouping for high-dimensional scalar-on-image regression, *The Annals of Applied Statistics*, *9*(2), 687–713.

Li, Y., H. Zhu, D. Shen, W. Lin, J. H. Gilmore, and J. G. Ibrahim (2011), Multiscale adaptive regression models for neuroimaging data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 559–578.

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

Lladó, X., A. Oliver, M. Cabezas, J. Freixenet, J. C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, and À. Rovira (2012), Segmentation of multiple sclerosis lesions in brain mri: a review of automated approaches, *Information Sciences*, *186*(1), 164–185.

Lövblad, K.-O., N. Anzalone, A. Dörfler, M. Essig, B. Hurwitz, L. Kappos, S.-K. Lee, and M. Filippi (2010), Mr imaging in multiple sclerosis: review and recommendations for current practice, *American journal of neuroradiology*, *31*(6), 983–989.

Lublin, F. D., S. C. Reingold, et al. (1996), Defining the clinical course of multiple sclerosis results of an international survey, *Neurology*, *46*(4), 907–911.

Lublin, F. D., et al. (2014), Defining the clinical course of multiple sclerosis the 2013 revisions, *Neurology*, *83*(3), 278–286.

MacKay Altman, R., A. J. Petkau, D. Vrecko, and A. Smith (2012), A longitudinal model for magnetic resonance imaging lesion count data in multiple sclerosis patients, *Statistics in medicine*, *31*(5), 449–469.

Marshall, R. S., J. J. Ferrera, A. Barnes, X. Zhang, K. A. O'Brien, M. Chmayssani, J. Hirsch, and R. M. Lazar (2008), Brain activity associated with stimulation therapy of the visual borderzone in hemianopic stroke patients, *Neurorehabilitation and Neural Repair*, *22*(2), 136–144.

Martin, N., and H. Maes (1979), *Multivariate analysis*, Academic press London.

Marx, B. D., and P. H. Eilers (2005), Multidimensional penalized signal regression, *Technometrics*, *47*(1), 13–22.

Mejia, A., Y. R. Yue, D. Bolin, F. Lindren, and M. A. Lindquist (2017), A bayesian general linear modeling approach to cortical surface fmri data analysis, *arXiv preprint arXiv:1706.00959*.

Montagna, S., S. T. Tokdar, B. Neelon, and D. B. Dunson (2012), Bayesian latent factor regression for functional and longitudinal data, *Biometrics*, *68*(4), 1064–1073.

Montagna, S., T. Wager, L. F. Barrett, T. D. Johnson, and T. E. Nichols (2018), Spatial bayesian latent factor regression modeling of coordinate-based meta-analysis data, *Biometrics*, *74*(1), 342–353.

Moodie, J., et al. (2012), Magnetic resonance disease severity scale (mrdss) for patients with multiple sclerosis: a longitudinal study, *Journal of the neurological sciences*, *315*(1), 49–54.

Morgan, C., I. Aban, C. Katholi, and G. Cutter (2010), Modeling lesion counts in multiple sclerosis when patients have been selected for baseline activity, *Multiple Sclerosis*.

Muir, J., and H. Tkalčić (2015), A method of spherical harmonic analysis in the geosciences via hierarchical bayesian inference, *Geophysical Journal International*, *203*(2), 1164–1171.

Müller, C. (2006), *Spherical harmonics*, vol. 17, Springer.

Müller, H.-G., and U. Stadtmüller (2005), Generalized functional linear models, *Annals of Statistics*, pp. 774–805.

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

Nasrallah, I., and J. Dubroff (2013), An overview of pet neuroimaging, in *Seminars in nuclear medicine*, vol. 43, pp. 449–461, Elsevier.

Natterer, F. (2001), *The mathematics of computerized tomography*, SIAM.

Neal, R. M., et al. (2011), *MCMC using Hamiltonian dynamics*, chap. 5, pp. 113–162, Chapman and Hall/CRC.

Ogawa, S., T.-M. Lee, A. R. Kay, and D. W. Tank (1990), Brain magnetic resonance imaging with contrast dependent on blood oxygenation, *proceedings of the National Academy of Sciences*, *87*(24), 9868–9872.

Ogawa, S., R. Menon, D. Tank, S. Kim, H. Merkle, J. Ellermann, and K. Ugurbil (1993), Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. a comparison of signal characteristics with a biophysical model, *Biophysical journal*, *64*(3), 803–812.

Ombao, H., M. Lindquist, W. Thompson, and J. Aston (2016), *Handbook of Neuroimaging Data Analysis*, Chapman and Hall/CRC.

Pham, D. L., C. Xu, and J. L. Prince (2000), Current methods in medical image segmentation, *Annual review of biomedical engineering*, *2*(1), 315–337.

Phan, K. L., T. Wager, S. F. Taylor, and I. Liberzon (2002), Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri, *Neuroimage*, *16*(2), 331–348.

Pinaud, O., O. Chadebec, L.-L. Rouve, J.-L. Coulomb, J.-M. Guichon, and A. Vassilev (2015), A bayesian approach for spherical harmonic expansion identification: Application to magnetostatic field created by a power circuitry, *IEEE Transactions on Electromagnetic Compatibility*, *57*(6), 1501–1509.

Polzehl, J., H. U. Voss, and K. Tabelow (2010), Structural adaptive segmentation for statistical parametric mapping, *NeuroImage*, *52*(2), 515–523.

Quiñonero-Candela, J., and C. E. Rasmussen (2005), A unifying view of sparse approximate gaussian process regression, *Journal of Machine Learning Research*, *6*(Dec), 1939–1959.

Rasmussen, C. E., and C. K. Williams (2006), *Gaussian processes for machine learning*, vol. 1, MIT press Cambridge.

Reiss, P. T., and R. T. Ogden (2010a), Functional generalized linear models with images as predictors, *Biometrics*, *66*(1), 61–69.

Reiss, P. T., and R. T. Ogden (2010b), Functional generalized linear models with images as predictors, *Biometrics*, *66*(1), 61–69.

Reiss, P. T., L. Huang, and M. Mennes (2010), Fast function-on-scalar regression with penalized basis expansions, *The International Journal of Biostatistics*, *6*(1).

Reiss, P. T., L. Huo, Y. Zhao, C. Kelly, and R. T. Ogden (2015), Wavelet-domain regression and predictive inference in psychiatric neuroimaging, *The annals of applied statistics*, *9*(2), 1076.

Röntgen, W. C. (1896), On a new kind of rays, *Science*, *3*(59), 227–231.

Rossi, F., et al. (2012), Relevance of brain lesion location to cognition in relapsing multiple sclerosis, *PloS one*, *7*(11), e44,826.

Rue, H., and L. Held (2005), *Gaussian Markov random fields: theory and applications*, CRC press.

Schröder, P., and W. Sweldens (1995), Spherical wavelets: Efficiently representing functions on the sphere, in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 161–172, ACM.

Seeger, M., C. Williams, and N. Lawrence (2003), Fast forward selection to speed up sparse gaussian process regression, in *Artificial Intelligence and Statistics 9*, EPFL-CONF-161318.

Sepulcre, J., J. Goñi, J. C. Masdeu, B. Bejarano, N. V. de Mendizábal, J. B. Toledo, and P. Villoslada (2009), Contribution of white matter lesions to gray matter atrophy in multiple sclerosis: evidence from voxel-based analysis of t1 lesions in the visual pathway, *Archives of neurology*, *66*(2), 173–179.

Shukla, A., and U. Kumar (2006), Positron emission tomography: An overview, *Journal of medical physics/Association of Medical Physicists of India*, *31*(1), 13.

Smola, A. J., and P. L. Bartlett (2001), Sparse greedy gaussian process regression, in *Advances in neural information processing systems*, pp. 619–625.

Snelson, E., and Z. Ghahramani (2006), Sparse gaussian processes using pseudo-inputs, in *Advances in neural information processing systems*, pp. 1257–1264.

Sormani, M. P., and M. Filippi (2007), Statistical issues related to the use of mri data in multiple sclerosis, *Journal of Neuroimaging*, *17*(s1), 56S–59S.

Suk, H.-I., S.-W. Lee, D. Shen, A. D. N. Initiative, et al. (2017), Deep ensemble learning of sparse regression models for brain disease diagnosis, *Medical Image Analysis*, *37*, 101–113.

Sweeney, E., R. Shinohara, C. Shea, D. Reich, and C. Crainiceanu (2013a), Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal MRI, *American Journal of Neuroradiology*, *34*(1), 68–73.

Sweeney, E. M., J. T. Vogelstein, J. L. Cuzzocreo, P. A. Calabresi, D. S. Reich, C. M. Crainiceanu, and R. T. Shinohara (2014), A comparison of supervised machine learning algorithms and feature vectors for ms lesion segmentation using multimodal structural mri, *PloS one*, *9*(4), e95,753.

Sweeney, E. M., et al. (2013b), Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri, *NeuroImage: clinical*, *2*, 402–413.

Tavor, I., O. P. Jones, R. Mars, S. Smith, T. Behrens, and S. Jbabdi (2016), Task-free MRI predicts individual differences in brain activity during task performance, *Science*, *352*(6282), 216–220.

Vehtari, A., A. Gelman, and J. Gabry (2016), Practical bayesian model evaluation using leave-one-out cross-validation and waic, *Statistics and Computing*, pp. 1–20.

Wang, X., H. Zhu, and A. D. N. Initiative (2017), Generalized scalar-on-image regression models via total variation, *Journal of the American Statistical Association*, *112*(519), 1156–1168.

Wills, A., and L. Hector (1924), The magnetic susceptibility of oxygen, hydrogen and helium, *Physical Review*, *23*(2), 209.

Wolfers, T., J. K. Buitelaar, C. F. Beckmann, B. Franke, and A. F. Marquand (2015), From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics, *Neuroscience & Biobehavioral Reviews*, *57*, 328–349.

Wood, A. T., and G. Chan (1994), Simulation of stationary gaussian processes in [0, 1] d, *Journal of computational and graphical statistics*, *3*(4), 409–432.

Worsley, K. J., and K. J. Friston (1995), Analysis of fmri time-series revisited—again, *Neuroimage*, *2*(3), 173–181.

Yan, B., and Y. Liu (2017), Smooth image-on-scalar regression for brain mapping, *arXiv preprint arXiv:1703.05264*.

Yang, C., R. Duraiswami, and L. S. Davis (2005), Efficient kernel machines using the improved fast gauss transform, in *Advances in neural information processing systems*, pp. 1561–1568.

Yotter, R. A., P. M. Thompson, I. Nenadic, and C. Gaser (2010), Estimating local surface complexity maps using spherical harmonic reconstructions, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 169–176, Springer.

Zhou, H., L. Li, and H. Zhu (2013), Tensor regression with applications in neuroimaging data analysis, *Journal of the American Statistical Association*, *108*(502), 540–552.

Zhu, H., M. Gu, and B. Peterson (2007), Maximum likelihood from spatial random effects models via the stochastic approximation expectation maximization algorithm, *Statistics and computing*, *17*(2), 163–177.

Zhu, J.-Y., R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman (2017), Toward multimodal image-to-image translation, in *Advances in Neural Information Processing Systems*, pp. 465–476.